A clustering framework to identify non-residential building archetypes using subsystem-level targeting

Rajeev Kotha[1], François Lédée[1], Mohammad Haris Shamsi[1], Ralph Evins[1]
[1]Energy in Cities group, Department of Civil Engineering, University of Victoria, Victoria, BC, Canada

Abstract

With the proliferation of building energy management systems and smart meters, high-resolution time-series data have become more easily accessible. The commercial building stock is often represented through aggregated representation that do not consider the time-series profiles. It is crucial to determine the types of patterns in these profiles to provide a deeper understanding of the building's operations. This study proposes a clustering framework that uses time-series trends to identify commercial building archetypes. The case study uses synthetic time-series data generated using US DOE building archetype EnergyPlus models. Results indicate that water heater use side outlet temperature as a feature is effective in formation of a building cluster which contains buildings that closely resemble a medium office building archetype (homogeneity score = 0.8). Stakeholders could use the identified clusters to model deep retrofits or perhaps run targeted energy-efficiency campaigns.

Introduction

The World Green Building Council and Architecture 2030 have advocated for 100% of existing buildings to be net zero by 2050 (Laski and Burrows 2017). The decision-making processes geared towards understanding the improvements in building energy performance need to be constructed on reliable scientific evidence. Over the past decade, Building Automation and Control System (BACS) have become immensely popular in the built environment for data acquisition and efficient building control. The Building Management System (BMS) market is poised to be valued at 19.25 billion USD by 2023; a staggering increase compared to 6.65 billion USD in 2016 (Jradi et al. 2021). Generating valuable insights from this voluminous source of data is a highly complex computational task.

Previous work using sensor point data originating from the building automation system (BAS) or building management system (BMS) was quite siloed due to lack of a homogeneous representation of data points within building systems (Gallaher et al. 2004). Recent developments in semantic model standards such as Project Haystack (Praire et al. 2016) and Brick (Balaji et al. 2016) potentially allows for a more effective analysis on the BAS/BMS sensor outputs. Shi et al. (2019) used multiple clustering techniques including agglomerative hierarchical clustering (AHC) on a large-scale dataset containing more than 40 buildings and 65,000 BMS points with manually translated ground-truth point labels to investigate the use of time series features for BMS point type inference and proposed that when raw metadata contains limited definitions of sensor points, time series data should be used to compliment the input data.

A critical issue facing the building industry is the capability of traditional building performance analysis techniques to handle the vast amount of detailed, temporal data streams (Miller, Nagy, and Schlueter 2018). While analyzing building information tagged with less known metadata, generating insights about the corresponding built environment is a challenge that could be addressed through unsupervised machine learning techniques. Such methods formulate the relationship between observations in an exploratory manner. Previous research pertaining to building performance analysis that used unsupervised learning methods resulted in a plethora of applications. These include portfolio analytics (such as targeting, classification), operations (such as energy management, occupancy detection), anomaly detection (such as whole building, subsystem), and smart meter (load profiling, customer classification) (Miller, Nagy, and Schlueter 2018).

Clustering is one of the most common general unsupervised techniques applied to building performance data. It is used to automatically generate subgroups of similar types of observations based on certain features. The two most common types of clustering applied to building performance data are k-means and hierarchical clustering (Miller, Nagy, and Schlueter 2018). (Park et al. 2019) performed clustering analysis using unsupervised clustering methods to un This study integrates concepts from the sub-domains of portfolio analytics and anomaly detection to identify crucial subsystem trends which can tie together a group of buildings into a cluster.

Portfolio analysis refers to a domain where a group of buildings, located in the same geographical area, are studied

for the purpose of managing or optimizing the group as a single entity. Within this domain, lies targeting, a concept that builds upon characterization and classification to recognize specific buildings or measures to be implemented in a portfolio to increase performance (Miller, Nagy, and Schlueter 2018). Petcharat, Chungpaibulpatana, and Rakkwamsuk (2012) used cluster analysis to estimate potential energy savings from the lighting systems across a group of commercial buildings based on the lighting power density measure. Additionally, the authors determined that that the clustering approach can also be applied to other parameters including equipment power density to accurately evaluate the impact of energy codes on building energy performance. Lévy et al. (2021) performed a cluster analysis on a group of sites using their time-series based energy consumption profiles to develop recommendations on energy savings.

Anomaly detection in buildings is focused on identification and diagnostics of issues that occur within a building. This field relies on the use of clustering methods to find anomalous behavior. Stamatescu et al. (2019) state that within a modular BAS structure, the HVAC subsystem plays a dominant role as both the main driver of energy consumption and a driver of user satisfaction with the working environment. Analyzing subsystem or individual component behavior allows for analysis using broader data set to identify and diagnose faults from a lower level. Yoshida et al. (2008) used a semi-supervised approach to determine the most influential variables that affect the overall building performance. Similarly, Le Cam et al. (2014) used PCA to create inverse models to detect issues in HVAC systems. According to Hensel (2013), the highest potential for innovation is at the systemic level where the positive system synergies can be exploited.

To measure the distance between various time series, different distance measures are used, such as Euclidean distance and its variants, Discrete Fourier Transform, and Dynamic Time Warping (DTW). Originally used in the field of text data matching and pattern recognition, DTW computes the similarity of two time series providing both a distance measure that is insensitive to local compression and stretches and warping, which optimally deforms one of the two input time series onto the other (Tormene et al. 2009). This paper computes DTW using *dtw-python*, a freely available Python implementation of the algorithm (Giorgino 2009). Bode et al. (2019) applied several unsupervised clustering algorithms on BACS time series data. One data set consisted of data for one day in high resolution and another for 20 days in reduced resolution. As the accuracies of clustering methods such as DTW and Mini batch K-means were found to be insensitive to the time frame or resolution of the data, the authors recommend focusing on larger time frames to capture representative time series behaviors of classes such as heat flow, power, temperature setpoints in a building.

Lack of reproducibility has been cited as one of the major issues when it comes to the use of unsupervised machine learning in building performance. Additionally, another challenge is the lack of clarity regarding the optimal technique for a particular application. Different studies use unique datasets and methodologies, hence do not render a straightforward comparison. Existing design methods and metrics do not fully capture or enable the quantification of the complex synergies that emerge during operation (Lumpkin, Horton, and Sinfield 2020).

This work proposes a generalized clustering-based methodology to identify prominent subsystem characteristics belonging to a specific building archetype (DOE medium office building in this study) using DTW as the distance measure. The generated distance matrix is then analyzed and visualized using various clustering methods such as agglomerative hierarchical clustering and k-medoids.

This paper is structured as follows: Section 2 describes the proposed workflow of mapping variables and performing clustering exercise. Section 3 lists the key results of a clustering exercise when considering a specific building archetype. Section 4 focuses on the discussions, interpretations, and limitations of the research. Section 5 lists the key conclusions of the research.

Methods

Multivariate time series data are widely available in daily life and research in multivariate time series classification has become more and more prominent during the past decades. BMS provides visibility and control over three broadly important aspects of a building, namely: energy consumption, security, and occupant comfort. Digital Twins (DTs) combine artificial intelligence, machine learning, and data analytics to create a simulation model that learns and updates in real time from heterogeneous and disjoint data sources as well as represent and predict the present and future conditions of physical counterparts. Time-series surrogate models using synthetic datasets potentially reduce the computation time of the physics-based building energy models employed in DTs. This study looks at the use of DTW-based clustering analysis in the identification of prominent subsystem features, available within a building BMS. The overall process involves a BMS output mapping procedure, synthetic data generation, distance matrix computation, and clustering analysis for the whole set of buildings and for the medium office building separately (Figure 1).

Output Mapping Procedure

This is the initial step in the workflow and involves the mapping of individual BMS outputs to the outputs of an EnergyPlus IDF. The mapping is crucial so that the synthetic data outputs retain a link to the real-time variables of the BMS (Table 1). These features are identified and

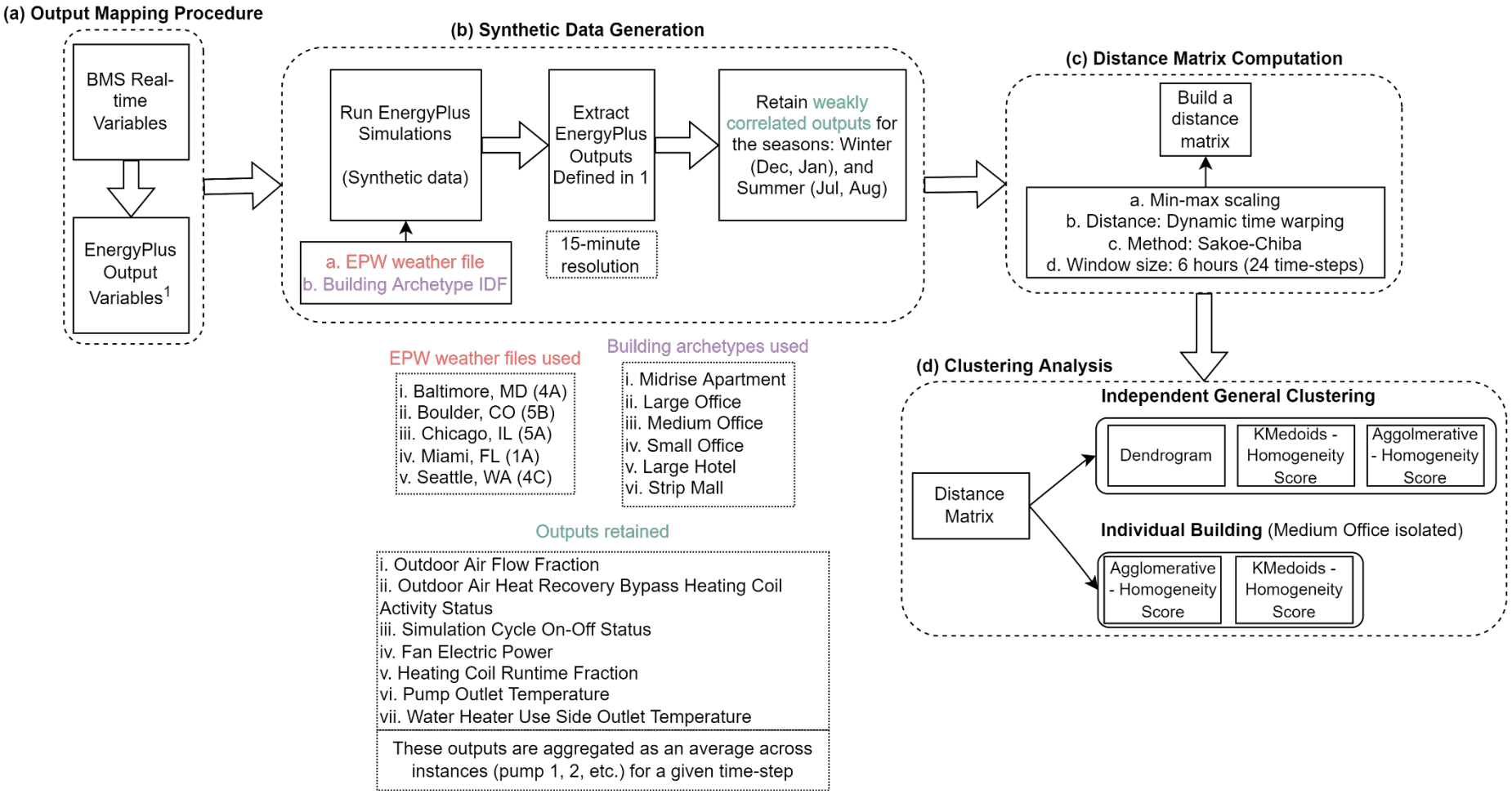

Figure 1 Process for running a clustering analysis on BMS data to identify representative features of a building archetype

selected after running a correlation matrix, thus eliminating strongly correlated features (Figure 2) for a sample of the matrix). It is worthwhile to note that some of the BMS output variables do not directly map to the EnergyPlus output variables (such as damper position), instead they represent easily extractable and relevant proxies to the originally intended BMS output variables.

Table 1 BMS-EnergyPlus variable mapping.

No.	BMS Output Variable	EnergyPlus Output Variable	Unit
1	Outdoor Air Damper Position Signal	Air System Outdoor Air Flow Fraction	-
2	AHU System Status	Air System Outdoor Air Heat Recovery Bypass Heating Coil Activity Status	-
3	AHU Run Mode	Air System Simulation Cycle On-off Status	-
4	Supply/Return Fan Amperage	Fan Electric Power	W
5	Heating Coil Valve Opening Fraction	Heating Coil Runtime Fraction	-
6	Pump Supply Water Temp.	Pump Outlet Temp.	°C
7	Boiler Leaving Water Temp.	Water Heater Use Side Outlet Temp.	°C

Synthetic Data Generation

This step involves the EnergyPlus simulation runs, and the extraction of outputs as identified in the output mapping procedure. The EnergyPlus simulation runs generate the synthetic dataset for an entire year on a 15-minute resolution basis, which corresponds to (35,040x7) data points for one building. The first number (35,040) corresponds to the yearly 15-min time steps and the second number (7) corresponds to the number of features. The simulations are generated for five weather files including Baltimore (Climate Zone 4A), Boulder (5B), Chicago (5A), Miami (1A), and Seattle (4C). The simulations are generated for six building archetypes including midrise apartment, large office, medium office, small office, large hotel, and strip mall. (35,040x7) datapoints are generated across each of the 30 buildings (5 climate zones x 6 building types). As calculation of the DTW distance measure is a computationally intensive process, the time series datapoints across the seven features are limited just to the representative summer (Jul, Aug) and winter (Dec, Jan) months.

Distance Matrix Computation

This step involves the creation of the distance matrix based on a list of results and indexes. Many clustering algorithms do not require the objects as input, but their relations. When it comes to time series mining, the relations can be defined as the distance between each pair of instances (Silva and Batista 2016). In this study, the DTW distance is computed for each feature, comparing the different building archetypes and locations. The measure of DTW is used in this paper as several independent studies have shown that for the problem of time series classification, the Nearest Neighbor DTW is hard to beat (Bagnall et al. 2016). Additionally, any effort to beat the DTW algorithm is at the cost of huge effort in coding or complexity of implementation, and a large time and space overhead. As DTW is a distance measure that is symmetric $\{D(A, B) = D(B, A)\}$ and has a constancy of self-similarity $\{D(A/A) = 0\}$, we calculate the results only for the upper triangle of the distance matrix, thus reducing the time and effort needed for the overall computation (Mueen and Keogh 2016). We select a month as the frequency for computation as it reduces the number of pairs to compare. Although weeks drastically reduce the difficulty for one comparison, it adds a lot of points for further analysis. The data is rescaled using the min-max scaler. Min-max normalization performs a linear alteration on the original data. The benefit of this normalization method is that all the values are annealed within certain range (Saranya and Manikandan 2013). Lastly, the DTW distances are calculated using the *Sakoe-Chiba* method and a window size of 6 hours (i.e., 24 time-steps). The *Sakoe-Chiba* band was shown to be quite promising towards improving the accuracy of time series classification, especially upon constraining the warping window (Geler et al. 2019). These distances are calculated simultaneously, and the resulting matrix is stored in a variable. The distance matrix creation in this study took approximately 71 minutes.

Clustering Analysis

This step involves running a series of analyses on the distance matrix computed during the previous stage. First, all buildings are labeled corresponding to their archetype (such as medium office building, strip mall). In this study, this is referred to as an *independent general clustering* exercise. Among the independent general clustering methods, the following ones were put to test: dendrograms, and two sets of homogeneity scores, one using k-medoids and the other using agglomerative clustering. Next, the buildings are labeled in a binary manner based on the target building archetype (i.e., medium office building vs. non-medium office building). This is referred to in this paper as *isolated building clustering*. In this study, clustering is done by isolating the medium office building (target building), followed by calculation of homogeneity scores. As before, the homogeneity scores in this case are computed using k-medoids and agglomerative clustering techniques.

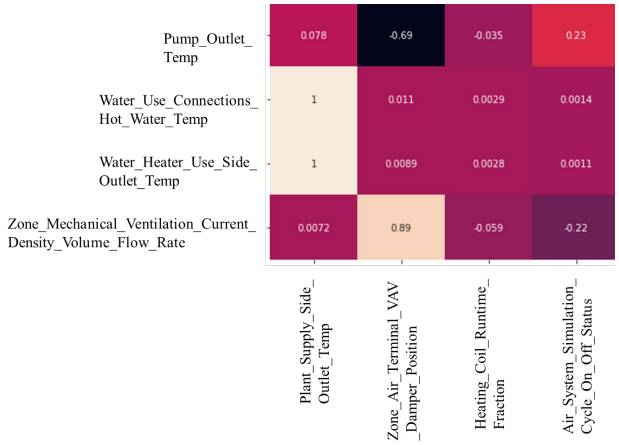

Figure 2 A snapshot of the correlation matrix represented as a heatmap.

Results and Discussion

First, relevant results pertaining to the independent general clustering methods are discussed, where all building types are tagged using their respective labels. A dendrogram is a type of tree diagram that shows relationships between similar sets of datapoints (Schonlau 2002). These structures are quite frequently used in biology to show clustering between genes and can be used to represent any type of grouped data. In this study, the well-formed dendrograms are represented by the following features and seasons: *Air system simulation cycle on-off status* (winter), *pump outlet temperature* (both seasons), and *water heater use side outlet temperature* (both seasons) (Figure 3). Majority of the office buildings across the selected climate zones form a unique cluster that differentiates these against other archetypes such as strip malls and apartment buildings.

Next, the homogeneity metric (Figure 4) is calculated by using 6 clusters. If the generated label matches the ground truth, the clustering result satisfies homogeneity, and higher the score the better the result. Comparing the results generated by using agglomerative and k-medoids, it appears that the k-medoids method identified more no. of prominent features being representative of the building archetype as compared to the agglomerative method, namely *pump outlet temperature*, *water heater use-side outlet temperature*, and *fan electric power* during the summer months. During the

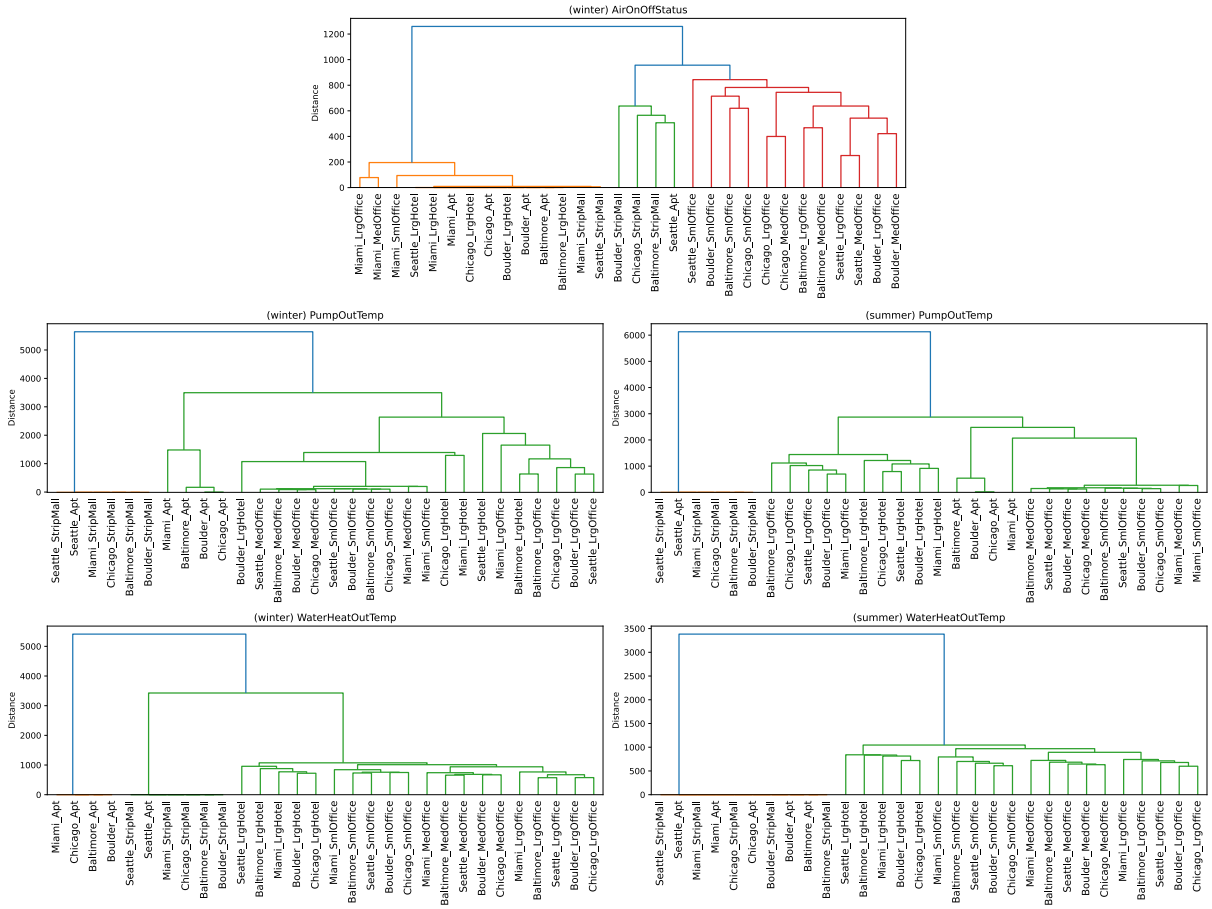

Figure 3 Dendrograms showing clusters formed from some building operational features.

winter months, only the *water heater use-side outlet temperature* scored high on the homogeneity metric.

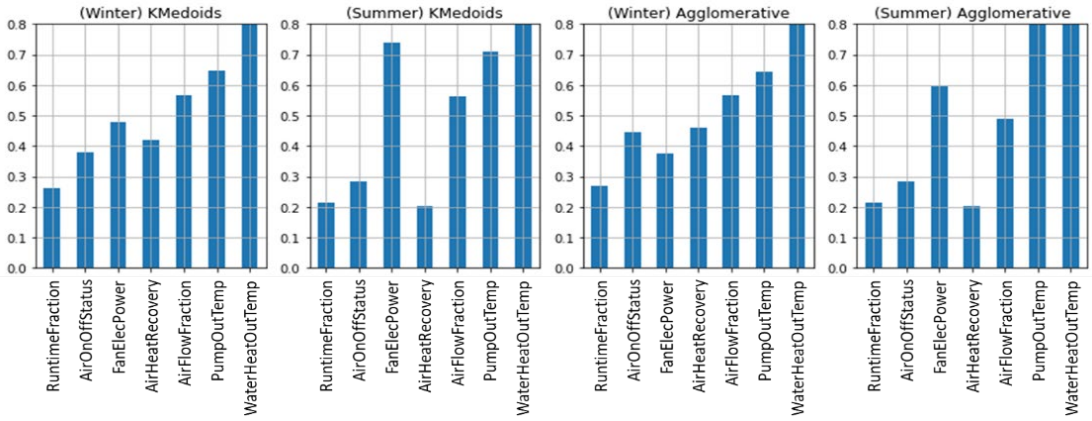

Figure 4 Homogeneity metrics across features clustered using k-medoids and agglomerative clustering methods (general).

Finally, the homogeneity scores are calculated again using the two clustering methods (k-medoids and agglomerative) but this time, the medium office building is isolated from the other building archetypes (Figure 5). This requires overwriting the labels for each building individually and creating a condition where the labels are either part of the expected class (medium office) or lie outside of it. The k-medoids method results in an observation that regardless of the season, the feature *water heater use side outlet temperature* is a prominent one for medium office buildings, meaning it could potentially be used for the identification of a medium office building when a set of unknown buildings are studied using this feature.

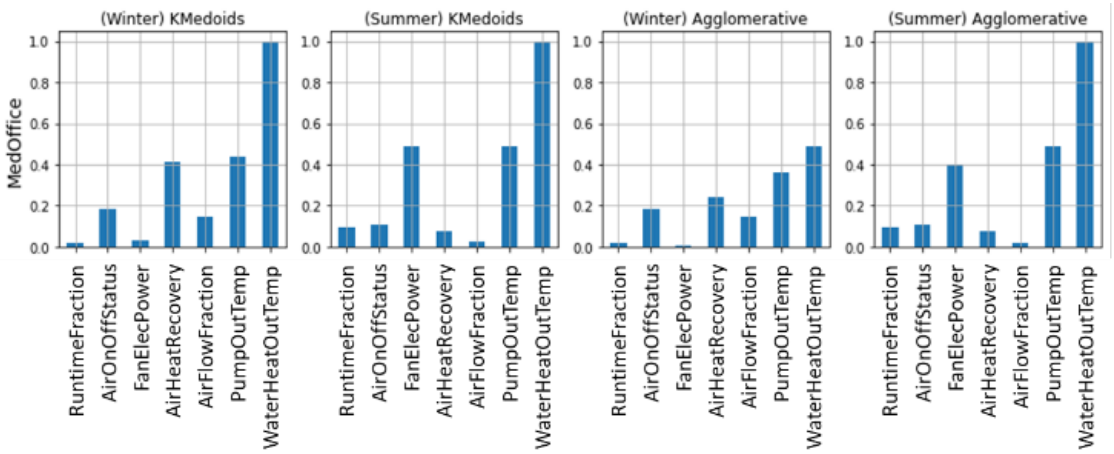

Figure 5 Homogeneity metrics across features clustered using k-medoids and agglomerative clustering methods (medium office isolated).

Conclusion and Future Work

In this paper, a clustering study was performed on seven time-series based building operational features. This study used a distance matrix where the distances between the different time-series data were calculated using DTW as the distance measure. Although this study used time-series corresponding to a month of operational data, it could be an interesting task to see the difference in analyzing weekly data instead. Additionally, the dataset used in this study is purely synthetic, generated using EnergyPlus. Inclusion of other building archetypes and climate zones could be added to the future scope of this study.

This study was set out with the premise that there are common prominent operational features across building archetypes and climate zones. These features could potentially be used to associate a building under study to a fully understood building archetype and further use a building energy prediction model which is fine-tuned to the archetype. The proposed methodology allows researchers to understand and recognize how a building subsystem trend can be used towards the identification of the building itself, purely from a data-driven perspective.

Although using DTW requires a greater computational effort than other distance measures (such as Euclidean), restricting the time window and the overall study period

(two months instead of annual) could help with expediting the overall process. Given that these intermediate data streams are readily available or can be used to derive a suitable proxy (coil runtime fraction derived from valve opening percentage) from a building BMS, constructing building energy prediction models becomes a more convenient task if a building can be bucketed within a larger group.

Acknowledgements

This publication has emanated from research conducted with the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC) under the 'ReBuild Initiative'.

References

Bagnall, Anthony, Aaron Bostrom, James Large, and Jason Lines. 2016. "The Great Time Series Classification Bake Off: An Experimental Evaluation of Recently Proposed Algorithms. Extended Version." *ArXiv:1602.01711 [Cs]*, February.

Bharathan Balaji, Yuvraj Agarwal, Mario Berges, David Culler, Rajesh Gupta, Mikkel Baun Kjærgaard, Mani Srivastava, Kamin Whitehouse, Arka Bhattacharya, Gabriel Fierro, Jingkun Gao, Joshua Gluck, Dezhi Hong, Aslak Johansen, Jason Koh, and Joern Ploennigs. 2016. Brick: Towards a Unified Metadata Model for Buildings. Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments - BuildSys '16 (2016), 41–50.

Bode, Gerrit, Thomas Schreiber, Marc Baranski, and Dirk Müller. 2019. "A Time Series Clustering Approach for Building Automation and Control Systems." *Applied Energy* 238 (March): 1337–45.

C. L. Stimmel, Big data analytics strategies for the smart grid. Boca Raton, FL: CRC Press/Taylor & Francis Group, 2015.

Dan Praire, John Petze, Marc Petock, and Mathan Kumar Gopan Samy. 2016. Project-Haystack. Technical Report. Continental Automated Building Association.

Gaitani, N., C. Lehmann, M. Santamouris, G. Mihalakakou, and P. Patargias. 2010. "Using Principal Component and Cluster Analysis in the Heating Evaluation of the School Building Sector." *Applied Energy* 87 (6): 2079–86.

Gallaher, Michael P., Alan C. O'Connor, John L. Dettbarn, Jr., and Linda T. Gilday. 2004. "Cost Analysis of Inadequate Interoperability in the U.S. Capital Facilities Industry." NIST GCR 04-867. National Institute of Standards and Technology.

Geler, Zoltan, Vladimir Kurbalija, Mirjana Ivanovic, Milos Radovanovic, and Weihui Dai. 2019. "Dynamic Time Warping: Itakura vs Sakoe-Chiba." In *2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*, 1–6. Sofia, Bulgaria: IEEE.

Giorgino, Toni. 2009. "Computing and Visualizing Dynamic Time Warping Alignments in *R*: The Dtw Package." *Journal of Statistical Software* 31 (7).

Hensel, Michael. *Design innovation for the built environment: Research by design and the renovation of practice*. Routledge, 2013.

Jradi, Muhyiddine, Niels Boel, Bo Eskerod Madsen, Jonas Jacobsen, Julie Strandesen Hooge, and Lars Kildelund. 2021. "BuildCOM: Automated Auditing and Continuous Commissioning of next Generation Building Management Systems." *Energy Informatics* 4 (1): 2.

Laski, Jonathan, and Victoria Burrows. 2017. "From Thousands to Billions." World Green Building Council.

Le Cam M, Zmeureanu R, Daoud A. Application of Data Mining techniques for energy modeling of HVAC sub-systems. in: Proceedings of eSim 2014 – IBPSA Canada Conference, IBPSA, Ottawa, Canada, 2014.

Lévy, Loup-Noé, Jérémie Bosom, Guillaume Guerard, Soufian Ben Amor, Marc Bui, and Hai Tran. 2021. "Application of Pretopological Hierarchical Clustering for Buildings Portfolio:" In *Proceedings of the 10th International Conference on Smart Cities and Green ICT Systems*, 228–35. Online Streaming, --- Select a Country ---: SCITEPRESS - Science and Technology Publications.

Lumpkin, Domenique R., W. Travis Horton, and Joseph V. Sinfield. 2020. "Holistic Synergy Analysis for Building Subsystem Performance and Innovation Opportunities." *Building and Environment* 178 (July): 106908.

Miller, Clayton, Zoltán Nagy, and Arno Schlueter. 2018. "A Review of Unsupervised Statistical Learning and Visual Analytics Techniques Applied to Performance Analysis of Non-Residential Buildings." *Renewable and Sustainable Energy Reviews* 81 (January): 1365–77.

Mueen, Abdullah, J. Keogh, Eamonn. Extracting Optimal Performance from Dynamic Time Warping. KDD 2016: 2129-2130.

Park, June Young, Xiya Yang, Clayton Miller, Pandarasamy Arjunan, and Zoltan Nagy. 2019. "Apples or Oranges? Identification of Fundamental Load Shape Profiles for Benchmarking Buildings Using a Large and Diverse Dataset." *Applied Energy* 236 (February): 1280–95.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2018. "Scikit-Learn: Machine Learning in Python." *ArXiv:1201.0490 [Cs]*, June.

Petcharat, Siriwarin, Supachart Chungpaibulpatana, and Pattana Rakkwamsuk. 2012. "Assessment of Potential

Energy Saving Using Cluster Analysis: A Case Study of Lighting Systems in Buildings." *Energy and Buildings* 52 (September): 145–52.

Saranya, C, and G Manikandan. 2013. "A Study on Normalization Techniques for Privacy Preserving Data Mining." *International Journal of Engineering and Technology* 5 (3): 4.

Shi, Zixiao, Guy R. Newsham, Long Chen, and H. Burak Gunay. 2019. "Evaluation of Clustering and Time Series Features for Point Type Inference in Smart Building Retrofit." In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 111–20. New York NY USA: ACM.

Silva, Diego F., and Gustavo E. A. P. A. Batista. 2016. "Speeding Up All-Pairwise Dynamic Time Warping Matrix Calculation." In *Proceedings of the 2016 SIAM International Conference on Data Mining*, 837–45. Society for Industrial and Applied Mathematics.

Stamatescu, Grigore, Iulia Stamatescu, Nicoleta Arghira, and Ioana Fagarasan. 2019. "Data-Driven Modelling of Smart Building Ventilation Subsystem." *Journal of Sensors* 2019 (March): 1–14.

Tormene, Paolo, Toni Giorgino, Silvana Quaglini, and Mario Stefanelli. 2009. "Matching Incomplete Time Series with Dynamic Time Warping: An Algorithm and an Application to Post-Stroke Rehabilitation." *Artificial Intelligence in Medicine* 45 (1): 11–34.

Yoshida, Keigo, Minoru Inui, Takehisa Yairi, Kazuo Machida, Masaki Shioya, and Yoshio Masukawa. 2008. "Identification of Causal Variables for Building Energy Fault Detection by Semi-Supervised LDA and Decision Boundary Analysis." In *2008 IEEE International Conference on Data Mining Workshops*, 164–73. Pisa, Italy: IEEE.