Islands of misfit buildings: Detecting uncharacteristic electricity use behavior using load shape clustering

Preprint · October 2019

DOI: 10.13140/RG.2.2.11489.86883

CITATIONS
0

READS
404

3 authors:

Matias Quintana
National University of Singapore
20 PUBLICATIONS **31** CITATIONS
SEE PROFILE

Pandarasamy Arjunan
Berkeley Education Alliance for Research in Singapore (BEARS) Limited
19 PUBLICATIONS **120** CITATIONS
SEE PROFILE

Clayton Miller
National University of Singapore
68 PUBLICATIONS **476** CITATIONS
SEE PROFILE

Islands of misfit buildings: Detecting uncharacteristic electricity use behavior using load shape clustering

Matias Quintana[a], Pandarasamy Arjunan[b], Clayton Miller[a],

[a]*Building and Urban Data Science (BUDS) Lab, Dept. of Building, School of Design and Environment (SDE), National University of Singapore (NUS), Singapore*

[b]*Berkeley Education Alliance for Research in Singapore (BEARS), Singapore*

Abstract

Many energy performance analysis methodologies assign buildings a descriptive label that represents their main activity, often known as the *primary space usage (PSU)*. This attribute comes from the intent of the design team based on assumptions of how the majority of the spaces in the building will be used. In reality, the way a building's occupants actually use the spaces can be different than what was intended. With the recent growth of hourly electricity meter data from the built environment, there is the opportunity to create unsupervised methods to analyze electricity consumption behavior to understand whether the PSU assigned is accurate. Misclassification or oversimplification of the use of the building is possible using these labels when applied to simulation inputs or benchmarking processes. To work towards accurate characterization of a building's utilization, we propose a modular methodology for identifying potentially mislabeled buildings using distance-based clustering analysis based on hourly electricity consumption data. This method seeks to segment buildings according to their daily behavior and predict which ones are *misfits* according to their assigned PSU label. The process assigns a flag indicating potential mixed-use or a misclassified PSU label based on uncharacteristic electricity use behavior. Our results on two public data sets, from the Building Data Genome (BDG) Project and Washington DC (DGS), with 507 and 322 buildings respectively, show that 26% and 33% of these buildings are potentially mislabelled based on their load shape behavior. Such information provides a more realistic insight into their true consumption characteristics, enabling more accurate simulation scenarios. Applications of this process and a discussion of limitations and reproducibility are included.

Keywords: Mixed-use buildings, Building energy use, Building energy benchmarking, Building performance rating, Primary-use-type analysis, Load profile clustering

1. Introduction

One of the major goals of researchers in the built environment is to reduce the electricity footprint of the built environment; this objective is crucial for meeting the required sustainability goals to address issues like climate change. As such, the importance of evaluating electricity consumption behavior of existing buildings, as well simulation results of retrofitted infrastructures, is paramount. In recent years, there has been a proliferation of unsupervised machine learning approaches and visual analytics [1, 2] for energy systems in diverse applications such as consumer segmentation [3], operations optimization [4, 5], energy forecasting [6] and anomaly detection based clustering [7]. Such applications benefit building owners by giving them the right tools to evaluate building performance and carry out well-informed inspections to mitigate failure and operational costs and predictive maintenance. These approaches contribute to the increasing collaboration in the interdisciplinary field of statistical learning and visual analytics in the building domain. This field uses Internet-of-Things (IoT) sensors such as smart meters with different temporal resolutions.

1.1. Buildings sometimes behave differently than their label

When it comes to classifying commercial buildings using human-made categories, the *primary space usage (PSU)* label is used to segment the population of buildings into groups of buildings that are theoretically being used for similar purposes. Evaluation methods for buildings rely on these building identifiers, or labels, to facilitate comparison among their industry or conventional use type. For example, if a building has a PSU tag of *office*, the performance analysis of this building uses the boundary conditions of typical office space: occupants are professionals who work 8 am-5 pm and use certain types of office equipment (e.g., computers, printers, etc.). Behavior or consumption that falls outside these predefined norms of this category will trigger further analysis, especially in the categories of building simulation and energy benchmarking. These behaviors ultimately serve as proxies of human activities and occupancy within the buildings; thus, atypical buildings electricity consumption patterns will correspond to atypical usage of appliances and loads by the users in such infrastructure. Understanding the users habits and consumption patterns is key to achieve better *and more sustainable* control strategies [8, 9].

1.2. Importance of labels to building simulation

Building performance simulation techniques require PSU labels as an accurate representation of real buildings. The labels are essential in physics-based models but equally crucial in data-driven simulations when it comes to the operational control schedule. A given building type will have a specific control schedule which determines its operational hours as well as consumption behavior and system boundaries. Thus, the importance of correctly knowing this label type helps designers and engineers perform a more accurate depiction of real-world buildings. These labels are the basis for initial input assumptions such as diversity schedules and plug load power density and are used to understand the electricity patterns that define a building [10]. If a building is mislabeled, the simulation process would be inaccurate and calibration of the model more tedious.

1.3. Importance of labels to building energy benchmarking

Building performance benchmarking is a growing field that includes useful tools for decision-making for portfolio analytics [11], policy-makers, and different stakeholders with variable domain expertise. Such building identification mechanisms provide a more holistic understanding of the different sources of uncertainty, particularly scenario-related ones where external conditions are imposed on the building [12], such as fixed operating hours based on an industry-defined schedule. The label that assigns the use of a building is vital for building energy benchmarking as they determine who the *peers* of a building will be in the analysis. However, as investigated by [13], these labels are inflexible to the reality of modern buildings since most of the current infrastructure cannot be classified entirely as one category. This situation is particularly challenging when we considered the plethora of existing buildings where their labels were assigned decades ago and have had served different uses and purposes over the years.

1.4. Contribution

To address the deficiencies in how a building is assigned a PSU, we propose a modular methodology for systematic validation of the human-made building classification. This methodology leverages its modularity for homogenizing power meter readings based on different temporal contexts and aggregation functions. These PSU labels might not be an accurate representation for most buildings with diverse uses and loads but they still carry enough information for detailed and grey-box simulations [12]. Hence, the building segmentation is carried out in an unsupervised way. The clustering of buildings will make use of raw electricity consumption aggregated over a specific time window, grouping them based on their load profiles. Finally, the resulting clusters are evaluated with commonly used validation metrics and then analyzed to identify potentially misclassified buildings based on the distribution of the buildings in each PSU type. We differentiate this work from previous PSU-focused investigations by focusing on the segmentation of existing labels as opposed to creating a new framework for labelling [13].

1.5. Previous work

In terms of building performance benchmarking and analysis, PSU labels play an essential role in the segmentation and homogenization of their groups. Conventional building benchmarking scenarios aim to establish how much better or worse a given building performs as compared to similar buildings (i.e., peers in the same group). Thus, it is important to specify these groups correctly, and making sure a building truly belongs to its label [14]. However, these human-made categories for grouping (PSU) are inflexible to the reality of modern infrastructures and are often inconsistent with the observed consumption behavior of the building [13].

Existing ways to analyze building electricity consumption and performance can be divided into direct and indirect clustering methods [15]. The former relies on analyzing the raw meter data (i.e., kWh hourly readings), and the latter uses features extracted from the meter data. When using raw meter data, the traditional way of representing such consumption patterns is with daily profiles, usually divided into hourly readings [16]. On the other hand, indirect clustering uses statistical or learned features from the readings themselves. Such features can be defined beforehand and applied to the data [17] or can be automatically extracted [18]. These daily profiles, or load profiles, have a wide range of applications, from building energy simulation to occupancy and load prediction. In terms of clustering, the two most common types of clustering algorithms used for building data are K-Means and Hierarchical Clustering [19, 20, 1]. However, other machine learning techniques such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and K-Shape have also been used primarily for forecasting [21, 19, 22, 23].

In the field of anomaly detection, an end-user perspective is usually the focus [24], interpretation is sought through inverse modeling with some discrepancies due to meta-data [25], or sometimes the process treated as a side-product of building benchmarking. Anomaly detection is sometimes treated as a way of understanding occupancy schedules and user demand [8, 26, 27, 28] and such building performance simulation are occupant-centric and close to real-world conditions [29]. Occupancy behavior is most often used to design better control strategies that suit the dynamic needs and conditions of a user and the indoor space, respectively, [9], but also as a design feature for future buildings yet to be built [30].

Another branch of research focuses more on the benchmarking and assessment of electricity consumption of such buildings [31]. Several studies exist addressing benchmarking models using the machine learning algorithms as mentioned earlier and other sophisticated tools as well (i.e., decision trees [32] and stochastic frontier analysis [14]). [13] summarizes this clearly and proposes three fundamental load shape profiles from raw meter data as a baseline for benchmarking based on a list of more than 3,000 non-residential and residential buildings.

In this paper, we present our proposed methodology in Section 2. Details about the experiments and results are presented in Section 3, and a discussion of the implications of the results and possible applications are found in Section 4. Finally, an

overview of the concluding remarks, reproducibility, and next steps are found in Section 5.

2. Methodology

To build upon the previous literature, we present a methodology that utilizes two open data sets of electrical meter data from buildings to demonstrate the clustering of daily load profiles to detect buildings that can be considered *misfits* as compared to their PSU label peers based on detection of uncharacteristic consumption behavior.

2.1. Datasets

This study focuses on two main data sources of non-residential buildings (see Table 1). The Building Genome Dataset (BDG) project is an open dataset from 507 non-residential buildings that includes hourly, whole building electrical meter data for one year as well as metadata such as industry type and primary space usage. The second dataset comes from the Department of General Services (DGS) from Washington, DC. This public dataset contains 15-minute interval electrical meter data and more building information of many buildings on the district. The overview of the data analysis methodology is presented in Figure 1.

2.2. Preprocessing

Since each dataset has a different data collection period and not all buildings had meter data from all the available periods, we first proceeded to find the time range with the most number of buildings metered. The summary of the re-sampled datasets is found in Table 1. We found that in the Building Genome Dataset (BGD), 368 buildings (roughly 72% of the dataset) have hourly data from January 1st to November 30th of 2015. On the other hand, buildings from the Washington D.C. Dataset (DGS) have hourly data (originally 15-minute readings that were re-sampled to hourly) from February 2nd, 2016 to March 2nd, 2018. Additionally, the DGS dataset had 22 unique labels, with many of them having only one building per category. We decided to drop the least frequent labels (less than 15 buildings) and keep 271 buildings (roughly 84% of the dataset). This step allows us to have the same number of samples for all buildings in their respective datasets. Figure 2 and Figure 3 show the PSU label distribution for the re-sample BDG and DGS dataset, respectively.

2.3. Context Extraction

After limiting the number of points of each dataset, we added a temporal context filter. This step allows us to generate a dataset based on specific situations that will make the data more homogeneous. We implemented three different types of contexts: *weekday*, *weekend*, and *fullweek*. These contexts will only keep the electricity consumption readings that occurred on a weekday, weekend, or any day of a week, respectively. In this way, we create different scenarios where we can analyze the buildings' electricity consumption behavior. This type of analysis is required since commercial buildings have a clearly unique consumption behavior due to operational hours. Figure 4 shows an example of two *Office* buildings from the *BDG* dataset where we can observe the different behaviors across different contexts. Additionally, this modularity allows the user to add any new context if needed.

2.4. Daily Load Profile Generation

As a final step, we extract the daily profile for each building based on an aggregation function. Let $t \in [0, 23]$ be the hour of the day, and $L_d(t)$ the electricity consumption of a building at time t on day d in kWh. The daily profile is expressed as 24 data points, i.e. $L_d(0), \ldots, L_d(23)$. Since the datasets have been filtered to the same time range, their datasets have the same number of daily profiles. Up to this point, each building's hourly readings can be reshaped as the following matrix where n is the number of available days for the building in the dataset (see Table 1):

$$\begin{bmatrix} L_0^1 & L_1^1 & \dots & L_{23}^1 \\ L_0^2 & L_1^2 & \dots & L_{23}^2 \\ \vdots & \vdots & \ddots & \vdots \\ L_0^n & L_1^n & \dots & L_{23}^n \end{bmatrix}$$

Finally, this matrix is aggregated into one daily profile 1x24 vector. We defined an *average* and *median* daily profile by calculating the column-wise mean or median, respectively.

2.5. Building Clustering

The objective of clustering is to find homogeneous groups (clusters) with significant differences among themselves. We use three different clustering algorithms in this paper, but other unsupervised learning algorithms can be used or added. Before applying any algorithm, the datasets were z-Normalized [35]. The purpose of this data standardization technique is to convert all data samples into a common scale for better comparison, a common practice when dealing with distance-based clustering algorithms that emphasis the shape of the profile rather than the value [36].

The first algorithm we studied is K-Shape clustering (see Algorithm 1). This algorithm was developed to be used with time-series data since it uses cross-correlation as a distance measure that is invariant to scaling and shifting. Thus, considering more the shape of the time series than actual values [37].

The second chosen algorithm is K-Means clustering (see Algorithm 2). This algorithm has been applied in many domains [38] and is the basis of K-shape [37]. Also, [1] showed that it is one of the most popular approaches for smart meter and portfolio analysis.

As for the final algorithm, Hierarchical clustering was selected (see Algorithm 3). This unsupervised learning method is also widely used for portfolio analysis alongside K-Means [1] and has two types of approaches: agglomerative and divisive. The latter is a bottom-up approach in which each data point starts as a cluster on its own and pairs of clusters are merged as the hierarchy moves up. The former, a top-down approach, follows the opposite strategy by starting with one big cluster and splits recursively as the hierarchy moves down. For our study, we opted for the agglomerative strategy.

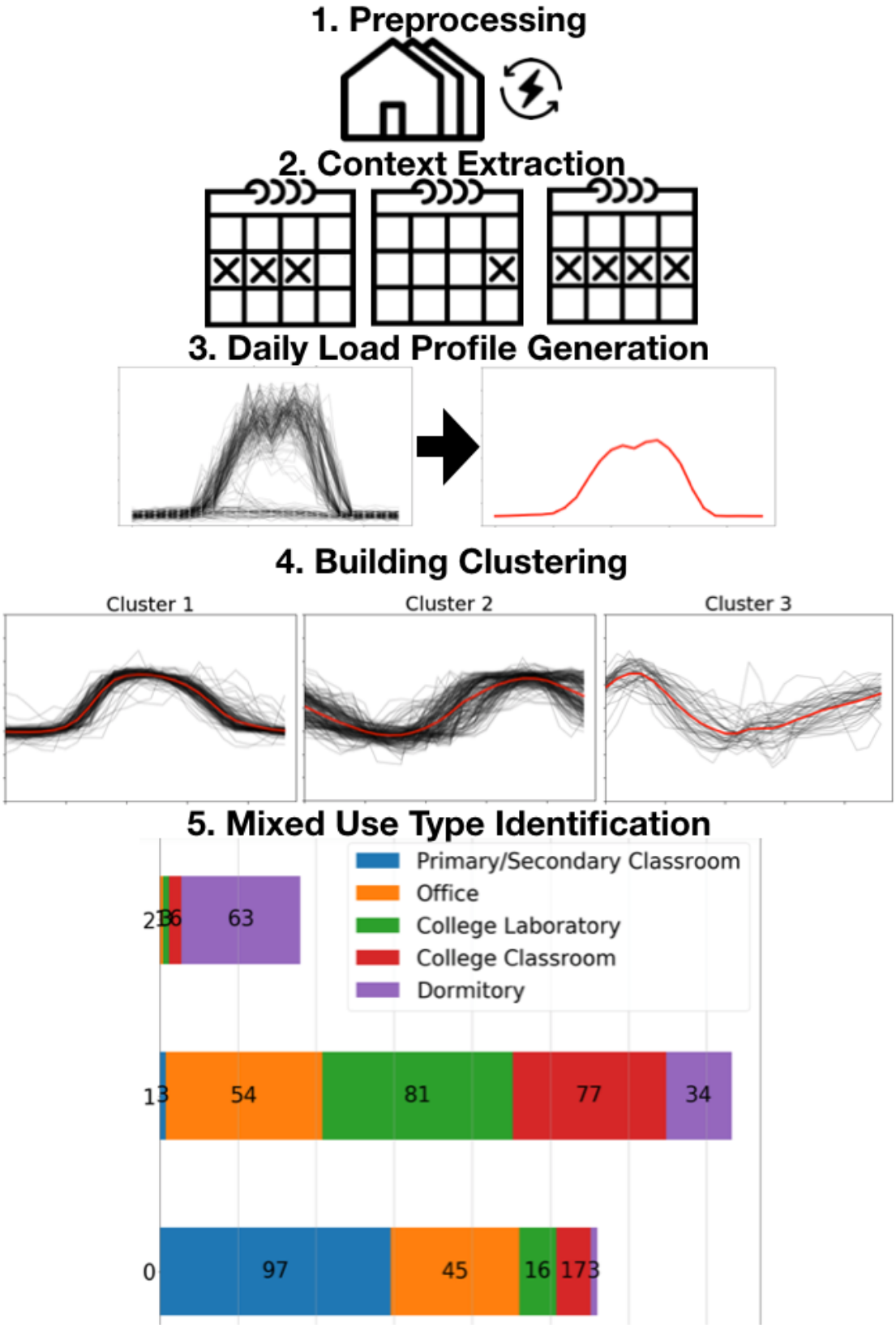


Figure 1: Overview of the proposed modular methodology. First, the temporal context is extracted, then the load profile is generated, later clustering takes place, and finally, the PSU distribution is evaluated to identify the mixed-use type buildings.

Table 1: Dataset details.

Properties	**Dataset 1**	**Dataset 2**
Name	Building Data Genome (BDG)	Washington DC (DGS)
Location	Various	Washington D.C., USA
Reference	[33]	[34]
Number of buildings (original)	507	322
Number of buildings (after cleaning)	368	271
Date range	2015-01-01 - 2015-11-30	2016-02-02 - 2018-03-02
Number of days	334	523
Number of load profiles	122,912	141,733
Number of building types	5	22 (filtered to 7)

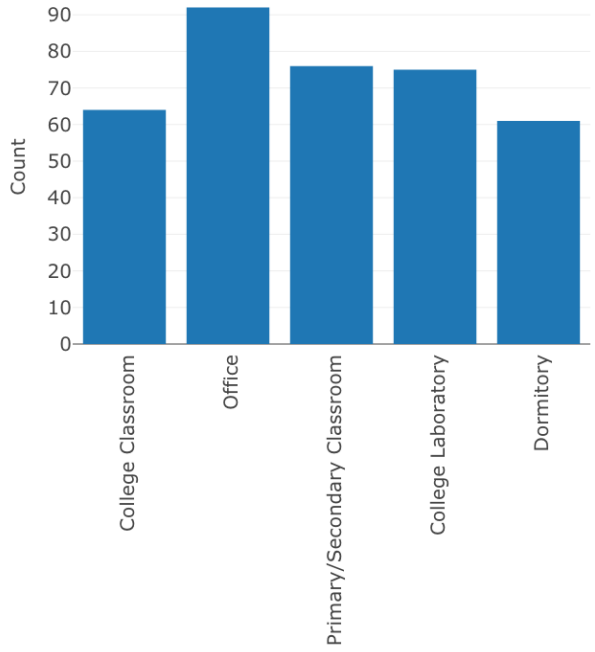


Figure 2: Primary-Space-Usage (PSU) label distribution in filtered Building Genome Dataset (BDG).

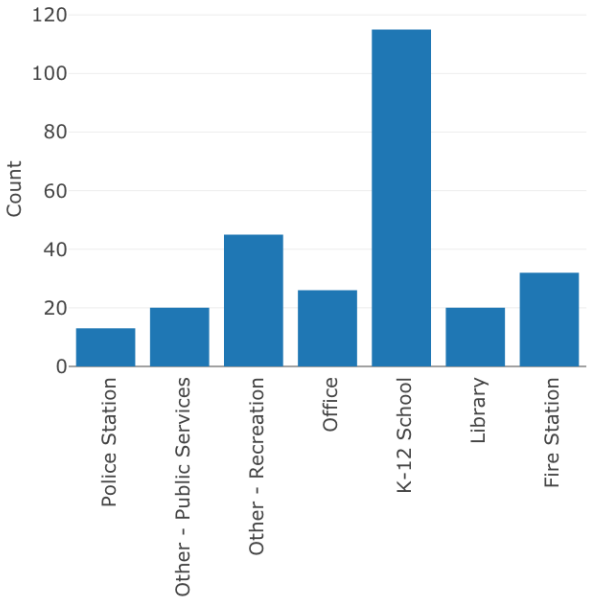


Figure 3: Primary-Space-Usage (PSU) label distribution in filtered Washington D.C. Dataset (DGS).

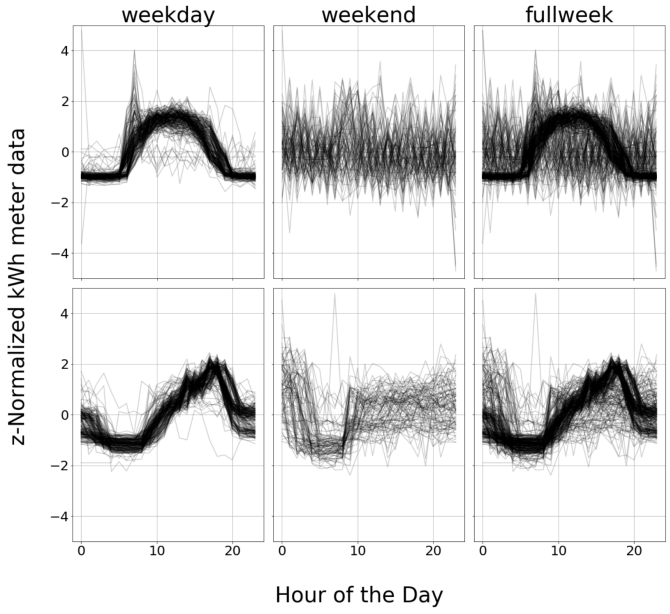


Figure 4: Z-Normalized daily profiles grouped by context for two *Office* buildings from the *BDG* dataset. The temporal contexts clearly show the difference in consumption behavior. Each building had 238, 96, and 334 *weekday*, *weekend*, and *fullweek* load profiles respectively

2.6. Clustering Validation

While many metrics can be used to quantify the quality of a cluster, we settled for the *Silhouette Score (SC)* since it captures the trade-off between within- and between- cluster distances (a and b respectively) and it is easier to interpret [39]. The within-cluster distance is the mean distance between a datapoint i and all other datapoints j in the same cluster *(d(i, j))*. It can be seen as a measure of how well i was assigned to its cluster (the smaller the value, the better).

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j) \tag{1}$$

On the other hand, the between-cluster distance is the smallest mean distance of a datapoint i to all other points j in any other remaining cluster.

$$b_i = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \tag{2}$$

The range of the *SC* is [-1, 1]: negative values indicate incorrect cluster assignment (as a different cluster is more similar), values near 0 indicate overlapping clusters, and values close to 1 highlight the homogeneity of a cluster.

$$SC_i = \frac{b_i - a_i}{max(a_i, b_i)}, \ \epsilon \ [-1, 1] \tag{3}$$

Such a metric was evaluated for a k clusters from 2 to 10. For each algorithm, the number of clusters was determined visually using the elbow method: number of the cluster at which the performance metric stops improving (reaches a plateau)).

2.7. Mixed Use Type Identification

Since all the algorithms perform unsupervised learning, the resulting clusters will not be labeled accordingly to a particular PSU but rather with numerical indices. Therefore, in order to assign a PSU label to such clusters, we assume that the building class determines the underlying PSU label of each cluster with highest number of instances in the cluster, i.e., in a cluster where 90% of all *Office* buildings are tallied, the cluster can be treated as an *Office* cluster.

Algorithm 1: K-Shape Clustering

```
Determine the number of clusters (k)
Initialize k number of centroid randomly
repeat
    for every data point do
        for every centroid do
            - calculate shape-based distance between
              datapoint and centroid
            - assign datapoint to cluster with lowest
              shape-based distance away
        end
    end
    for every cluster do
        - extract cluster shape
        - update centroid to the cluster shape
    end
until no data point has changed cluster assignment
```

Algorithm 2: K-Means Clustering

```
Determine the number of clusters (k)
Initialize k number of centroids randomly
repeat
    for every data point do
        for every centroid do
            - calculate distance between datapoint and
              centroid
            - assign datapoint to cluster with lowest
              distance away
        end
    end
    for every cluster do
        - calculate the cluster mean
        - update centroid to the cluster mean
    end
until no data point has changed cluster assignment
```

Algorithm 3: Agglomerative Clustering

```
Determine the number of clusters (k)
Initialize every data point as a unique cluster
repeat
    for every pair of clusters do
        - calculate distance between clusters
        - merge pair of clusters with the lowest distance
          away
    end
    decrease total number of clusters
until total number of clusters reached k
```

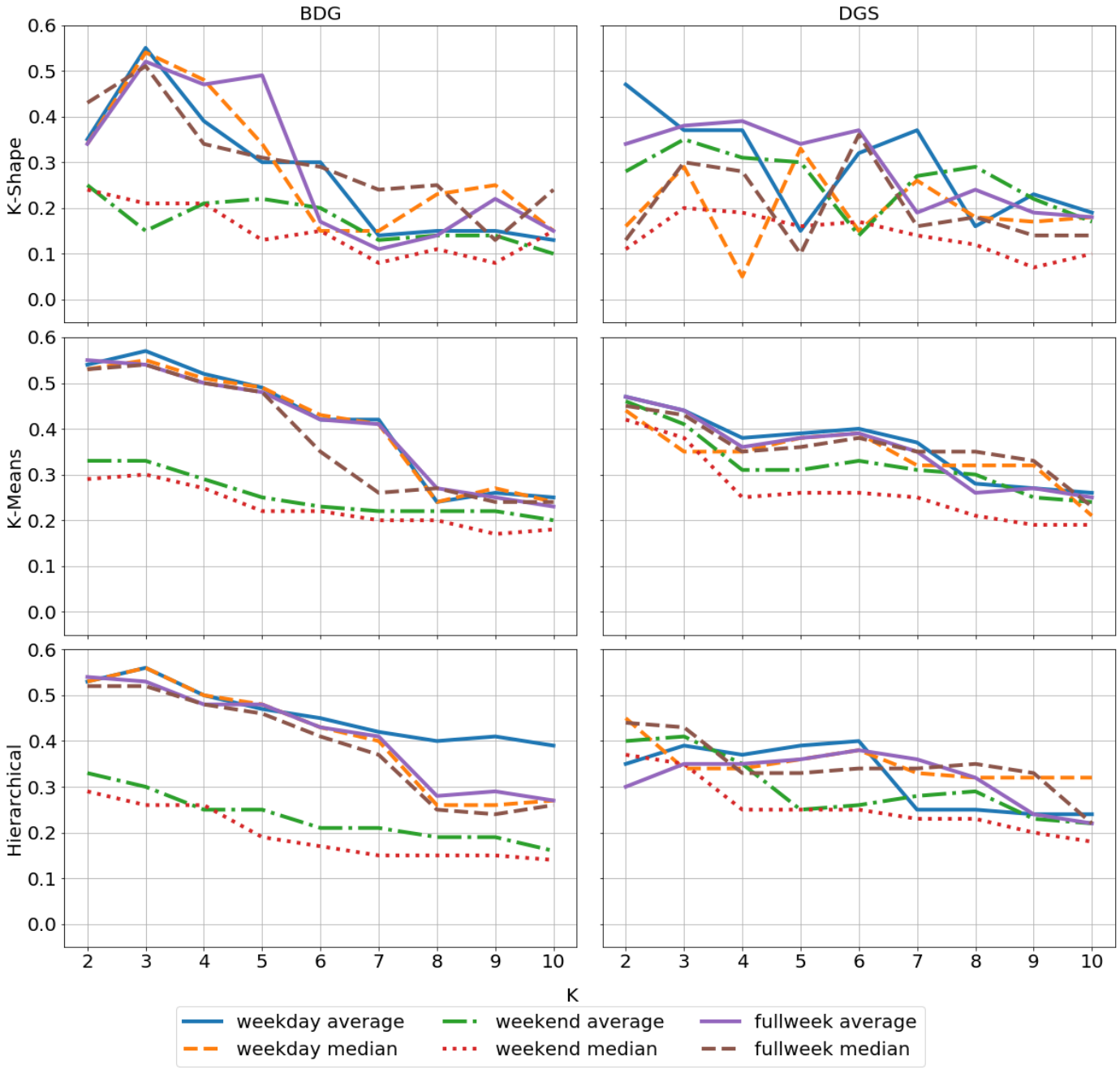


Figure 5: Silhouette Scores for all possible scenarios and datasets. In the BDG dataset, weekday contexts outperform weekends whereas, in the DGS dataset, all temporal contexts are closer in score value.

Table 2: Summary of methodology options for the experiments carried out.

Methodology Step	Options
Dataset	BDG, DGS
Context	weekday, weekend, fullweek
Aggr. Function	average, median
Algorithm	K-shape, K-means, Agglomerative Clustering
Algo. Parameter	k ϵ [2, 10]

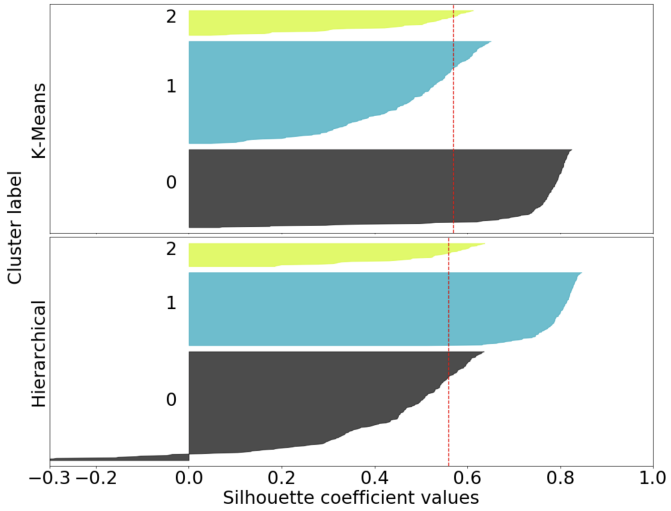


Figure 6: Silhouette scores sorted in each cluster for K-Means and Hierarchical clustering with k = 3. The average score of the algorithm is represented by a dashed red line. K-Means achieves a slightly higher average score mainly because all samples have a positive score, indicating a more homogeneous clustering.

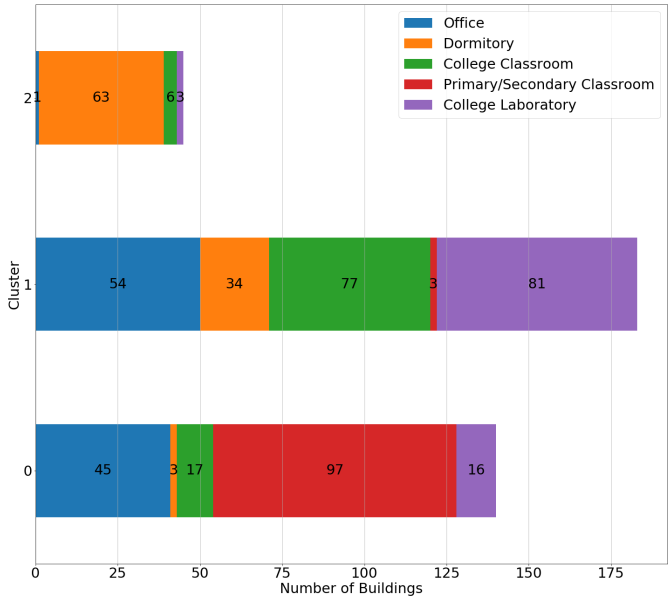


Figure 7: PSU distribution in each cluster for the BDG dataset with a weekday context and average as aggregation function and K-means with k = 3. The number in each color section represents the overall membership percentage of buildings in such PSU.

3. Results

The current methodology allows us to make a plethora of combinations based on datasets, contexts, daily profile aggregation functions, algorithms, and the different parameters for the algorithm. The total number of possible experiments will be determined by multiplying their available options. Table 2 shows the different options we used for a total of 486 experiments; this number is the result of all the possible combinations of the methodology steps. The modularity in this methodology allows the incorporation of more options in any step of the pipeline, i.e., if holidays or a regression profile needed to be assessed, the user would only need to define them as a new context and daily profile aggregation function respectively.

As mentioned previously, to evaluate the performance of the cluster for different values of k across the different experiment setups, we focused our attention on the Silhouette Score. While other metrics are also widely used to compare clustering performance (i.e., cohesion, separation, R-squared, etc.), in our results, the Silhouette Score was more natural to interpret [39].

Figure 5 shows the tendency of the metric (y-axis) with different values of k (x-axis) for all datasets and algorithms across the combinations of context and daily load profile aggregation functions (colored curves). For the *BDG* dataset, we see that all contexts but the ones including *weekend* perform relatively similar, with the highest achieved at $k = 3$. Conversely, for the *DGS* dataset, all contexts follow the same tendency and similar variations, with K-Shape showing more variations. In this case, although $k = 2$ has the highest value, at $k = 4$ we notice the elbow of such metric, thus indicating the scores stabilization. In terms of choosing the algorithm for the *BDG* dataset, a more detailed analysis is reflected in Figure 6, where the average Silhouette scores for both algorithms with their respective setups are represented by the dashed red line and all the data points' Silhouette scores are sorted within their respective cluster. From these plots, it can be argued that both average scores for K-Means and Hierarchical Clustering, 0.57 and 0.56 respectively, are so close that either can be chosen as the clustering algorithm. However, Hierarchical Clustering results show some instances within Cluster 0 with negative silhouette scores. As stated in Section 2.6, this most likely indicates that those instances are closer to instances in other clusters rather than instances in their current cluster.

On the other hand, K-Means show all data points with positive silhouette scores, making it our selected algorithm for this particular dataset. The selection for the *DGS* algorithm is more

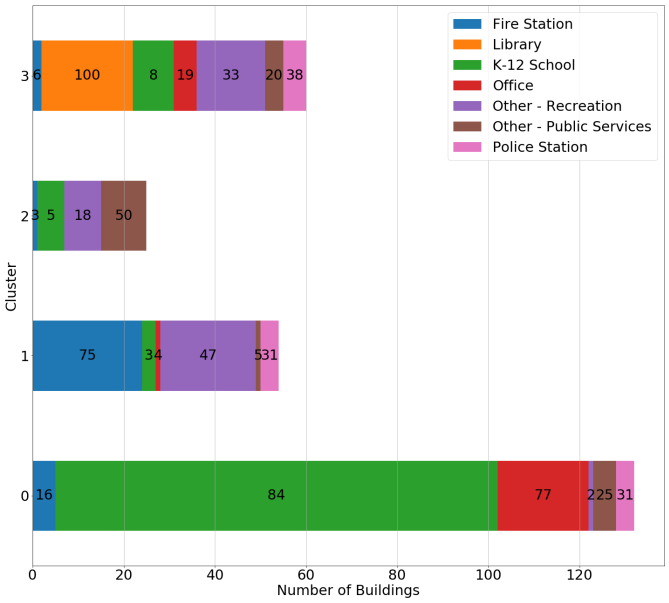


Figure 8: PSU distribution in each cluster for the DGS dataset with a weekday context and average as aggregation function and K-means with k = 4. The number in each color section represents the overall membership percentage of buildings in such PSU.

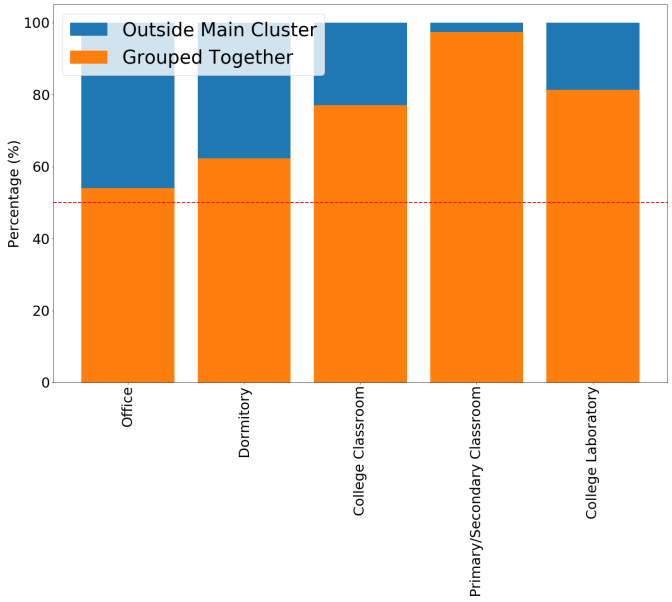
straightforward since K-Means shows more consistent performance and a clear elbow. While K-Shape is known to perform well on time series data [37], the authors believe it does not perform as expected on these time series datasets because of the aggregation functions that are smoothing details and intricacies of the raw data, which are usually exploited by the algorithm.

Figure 7 shows the distribution of PSU in the resulting cluster, as well as the percentage across all buildings in their PSU. As per the final step in our methodology, the PSU label of each cluster matches the label of the majority of its members. However, since the optimal number of clusters is different from the total number of labels (i.e., three clusters versus five PSU labels in the *BDG* dataset), some clusters will have a mixed PSU. Cluster 0 can be treated as a *Primary/Secondary Classroom* cluster. Conversely, Cluster 1 shows three predominant PSU label, meaning the Cluster is a mix of *College Laboratory*, *College Classroom*, and *Office*. Finally, Cluster 2 can be treated as *Dormitory*. Figure 8 shows the same type of results for the *DGS* dataset. In this scenario, Cluster 0 shows predominantly *K-12 School* and *Office* buildings. Cluster 1 also shows a mixture of *Fire Station* and *Other - Recreation*. Cluster 2 clearly shows a predominance of *Other - Public Services* buildings, and finally, Cluster 3 is mostly represented by *Library* buildings.

4. Discussion

Based on our initial assumption that the PSU label of a formed cluster will be the same PSU label of the majority of buildings in the cluster, and looking at Figure 9 and 10, we find that on average 74% and 67% of buildings in the *BDG* and *DGS* dataset had a PSU that matches their profile consumption with their peers, respectively. While we do not have information to

Figure 9: Summary of membership classification as one main cluster for each PSU label for the BDG data set with with K-Means and k = 3. A horizontal 50% line is represented by the red dashed line. The buildings falling outside the main cluster for each of these buildings can be interpreted as having uncharacteristic energy use behavior as compared to their peers, and could thus be labelled as *misfits*.



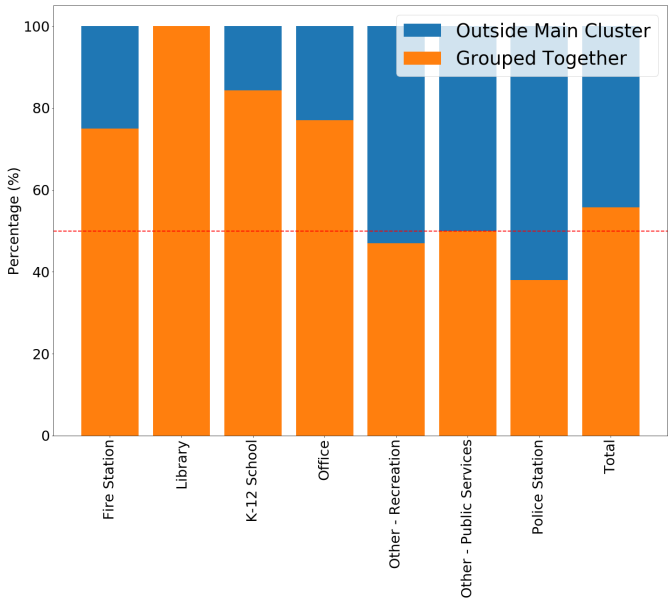


Figure 10: memberships classification as one main cluster for each PSU: dataset DGS, context and function weekday-average with K-Means and k = 4. A horizontal 50% line is represented by the red dashed line. Once again, many of the buildings can labelled as *misfits*.

explain why other buildings are tallied in different clusters and thus is split into multiple PSU groups, we do have a hint from the figures as mentioned earlier. In Figure 7, we see that while the predominant PSU's in each cluster has a significantly high percentage value, meaning the vast majority of those buildings are present in such cluster, the buildings labeled *Office* are split into two clusters. This split is not entirely skewed towards one cluster, which could suggest that many buildings labeled as *Office* are actually mixed-used buildings with similar load profiles (during weekdays, since that is the context we are analyzing) to *College Laboratories* and *College Classrooms* (cluster 1) and *Primary/Secondary Classrooms* (cluster 0). This situation can be highlighted in Figure 4, where the first *Office* building (first row in the Figure) was found to be in cluster 0 and the other *Office* building (second row) in cluster 1, both showing different consumption behaviors for the same PSU. We can argue that similar situations happen in Figure 8. However, we have to highlight that the distribution of human-made PSU in this dataset is heavily imbalanced (Figure 3), making it hard to have stronger claims on such results. With a more evenly distributed number of buildings per label or by using undersampling techniques, the distribution of PSU in each cluster showed in Figure 8 could vary significantly.

Overall, this analysis is focused on the identification of electricity use behavior that is divergent from the *status quo* as compared to the peer group that the PSU label represents. Many of the buildings in the two data sets exhibited load profile shapes that indicated that they could require more detailed investigation through collection of sub-meter data or a walk-through analysis to understand these discrepancies. Discussion of the impact on of building benchmarking and simulation are most relevant to this analysis.

4.1. Building energy performance benchmarking

The expansion of building performance benchmarking programs have made a positive impact on the quest to identify energy savings opportunities in the built environment. The methodologies of these platforms are built on data from energy use surveys and utility disclosure programs. A significant part of the benchmarking process is to assign a building to a peer group based on the building attributes. This peer group makes up the set of buildings that the targeted facility is being compared to create a rating. If the building is performing poorly, then it will fall at the bottom of that group. However, if the building has atypical uses that are different from its peers, then it might be misclassified as a poor performer. A classic example of this type of behavior is when an office building has occupants who work more hours than the usual office worker. The company leasing the building will often not discourage its employees form putting in extra hours if needed for the sake of energy performance. These types of behaviors are often not formally documented nor used as input for the benchmarking systems. Our methodology could be used to shine some light onto those atypical buildings and set the stage to a more in-depth understanding of why their performance is more unsatisfactory when compared to their peers.

4.2. Building simulation assumptions

The importance of the mixed-use or misfit labels, alongside their most representative load profile, is crucial for the development and calibration of physics-based building performance simulation models. A modeler would use this insight to choose the space use classifiers and their associated inputs more carefully. These decisions can impact the speed of calibration, especially in the case of setting input parameters initial distributions for optimization techniques focused on calibration. The mixed-use information can be further applied in benchmarking scenarios and in detailed and grey-box modeling applications: when operations schedules are needed, a real-world depiction of the building, most of the time, a building with diverse loads and uses, could lead to better convergence of measured and simulated data.

5. Conclusion

In this paper, we presented a modular methodology to assess the validity of human-made primary-space-use (PSU) labels and the uncharacteristic behavior in hourly load shape data that can identify misfit buildings. This methodology allows a plethora of comparisons and experiments to understand the best homogeneous scenarios further to compare building load profiles and group them in an unsupervised way. The resulting clusters are renamed based on the majority of buildings inside it. The latter is the starting assumption the authors made for this work since it is expected that buildings with similar consumption patterns (load curves) follow the same primary usage type. This method allowed us to find 74% and 67% of buildings with the same label group together in the *BDG* and *DGS* dataset, respectively. These results show that 26% and 33% of buildings in the datasets respectively can be treated as mixed-use or PSU outliers with uncharacteristic behavior. It is this methodology and its results that these buildings can be pin-pointed to the engineers and designers to further understanding their circumstances and drill down on what makes their electricity consumption so different than those from their respective peers.

5.1. Limitations

One limitation of this work is further assessment as to why the remaining buildings have this behavior. Hints, such as mix-use type, can be found in Figures 7 and 8. Secondly, the identification of clusters relies entirely on the algorithm selected and the different parts of the experiment parameters (context and aggregation function). Thirdly, since *holidays* were not filtered out or treated as *weekends*, they could have inserted noise if they happen to fall on a *weekday*; future work will address this scenario. Finally, it is possible that another unsupervised, or supervised learning technique, achieves a better Silhouette Score than the three algorithms presented here. Regardless, these algorithms were chosen due to their extended use in the related literature. On the other hand, the modularity of the methodology simplifies the testing and incorporation of new parameters in any stage of the pipeline: temporal context, aggregation function, clustering technique, and its different parameters.

5.2. Reproducibility

This publication is fully reproducible using the codebase from a Github repository[1] and data publicly available from the Building Data Genome Project[2].

References

[1] C. Miller, Z. Nagy, A. Schlueter, A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings, Renewable and Sustainable Energy Reviews 81 (2018) 1365–1377.

[2] C. Fan, F. Xiao, Z. Li, J. Wang, Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review, Energy and Buildings 159 (2018) 296–308.

[3] S. Xu, E. Barbour, M. C. González, Household Segmentation by Load Shape and Daily Consumption, In Proceedings of. ACM SigKDD 2017 conference (2017) 1–9.

[4] F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, Energy and Buildings 75 (2014) 109–118.

[5] S. Wang, C. Yan, F. Xiao, Quantitative energy performance assessment methods for existing buildings, Energy and buildings 55 (2012) 873–888.

[6] P. Lusis, K. R. Khalilpour, L. Andrew, A. Liebman, Short-term residential load forecasting: Impact of calendar effects and forecast granularity, Applied Energy 205 (2017) 654–669.

[7] P. Arjunan, H. D. Khadilkar, T. Ganu, Z. M. Charbiwala, A. Singh, P. Singh, Multi-User Energy Consumption Monitoring and Anomaly Detection with Partial Context Information, Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments - BuildSys '15 (2015) 35–44.

[8] R. P. Richner, Research Collection, BRISK Binary Robust Invariant Scalable Keypoints (2011) 12–19.

[9] J. Y. Park, M. M. Ouf, B. Gunay, Y. Peng, W. O'Brien, M. B. Kjærgaard, Z. Nagy, A critical review of field implementations of occupant-centric building controls, Building and Environment (2019) 106351.

[10] M. Heidarinejad, J. G. Cedeño-Laurent, J. R. Wentz, N. M. Rekstad, J. D. Spengler, J. Srebric, Actual building energy use patterns and their implications for predictive modeling, Energy Conversion and Management 144 (2017) 164–180.

[11] J. Ploennigs, B. Chen, P. Palmes, R. Lloyd, E2-diagnoser: A system for monitoring, forecasting and diagnosing energy usage, IEEE International Conference on Data Mining Workshops, ICDMW 2015-Janua (2015) 1231–1234.

[12] D. Coakley, P. Raftery, M. Keane, A review of methods to match building energy simulation models to measured data, Renewable and Sustainable Energy Reviews 37 (2014) 123–141.

[13] J. Y. Park, X. Yang, C. Miller, P. Arjunan, Z. Nagy, Apples or Oranges? Identification of fundamental loadshape profiles for benchmarking buildings using a largeand diverse dataset, Applied Energy (2018) 1–37.

[14] Z. Yang, J. Roth, R. K. Jain, DUE-B: Data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis, Energy and Buildings 163 (2018) 58–69.

[15] Y. Wang, Q. Chen, T. Hong, C. Kang, Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges, IEEE Transactions on Smart Grid (2018) 1–24.

[16] R. Granell, C. J. Axon, D. C. Wallom, Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles, IEEE Transactions on Power Systems 30 (2015) 3217–3224.

[17] C. Miller, Screening Meter Data: Characterization of Temporal Energy Data from Large Groups of Non-Residential Buildings, Ph.D. thesis, 2016.

[18] M. Christ, N. Braun, J. Neuffer, A. W. Kempa-Liehr, Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package), Neurocomputing 307 (2018) 72–77.

[19] D. Hsu, Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data, Applied Energy 160 (2015) 153–163.

[20] A. Al-Wakeel, J. Wu, N. Jenkins, K -Means Based Load Estimation of Domestic Smart Meter Measurements, Applied Energy 194 (2017) 333–342.

[21] A. S. Ahmad, M. Y. Hassan, M. P. Abdullah, H. A. Rahman, F. Hussin, H. Abdullah, R. Saidur, A review on applications of ANN and SVM for building electrical energy consumption forecasting, Renewable and Sustainable Energy Reviews 33 (2014) 102–109.

[22] G. Tardioli, R. Kerrigan, M. Oates, J. O'Donnell, D. Finn, Data driven approaches for prediction of building energy consumption at urban level, Energy Procedia 78 (2015) 3378–3383.

[23] J. Yang, C. Ning, C. Deb, F. Zhang, D. Cheong, S. E. Lee, C. Sekhar, K. W. Tham, k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement, Energy and Buildings 146 (2017) 27–37.

[24] J. Wu, J. Zhao, Evaluation on Building End-user Energy Consumption Using Clustering Algorithm, Procedia Engineering 121 (2015) 1144–1149.

[25] H. Burak Gunay, W. Shen, G. Newsham, A. Ashouri, Detection and interpretation of anomalies in building energy use through inverse modeling, Science and Technology for the Built Environment 25 (2019) 488–503.

[26] I. E. Bennet, W. O'Brien, Office building plug and light loads: Comparison of a multi-tenant office tower to conventional assumptions, Energy and Buildings 153 (2017) 461–475.

[27] F. Kung, S. Frank, S. Pless, R. Judkoff, Meter-based synthesis of equipment schedules for improved models of electrical demand in multifamily buildings, Journal of Building Performance Simulation 12 (2019) 388–403.

[28] D. Malekpour Koupaei, F. Hashemi, V. Tabard-Fortecoëf, U. Passe, A Technique for Developing High-Resolution Residential Occupancy Schedules for Urban Energy Models, The Symposium on Simulation for Architecture and Urban Design (2019).

[29] M. M. Ouf, H. B. Gunay, W. O'Brien, A method to generate design-sensitive occupant-related schedules for building performance simulations, Science and Technology for the Built Environment 25 (2019) 221–232.

[30] B. Dong, D. Yan, Z. Li, Y. Jin, X. Feng, H. Fontenot, Modeling occupancy and behavior for better building design and operation—A critical review, Building Simulation 11 (2018) 899–921.

[31] W. O'Brien, H. B. Gunay, Do building energy codes adequately reward buildings that adapt to partial occupancy?, Science and Technology for the Built Environment 25 (2019) 678–691.

[32] H. S. Park, M. Lee, H. Kang, T. Hong, J. Jeong, Development of a new energy benchmark for improving the operational rating system of office buildings using various data-mining techniques, Applied Energy 173 (2016) 225–237.

[33] C. Miller, F. Meggers, The Building Data Genome Project: An open, public data set from non-residential building electrical meters, Energy Procedia 122 (2017) 439–444.

[34] B. DC, Building directory, "http://www.buildsmartdc.com/ buildings", n.d., Accessed: 2018-12-28, 2018.

[35] C. Miller, Z. Nagy, A. Schlueter, Automated daily pattern filtering of measured building performance data, Automation in Construction 49, Part A (2015) 1 – 17.

[36] D. Q. Goldin, P. C. Kanellakis, On similarity queries for time-series data: Constraint specification and implementation, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 976 (1995) 137–153.

[37] J. Paparrizos, L. Gravano, k-Shape: Efficient and Accurate Clustering of Time Series (2015).

[38] A. K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognition Letters 31 (2010) 651–666.

[39] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational North-Holland and Applied Mathematics (1987) 53–65.

[1] `https://github.com/buds-lab/island-of-misfit-buildings`

[2] `https://github.com/buds-lab/the-building-data-genome-project`