Energy and Buildings 156 (2017) 360–373

Contents lists available at ScienceDirect

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings

Clayton Miller[a,b,*], Forrest Meggers[c]

[a] Building and Urban Data Science (BUDS) Group, Department of Building, National University of Singapore, 117566 Singapore, Singapore
[b] Institute of Technology in Architecture (ITA), Architecture and Building Systems (A/S), ETH Zürich, 8093 Zürich, Switzerland
[c] Cooling and Heating for Architecturally Optimized Systems (CHAOS) Lab, Andlinger Center for Energy and Environment, Department of Architecture, Princeton University, Princeton, NJ 08544, USA

ARTICLE INFO

Article history:
Received 26 July 2017
Received in revised form
18 September 2017
Accepted 19 September 2017
Available online 4 October 2017

Keywords:
Data mining
Building performance
Performance classification
Energy efficiency
Smart meters

ABSTRACT

This study focuses on the inference of characteristic data from a data set of 507 non-residential buildings. A two-step framework is presented that extracts statistical, model-based, and pattern-based behavior. The goal of the framework is to reduce the expert intervention needed to utilize measured raw data in order to infer information such as building use type, performance class, and operational behavior. The first step is temporal feature extraction, which utilizes a library of data mining techniques to filter various phenomenon from the raw data. This step transforms quantitative raw data into qualitative categories that are presented in heat map visualizations for interpretation. In the second step, a random forest classification model is tested for accuracy in predicting primary space use, magnitude of energy consumption, and type of operational strategy using the generated features. The results show that predictions with these methods are 45.6% more accurate for primary building use type, 24.3% more accurate for performance class, and 63.6% more accurate for building operations type as compared to baselines.

1. Introduction

The built and urban environments have a significant impact on resource consumption and greenhouse gas emissions in the world. The United States is the world's second-largest energy consumer, and buildings there account for 41% of energy consumed.[1] The most extensive meta-analysis thus far of non-residential existing buildings showed a median opportunity of 16% energy savings potential by using cost-effective measures to remedy performance deficiencies [1]. Simply stated, roughly 6% of the energy consumed in the U.S. could be easily mitigated – a figure that would eventually grow to an annual energy savings potential of $30 billion and 340 megatons of CO_2 by the year 2030. Beyond saving energy, money and mitigating carbon, the impact of building performance improvement also extends to the health, comfort and satisfaction of the people who use buildings.

It is mysterious that these performance improvements are not rapidly being identified and implemented on a massive scale across the worlds building stock given the incentives and amount of research focused on building optimization in the fields of Architecture, Engineering and Computer Science. A comprehensive study of building performance analysis was completed by the California Commissioning Collaborative (CACx) to characterize the technology, market, and research landscape in the United States. Three of the key tasks in this project focused on establishing the state of the art [2], characterizing available tools and the barriers to adoption [3], and developing standard performance metrics [4]. These reports were accomplished through investigation of the available tools and technologies on the market as well as discussions and surveys with building operators and engineers. The common theme amongst the interviews and case studies was the *lack of time and expertise* on the part of the dedicated operations professionals. The findings showed that installation time and cost was driven by the need for an engineer to develop a full understanding of the building and systems. These barriers reduce the implementation of performance improvements.

From these studies, it becomes apparent that the biggest barrier to achieving performance improvement in buildings is scalability.

* Corresponding author at: 4 Architecture Drive, Department of Building, School of Design and Environment, National University of Singapore, 117566, Singapore.
E-mail address: clayton@nus.edu.sg (C. Miller).

[1] As of 2014, according to http://www.eia.gov/.

http://dx.doi.org/10.1016/j.enbuild.2017.09.056
0378-7788/

Architecture is a discipline founded with aesthetic creativity as a core tenet. Frank Lloyd Wright once stated, "The mother art is architecture. Without an architecture of our own, we have no soul of our civilization." Designers rightfully strive for artistic and meaningful creations; this phenomenon results in buildings with not only distinctive aesthetics but also unique energy systems design, installation practices and different levels of organization within the data-creating components. This paper shows that an emerging mass of data from the built environment can facilitate better characterization of buildings by through automation of meta-data extraction. These data are temporal sensor measurements from performance measurement systems.

1.1. Growth of raw temporal data sources in the built environment

As entities of analysis, buildings are less on the level of a typical mass-produced manufactured device in which each unit is the same in its components and functionality; and more on the level of customers of business, entities that are similar and yet have many nuances. Conventional mechanistic or model-based approaches, typically borrowed from manufacturing, have been the status quo in building performance research. As previously discussed, scalability among the heterogeneous building stock is a significant barrier to these approaches. More appropriate means of analysis lies in statistical learning techniques more often found in the medical, pharmaceutical and customer acquisition domains. These methods rely on extracting information and correlating patterns from large empirical data sets. *The strength of these techniques is in their robustness and automation of implementation – concepts explicitly necessary to meet the challenges outlined.*

This type of research on buildings would have been difficult even a few years ago. The creation and consolidation of measured sensor sources from the built environment and its occupants is occurring on an unprecedented scale. The Green Button Ecosystem now enables the easy extraction of performance data from over 60 million buildings.[2] Advanced metering infrastructure (AMI), or smart meters, have been installed on over 58.5 million buildings in the US alone.[3] A recent press release from the White House summarizes the impact of utilities and cities in unlocking these data [5]. It announces that 18 power utilities, serving more than 2.6 million customers, will provide detailed energy data by 2017. This study also suggests that such accessibility will enable improvement of energy performance in buildings by 20% by 2020. A vast majority of these raw data being generated are sub-hourly temporal data from meters and sensors.

1.2. Previous work

A significant amount of work has been undertaken in the field of building characterization using measured meter data. A comprehensive review of unsupervised learning techniques for various portfolio analysis and smart meter data was recently completed that includes much of the previous work in this area [6]. The key studies in the field of building characterization often deal with segmentation of large numbers of buildings, usually within the realm of smart meter analytics. Customer segmentation has been studied using various extracted temporal features from smart meter data for targeting programs [7–10]. Feature-based clustering of time-series performance data from building is another key field that precedes the current work. This field seeks to group various types of buildings or meters into similar clusters for analysis [11–18]. Various studies have looked at classification of building with various objectives using temporal meter data as a source of features [19–21,16,22]. Several other studies have extracted temporal features that enhance the ability to forecast consumption [23–25]. Several studies have analyzed larger than usual datasets from devices such as water heaters [26] and retrofit analysis at the city scale [27].

1.3. A framework for automated characterization of large numbers of non-residential buildings

This paper discusses a framework to investigate which characteristics of whole building electrical meter data are most indicative of various meta-data about buildings among large collections of commercial buildings. This structure is designed to *screen* electrical meter data for insight on the path towards deeper data analysis. The screening nature of the process is motivated by the scalability challenges previously outlined. An initial component of the methodology was a series of case study interviews and data collection processes to survey field data from numerous buildings around the world. A significant portion of this work was completed as part of a Ph.D. dissertation entitled "Screening Meter Data: Characterization of Temporal Energy Data from Large Groups of Non-Residential Buildings" [28].

The contributions of this study are related to its development and testing of a library of temporal machine learning features within the domain of non-residential buildings. To the author's best knowledge, no previous study has taken such a large number of buildings (507) and applied temporal feature engineering approaches from such a wide range of sources. Temporal features are extracted using techniques such as Seasonal Decomposition of Time Series by Loess (STL) and Symbolic Aggregate approXimation (SAX) using Vector Space Models (VSM) that have never been applied to electrical meter data from buildings. This study is also unique in that the objective is prediction of meta-data about buildings. This target is related to the contemporary challenge of large, raw temporal datasets from thousands of buildings with a significant amount of missing information; such is the case with large campuses, portfolios and utility-scale smart meter implementations.

2. Methodology

A two-step process is presented as a means of extracting knowledge from whole building electrical meters. Fig. 1 illustrates the intermediate steps in each of the phases. The first step is to create temporal features that produce quantitative data to describe various phenomenon occurring in the raw temporal data. This action is intended to transform the data into a more human-interpretable format and visualize the general patterns in the data. In this step, the data are extracted, cleaned, and processed with a library of temporal feature extraction techniques to differentiate various types of behavior. These features are visualized using an aggregate heat map format that can be evaluated according to expert intuition, comparison with design intent metrics, or with outliers detection.

The second step is focused on the characterization of buildings using the temporal features according to several objectives. This process allows an analyst to understand the impact each feature has upon the discrimination of each objective. Five test objectives are implemented in this study: principal building use, performance class, and operations strategy. One of the key outputs of this supervised learning process is the detection and discussion of what input features are *most important* in predicting the various classes. This approach gives exploratory insight into what features are important in determining various characteristics of a particular building

[2] According to http://www.greenbuttondata.org/.
[3] As of 2014, according to http://www.eia.gov/tools/faqs/faq.cfm?id=108&t=3.

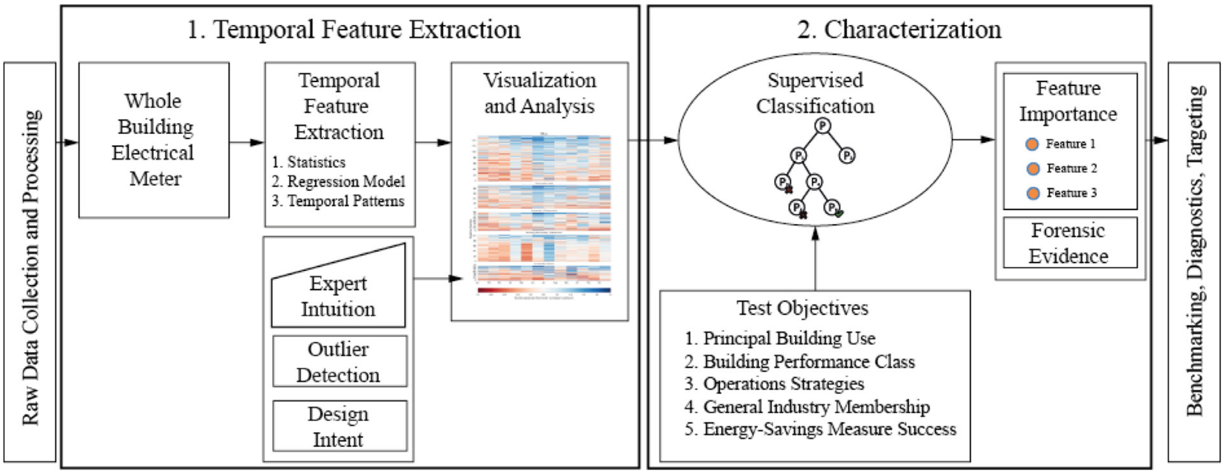

Fig. 1. Overview of data analytics framework.

amongst a large set of its peers. These metadata are building blocks for many other techniques such as benchmarking, diagnostics and targeting. The motivation for choosing these particular objectives centers around the consistently available meta-data from the collected case study data and their relation to various other techniques in the building performance analysis domain.

2.1. Case study buildings and collected data

An open data set of 507 whole building electrical meters were utilized in this study for implementation of the two-step process. These buildings are from university campuses from around the world. The origin and development of these data are found in Miller and Meggers [29]. A broad range of descriptive statistics and meta-data explanation are available in the previous literature.

2.2. Temporal feature extraction

Feature extraction is an essential process of machine learning and is the means by which objects are described quantitatively in a way that algorithms can differentiate between different types or classes. Much of these data are needed when creating an energy simulation model, when setting thresholds for automated fault detection and diagnostics, or benchmarking a building. When performing analysis on a single building, this meta-data might be easy to accumulate. However, when such a process is scaled across hundreds or potentially thousands of buildings, a collection of these data is not a trivial procedure.

The goal of temporal feature extraction and analysis is to use various techniques to convert all these *qualitative* terms into a *quantitative* domain. For example, the descriptor *weather-dependency* can be quantified through the utilization of the Spearman rank order correlation coefficient with outdoor air temperature. Consistency or volatility of daily, weekly, or annual behavior can be quantified using various pattern recognition techniques. The primary focus of this study is to create and apply some temporal feature extraction techniques on commercial buildings for characterization. Fig. 2 illustrates a hierarchy of the conventional categories of temporal features and the new category of temporal features that include a few examples that are outlined in this study.

Temporal features are aggregations of the behavior exhibited in time-series data. They are characteristics that summarize sensor information in a way to inform an analyst through visualization or to use as training data in a predictive classification or regression model. Feature extraction is a step in the process of machine learning and is a form of dimensionality reduction of data. This process

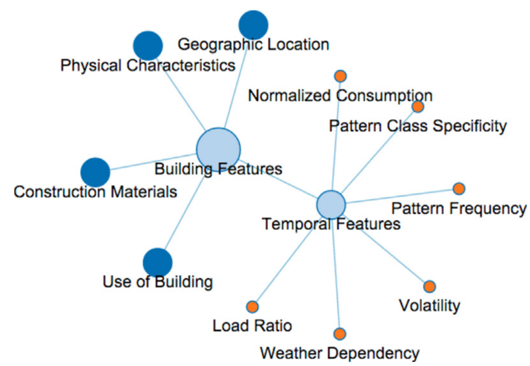

Fig. 2. Temporal features extracted solely from raw sensor data.

Table 1
Overview of feature categories.

Feature category	General description
Statistics-based	Aggregations of time series data using mean, median, max, min, standard deviation
Regression model-based	Development of a predictive model using training data and using model parameters and outputs to describe the data
Pattern-based	Extraction of frequent and useful daily, weekly, monthly, or long-term patterns

seeks to quantify various qualitative behaviors. This section provides an overview of the categories of temporal features extracted from the case study building data, the methods used to implement them, and visualized examples of a selected subset of features manifest themselves over a time range. Table 1 gives an overview the temporal features outlined in this section.

2.3. Characterization and prediction of meta-data

The primary goal of this research is to get a better sense of what behavior in time-series sensor data is most characteristic of various *types* of buildings. As mentioned in the introduction, if this meta-data can be discriminated, the process of characterizing a building can be automated. In this section, the operation of using random forest classification models and the input variable importance feature. An overview of this process is found in Fig. 3.

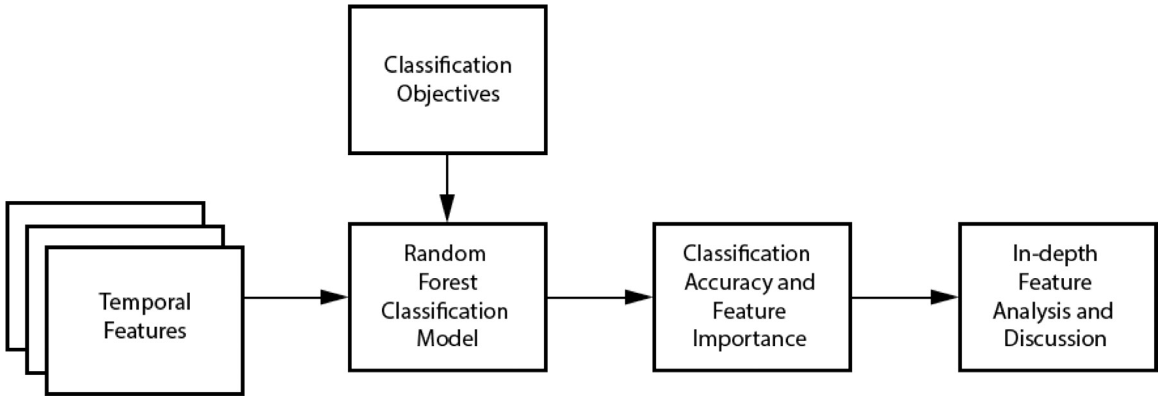

Fig. 3. Characterization process to investigate the ability for various features to describe the classification objectives.

For each objective, two steps are taken to predict each objective and then to investigate the influence of the input features on class differentiation. In the first step, a random forest classification model is built using subsets of the generated features to predict the objectives class. In the second phase, the classification model indicates the ability of the temporal features in describing the class based on its accuracy.

Random forest classification models were chosen based on their capacity to model diverse and large data sets in a robust way [30]. These models use an ensemble of decision trees to predict various characteristic labels about each building based on its features. The literature describes decision trees as the "closest to meeting the requirements for serving as an off-the-shelf procedure for data mining" [31]. Decision trees often over-fit data due to high variance. Random forest models work by creating a set of decision trees and averaging all of their predictions to overcome this variance.

Random forests use a form of cross-validation by training and testing each tree using a different bootstrapped sample from the data. This process produces an *out-of-bag error (OOB)* that acts as a generalized error for understanding how well each class can be predicted. This accuracy is used to determine how well the generated temporal features can delineate the class objectives. Random forests can also calculate the importance of the input features and how well they lend themselves to predicting the targets. This attribute is useful in that it allows us to understand exactly which temporal features are most characteristic of various objectives. Variable importance is calculated using Eq. (1). The importance of input feature X_m for predicting Y by adding up the weighted impurity decreases $p(t)\Delta i(s_t, t)$ for all nodes t where X_m is used, averaged over all N_T trees in the forest [32].

$$Imp(X_m) = \frac{1}{N_T}\sum_{T}\sum_{t\in T:v(s_t)=X_m} p(t)\Delta i(s_t, t) \quad (1)$$

3. Statistics-based features

Statistics-based temporal features are the first and most simplified category of temporal features developed. The main classes of features are essential temporal statistics, ratio-based, and the Spearman rank order correlation coefficient.

3.1. Basic statistics

The first set of temporal features to be extracted are basic statistics-based metrics that utilize the time-series data vector for various time ranges to obtain information using mean, median, maximum, minimum, range, variance, and standard deviation. Many of these features are developed through the implementation of the VISDOM package in the R programming language [33]. As a simple example, if a time-series vector is described as X, with N values of $X = x_1, x_2, \ldots, x_n$, the most common statistical metric, mean (or μ), can be calculated using Eq. (2) [34].

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad (2)$$

The mean is taken not just for the entire time series, but also from the summer and winter seasons. The variance of the values is taken for the whole year, the summer and winter seasons as well. The variance of daily mean, minimum, and maximum values are determined to understand the breadth of values across the time range. Variance is calculated according to Eq. (3).

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} \quad (3)$$

The maximum and minimum electrical demand are calculated. Additionally, the hour and date at which the maximum demand occurs are determined to understand when peak consumption occurs. The 97th and 3rd percentiles are calculated to exclude any extreme outliers, a value that's often more useful than the maximum and minimum.

A series of hour-of-day (HOD) metrics are calculated that relate to aggregating the behavior occurring at each of hour of the day. The first of these calculate the most current hour of the top demand of the 10% hottest days and the most common hour of the top 10% temperatures to inform roughly about cooling energy consumption. These metrics are repeated from the bottom 10% coldest days and temperatures. Another set of twenty-four parameters is calculated to account directly for the mean demand of each hour of the day.

3.2. Ratio-based statistical features

The second major category of statistical features is ratio-based features. Simply, these are metrics in which two or more of the previously calculated statistical parameters are combined as a ratio. These features often have a *normalizing effect* in which buildings can be more appropriately compared to each other. The first extracted metric of this type is one of the most commonly calculated for building performance analysis: the consumption magnitude of electricity normalized by the floor area of the building. This metric seeks to provide a basis for comparison between buildings and is used as a key metric within numerous benchmarking and performance analysis techniques.

3.3. Spearman rank order correlation coefficient

Another useful metric to calculate is related to how much influence outside air temperature has on the consumption of a building. Miller et al. describes a process of utilizing the Spearman Rank Order Correlation (ROC) coefficient to approximate the correlation between outside conditions and the electrical consumption [35]. The ROC essentially ranks the items in two different lists the ratio quantifies whether those lists are correlated positively or negatively. In this case, the two variables are outside air temperature and electrical consumption. The coefficient range is *-1* (highly negatively correlated) and +1 (highly positively correlated). If the correlation is positive, the ROC is positive and closer to *+1* and the electrical consumption is cooling sensitive as the consumption goes up with higher temperature. If the correlation is negative, the ROC is negative and closer to −1, and the time range is heating sensitive as the consumption goes up with lower temperature.

The correlation coefficient can be visualized for a single case as seen in Fig. 4. The factor, in this instance, is calculated individually for each month. This process results in twelve calculations of the metric using between 29 and 31 samples. In this case, consumption in January to May is noticeably more heating sensitive, a fact that can be observed clearly from the line chart, as well as the one dimension heat map. May to November is more cooling sensitive. It is interesting that September appears to be the most cooling sensitive month, a fact perhaps related to using schedules during that month. This coefficient is not a perfect indicator of HVAC consumption; it just detects a correlation. However, it is fast and easy to calculate and is the first phase of detecting weather dependency.

3.4. Implementation of stats-based features

Fig. 5 illustrates the same normalized consumption metric as applied to all of the case study buildings for three examples of the screening parameters: area-normalized consumption, ratio-based daily load max vs. min, and monthly Spearman rank order correlation coefficient. There are five segments of buildings based on the primary use types within the set: offices, university laboratories, college classrooms, primary/secondary schools, and university dormitories. These metrics are visualized in this way to understand the difference between each of these use types for each of the presented metrics. Each row of the heatmap for each segment is the values of the feature for a single building, while the *x*-axis is the time range for all buildings. Not all of the case study buildings have a January to December time range. For these cases, the data was rearranged so that a continuous set of January to December data is available to be visualized in the heat map. The aggregation metrics themselves are not calculated with this rearranged vector; it is only for visualization purposes.

4. Regression model-based features

Semi-physical behavior about a building can be extracted by using performance prediction models and using output parameters and goodness-of-fit metrics for characterization. This section covers the use of several common electrical consumption prediction models to create sets of temporal features useful for characterization of buildings.

Prediction of electrical loads based on their shape and trends over time is a mature field developed to forecast consumption to detect anomalies and analyze the impact of demand response and efficiency measures. The most common technique in this category is the use of heating and cooling degree days to normalize monthly consumption [36]. Over the years, various other methods have been developed using techniques such as neural networks, ARIMA models, and more complex regression [37]. However, simplified methods have retained their usefulness over time due to ease of implementation and accuracy. In the context of temporal feature creation, a regression model provides various metrics that describe how well a meter conforms to conventional assumptions. For example, if actual measurements and predicted consumption match well, the underlying behavior of energy-consuming systems in the building has been captured adequately. If not, there is the uncharacteristic phenomenon that will need to be obtained with a different type of model or feature.

4.1. Load shape regression-based features

A contemporary, simplified load prediction technique is selected to create temporal features that capture whether the electrical measurement is simply a function of time-of-week scheduling. This model was developed by Matthieu et al. and Price and implemented mostly in the context of electrical demand response evaluation [38,39]. The premise of the model is based on two features: a time-of-week indicator and an outdoor air temperature dependence. This model is also known as the *Time-of-week and Temperature or (TOWT)* model or *LBNL regression model* and is implemented in the *eetd-loadshape* library developed by Lawrence Berkeley National Laboratory.[4]

According to the literature, the model operates as follows [38]. The time of week indicator is created by dividing each week into a set of intervals corresponding to each hour of the week. For example, the first interval is Sunday at 01:00, the second is Sunday at 02:00, and so on. The last, or 168th, interval is Saturday at 23:00. A different regression coefficient, α_i, is calculated for each interval in addition to temperature dependence. The model uses outdoor air temperature dependence to divide the intervals into two categories: one for occupied hours and one for unoccupied. These modes are not necessarily indicators of exactly when people are inhabiting the building, but merely an empirical indication of when occupancy-related systems are detected to be operating. Separate piecewise-continuous temperature dependencies are then calculated for each type of mode. The outdoor air temperature is divided into six equally sized temperature intervals. A temperature parameter, β_j, with $j = 1\ldots6$, is assigned to each interval. Within the model, the outdoor air temperature at time, t, occurring at time-of-week, i, (designated as $T(t_i)$) is divided into six component temperatures, $T_{c,j}(t_i)$. Each of these temperatures is multiplied by β_j and then summed to determine the temperature-dependent load. For occupied periods the building load, L_o, is calculated by Eq. (4).

$$L_0(t_i, T(t_i)) = \alpha_i + \sum_{j=1}^{6} \beta_j T_{c,j}(t_i) \tag{4}$$

Prediction of unoccupied mode occurs using a single temperature parameter, β_u. Unoccupied load, L_u, is calculated with Eq. (5).

$$L_0(t_i, T(t_i)) = \alpha_i + \beta_u T_{c,j}(t_i) \tag{5}$$

The primary means of temporal feature creation from this process is through the analysis of model fit. The first metric calculated is a normalized, hourly residual, R, that can be used to visualize deviations from the model. It is calculated from the actual load, L_a, and the predicted load, L_p. The residual at a particular hour, t, is calculated using Eq. (6).

$$R_t = \frac{L_{t,a} - L_{t,p}}{\max_{L_a}} \tag{6}$$

[4] https://bitbucket.org/berkeleylab/eetd-loadshape.

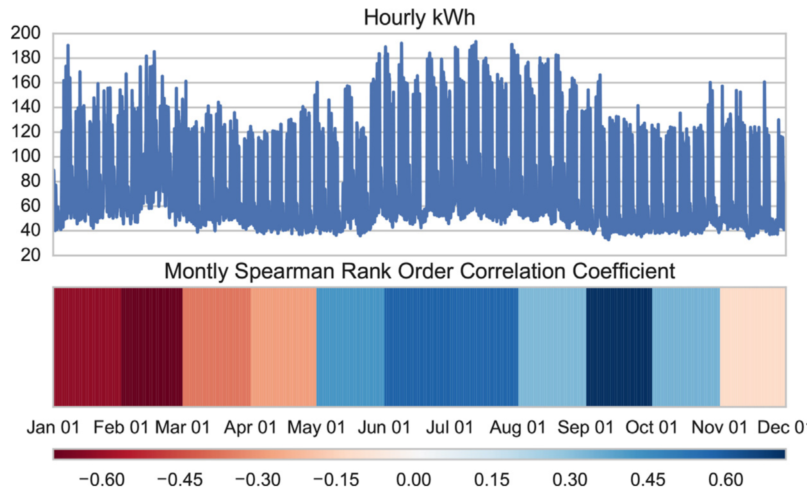

Fig. 4. Single building example of the spearman rank order correlation coefficent with weather.

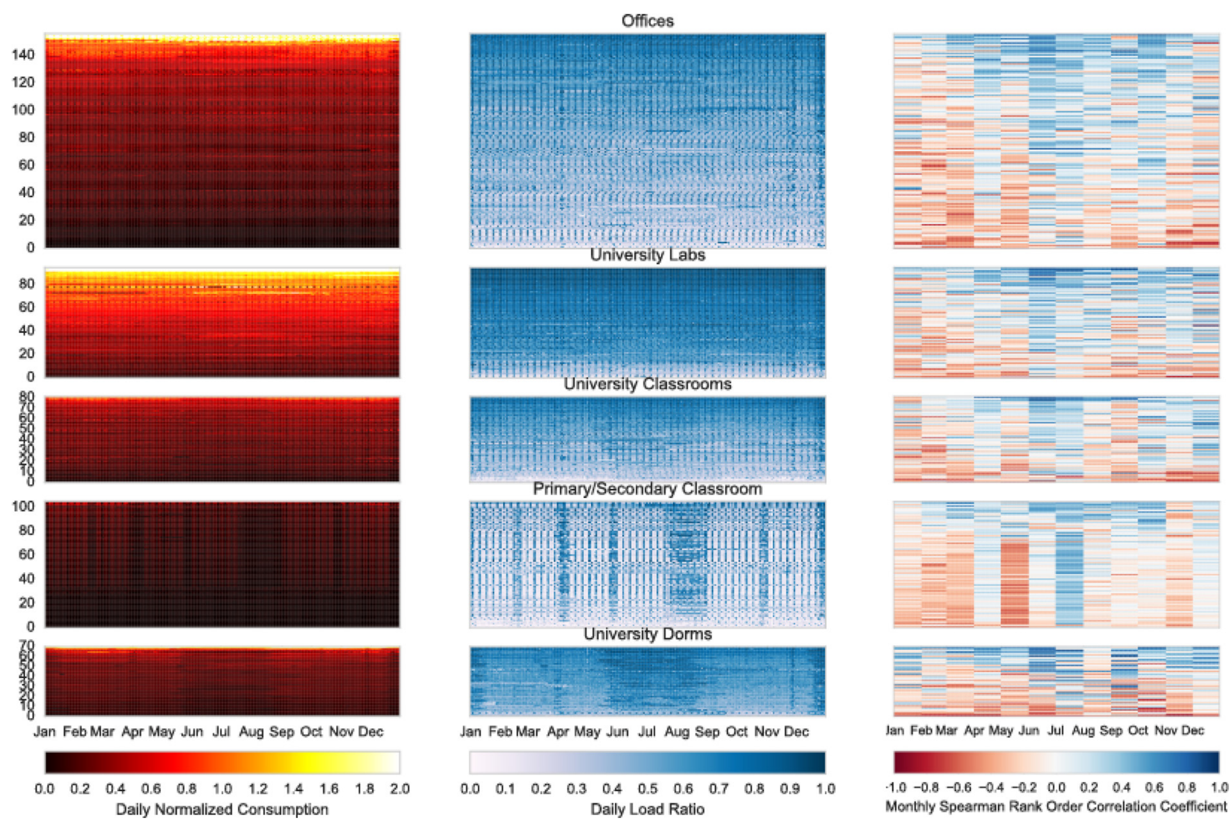

Fig. 5. Heat map of a selection of statistics-based temporal features: area-normalized consumption (left), ratio-based daily load max vs. min (center), and monthly spearman rank order correlation coefficient (right).

An example of the TOWT model implemented on one of the case study buildings is seen in Fig. 6. Two primary characteristics are captured from a model residual analysis. The first is the building's deviation from a set time-of-week schedule and behavior causing the model to highly over-predict. These deviations are most often attributed to public holidays, breaks in normal operation, or changes in normal operating modes. In the single building study, one of the most visible daily deviations, Christmas Day, is observed. This day is significantly over-predicted due to the model not being informed of the Christmas Day holiday. The automated capture of this phenomenon can report whether the building is of a certain use-type or in an individual jurisdiction. The second characteristics obtained are periods of underprediction when the building is consuming more electricity than expected. These data inform whether a building is being consistently utilized, or whether there is volatility in its normal operating schedule from week-to-week.

4.2. Change point model regression

Another means of performance modeling that considers weather characterization is the use of linear change point models. The outputs of these models can be interpretable in approximating the amount of energy being used for heating, ventilation, and air-conditioning (HVAC). This type of model has its basis in the

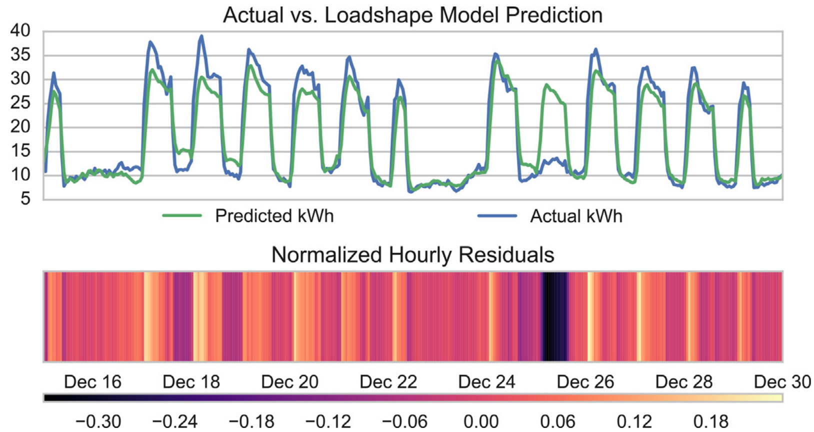

Fig. 6. Single building example of TWOT model with hourly normalized residuals.

previously-mentioned PRISM method and has been continuously utilized, recently by [40]. This multivariate, piece-wise regression model is developed using daily consumption and outdoor air dry-bulb temperature information. A linear regression model is fitted to data detected to be correlated with outdoor dry-bulb air temperature, either positively for cooling energy consumption or negatively for heating energy consumption. For example, as the outdoor air temperature climbs above a certain point, the relationship between electricity consumption and every degree increase in temperature should be a straight line with a certain slope if the building has an electrically-driven cooling system. The point at which this change occurs is considered the cooling balance point of the building, and the slope of the line is the rate of cooling energy increase due to outdoor air conditions.

Eqs. (7) and (8) are used to predict energy consumption based on an outdoor air temperature, T. This equation can also predict the heating ($\beta_2(T-\beta_3)$) or cooling ($\beta_2(\beta_3-T)$) components of the electrical consumption to a certain level of accuracy.

$$E_c = \beta_1 + \beta_2(T - \beta_3) \tag{7}$$

$$E_h = \beta_1 + \beta_2(\beta_3 - T) \tag{8}$$

4.3. Seasonality and trend decomposition

Temporal data from different sources often exhibit similar types of behavior that are studied within the field of forecasting and temporal data mining. Electrical building meter data fits into this category, and the same feature extraction techniques can be applied as what is commonly done for financial or social science analysis. These techniques often seek to decompose time-series data into several components that represent the underlying nature of the data [34]. For example, the electrical meter data collected from buildings is often cyclical in its weekly schedule. People are utilizing buildings each day of the week in a relatively predictable pattern. A prevalent example of this behavior is found in office buildings where occupants are typical white-collar professionals who come into work on weekdays at a particular time and leave to go home at a certain time. Weekends are unoccupied periods in which there is little to no activity. This behavior is an example of what's known as seasonality within time series analysis. Seasonality is a fixed and known period of consistent modulation and is a feature that is often extracted before creating predictive models.

Trends are another feature commonly found in temporal data. A trend is a long-term increase or decrease in the data that often doesn't follow a particular pattern. Trends are commonly due to factors that are less systematic than seasonality and are often due to external influences. For building energy consumption, trends manifest themselves as gradual shifts in consumption over the course of weeks or months. Often these variations are due to weather-related factors influencing the HVAC equipment. Other causes of trends are changes in occupancy of degradation of system efficiency.

To capture these features to understand their impact on characterizing buildings, the seasonal-trend decomposition procedure based on loess is used to extract each of these features from the case study buildings [41]. This process is used to remove the weekly *seasonal* patterns from each building, the long-term trend over time, and the residual remainders from the model developed by those two components. The input data is aggregated to daily summations and weather-normalized by subtracting the calculated heating and cooling elements from the change point model described in Section 4.2. This step is done to reduce the influence weather plays in the trend decomposition. The *STL* package in R is used for this process to extract the seasonal, trend, and irregular components.[5]

The details of the internal algorithms of the *STL* procedure are described by [41]. The process uses an inner loop of algorithms to detrend and deseasonalize the data by creating a trend component, T_v, and a seasonal component, S_v. The remainder component, R_v, is a subtraction of the input values, Y_v as seen in Eq. (9).

$$R_v = Y_v - T_v - S_v \tag{9}$$

4.4. Implementation of model-based features

Fig. 7 illustrates an overview of an implementation of three examples of model-based features on all the buildings across the various building use types in the study. The heat map at the far left illustrates normalized residuals from the load shape regression model. The differences between each use type can be noticed from a high level due to the nature of residuals. The darker areas of the visualization indicate when the model is highly over-predicting consumption and lighter areas indicate when the model is under-predicting. Typical holiday periods such as spring, summer and winter breaks and holidays such as the American Labor Day and Thanksgiving are seen as darker areas. Offices, labs and classrooms seem to have similar residual patterns, likely due to their scheduling being similar. Slight fundamental differences are seen such as the fact that classrooms have more general areas of over-prediction, likely due to less consistent occupancy. Primary/Secondary schools and dormitories are less predictable on an annual basis due to their strong seasonal patterns of use; this fact is intuitive, and model residuals of this type are accurate in automatically characterizing this behavior. The center figure illustrates heating energy regres-

[5] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/stl.html.

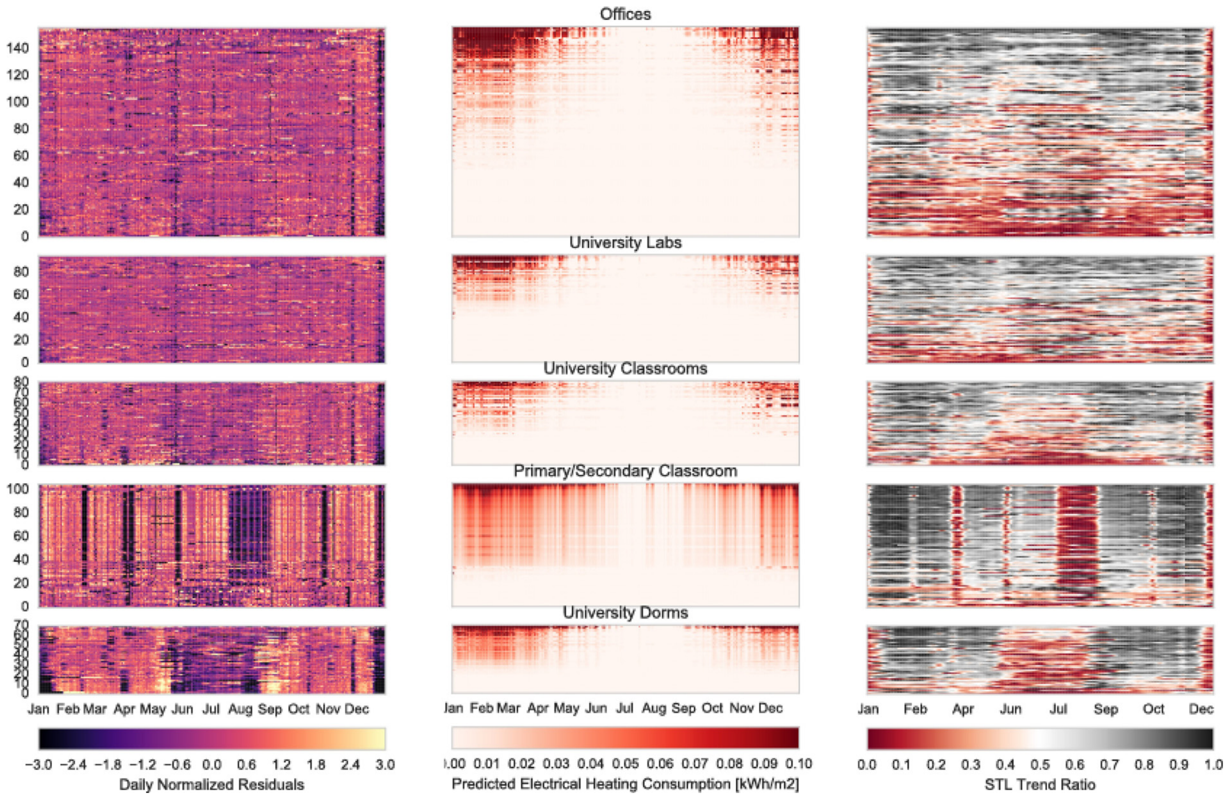

Fig. 7. Heat map of a selection of model-based temporal features: daily normalized residuals from load shape regression models (left), heating energy prediction using change point model regression (center), and seasonal trends using the STL package (right).

sion for all case study buildings. These figures have been normalized according to floor area. Each building's response to outdoor air temperature is indicative of the type of systems installed in addition to the efficiency of energy conversion of those systems. The far right heat map illustrates the trend decomposition as applied to the entire case study set of buildings. Offices appear to have quite a bit of diversity over time, with a few observable systematic low spots in the spring and autumn periods at the bottom of the heat map. Laboratories reflect that behavior, while university visually has an opposite effect with less than the average trend in the summer months. Primary/Secondary school classrooms have a very distinct delineation between when school is in session and out of session during the summer and various breaks. As many of these schools are in the UK, their out-of-session periods appear to line up naturally. University dormitories also have clear delineations between occupied and unoccupied periods, and they also seem to match up quite well, despite the diversity of data sources of these buildings.

5. Pattern-based features

The third category of temporal features that is developed in this study is related to capturing the typical and atypical patterns of use from building performance data. The goal of these features is to quantify whether a building has consistency on a daily or weekly basis, whether certain building types have certain types of patterns of use and to inform how they can be used to predict various kinds of meta-data. In temporal feature mining, two concepts are relevant in this analysis: motifs and discords. A motif is a typical pattern that occurs on a regular basis within a data set [42]. A discord is an unusual pattern within a data set that identifies infrequent behavior [43]. Several of the temporal features developed in this process is designed to leverage these concepts. The pattern-based feature categories outlined in this section include diurnal pattern extraction, pattern specificity, and long-term consistency.

5.1. Dirunal pattern extraction

The first temporal feature outlined is based on the *DayFilter* process which extracts the motifs and discords from raw meter data based on 24 h periods [44]. This process heavily utilizes the Symbolic Aggregate approXimation (SAX) representation of time-series data [45]. SAX is a process of time-series data discretization that converts temporal data into the string data type. This process empowers various text mining and visualization techniques. The primary feature extracted from this process for this study is diurnal pattern frequency which quantifies the number and size of motifs found from a particular meter.

5.2. Pattern specificity

Another method to leverage SAX to characterize the case study data is to use it to extract which patterns are most indicative of a particular building use type. This information is obtained using the SAX-VSM process pioneered by Senin and Malinchik that uses SAX and Vector Space Model technique from the text mining field [46]. Conventionally this method is utilized as a classification model to predict which class a certain time-series belongs. A by-product of the process is that the subsequences of each data stream are assigned a metric indicating their specificity. Pattern specificity is a concept that quantifies how well a meter *fits within its class*. This technique is used to determine whether a building is operating similar to other supposed peer buildings of the same type.

5.3. Long-term pattern consistency

The concept of long-term consistency is related to how volatile a building's electrical consumption is over the course of a long-range period such as one year. A building that is considered more volatile will have significant shifts in steady-state operation over the process of a year. Often these changes are related to seasonality of scheduling that can be the case in buildings like schools and universities. A less volatile building will be more consistent in overall magnitude of consumption over the course of a year. This behavior is more often the case in offices and laboratories. In this analysis, a concept known as *breakout detection* is utilized to quantify the difference between these behaviors. A metric is created to detect the number of shifts in relative steady-state over the course of the time range. This metric was developed in a previous study focused on data from a single campus [35]. An R programming package, *BreakoutDetection*, is used to create this parameter. This package was developed by the social media company Twitter to process their time-series data.[6] The details of the algorithms utilized in this package can be found in a study by James et al. [47].

Fig. 8 illustrates the breakout detection process from a single building data stream. This dataset includes hourly data from an entire year from a university dormitory building. A minimum threshold of 30 days is chosen in this case, which explains the lack of threshold shift in April, a break that may be attributed to spring break for this building. Seven total steady-state variations were detected by the algorithm in this case, and many of them occur in the conventional university scheduled summer, spring, fall and winter breaks.

5.4. Implementation of pattern-based features

Fig. 9 shows three pattern-based features as applied to all the case study buildings. The far right heat map shows the pattern frequency metric as applied to all the case study buildings extracted from the *DayFilter* process. One will notice that there is a range of pattern frequencies occurring across each of the building use types. Offices and Primary/Secondary Classrooms seem to have larger regions of darker, more consistent behavior. Labs and Classrooms seem to be more volatile across the time ranges. The center heat map illustrates breakout detection across the building use types in this study. This implementation uses the same input parameter of a 30 day minimum between breakouts. One notices somewhat of consistency among offices, labs, and classrooms regarding the distribution of breakout numbers, while university dormitories and primary/secondary classrooms have a noticeably higher number of breakouts across the range of behavior. The far right heat map illustrates the daily specificity calculation process applied to all 507 case studies as divided among the use types. Clear differences in patterns across the time ranges are visible for each of the building use types. Offices, university laboratories, and university classrooms all seem to have similar phases of specificity at similar times of the year, while their breaks often differentiate dorms and primary/secondary schools.

6. Prediction of building use, performance class, and operational strategy

Visualization of temporal features on their own is a means of understanding the range of values of the various phenomenon across a time range. This situation gives an analyst the basis to begin understanding what discriminates a building based on different objectives. The next step is to utilize the features to predict whether a building falls into a particular category and test the importance of various elements in making that prediction. Understanding which features are most characteristic to a particular objective is the fundamental tenet of this study. In this section, three classification objectives are tested:

1. Principle building use – The primary use of the building is designated for the principal activity conducted by percentage of space designated for that activity. It is rare for a building to be devoted specifically to a single task, and mixed-use buildings pose a specific challenge to prediction.
2. Performance class – Each building is assigned to a particular performance class according to whether its area-normalized consumption in the bottom, middle, or top 33% percentiles within its principle building use-type class.
3. General operation strategy – Buildings that are controlled by the same entity, such as those on a University campus, often have similar schedules, operating parameters, and use patterns. This objective tests to understand how distinct these differences are between different campuses.

6.1. Principal building use

The first scenario investigated is the characterization of primary building use type. The goal of this effort is to quantify what temporal behavior *is most characteristic in a building being used for a certain purpose*. For example, what makes the electrical consumption patterns of an office building unique as compared to other purposes such as a convenience store, airport, or laboratory. This objective is necessary to understand who are the *peers* of a building. Whatever category a building is assigned determines what benchmark is used to determine the performance level of a building. The EnergyStar Portfolio Manager is the most common benchmarking platform in the United States and the first step in its evaluation is identifying the property type. There are 80 *property types* in portfolio manager and each one is devoted to a particular primary building use type. Twenty-one of those property types are available for submission to achieve a 1–100 ENERGYSTAR score in the United States.

Allocation of the primary use type of a building is often considered a trivial activity when analyzed from a smaller set of buildings. As the number of building being analyzed grows, so does the complexity of space use evaluation. The use of buildings changes over time and these changes are not always documented. In several of the case studies, this topic was discussed and highlighted as an issue concerning benchmarking a building.

Discriminatory features have already been visualized extensively in previous sections and the differences between the primary use types are apparent in the overview heat maps of each feature. Fig. 10 is the first such example of the output results of the classification model in predicting the building's primary use type using the temporal features created in this study. This visualization is a kind of error matrix, or confusion matrix, that illustrates the performance of a supervised classification algorithm. The y-axis represents the correct label of each classification input and the x-axis is the predicted label. An accurate classification would fall on the left-to-right diagonal of the grid. This grid is normalized according to the percentage of buildings within each class. The model was built using the scikit-learn Python library[7] with the number of estimators set to 100 and the minimum samples per leaf set to 2. The overall general accuracy of the model is 67.8% as compared to a baseline model of 22.2%. The baseline model using a strati-

[6] https://github.com/twitter/BreakoutDetection.

[7] http://scikit-learn.org/.

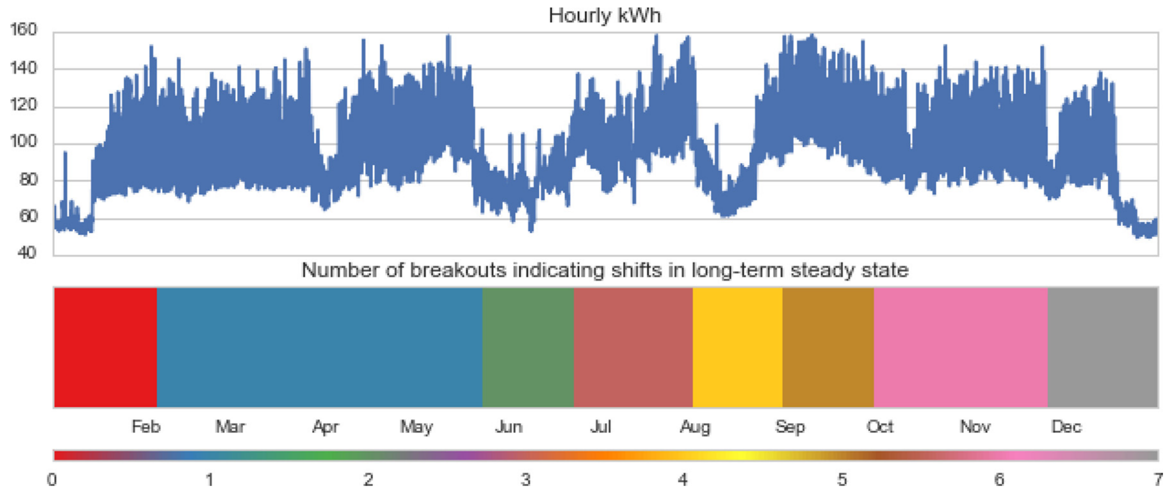

Fig. 8. Single building example of breakout detection to test for long-term volatility in an university dormitory building.

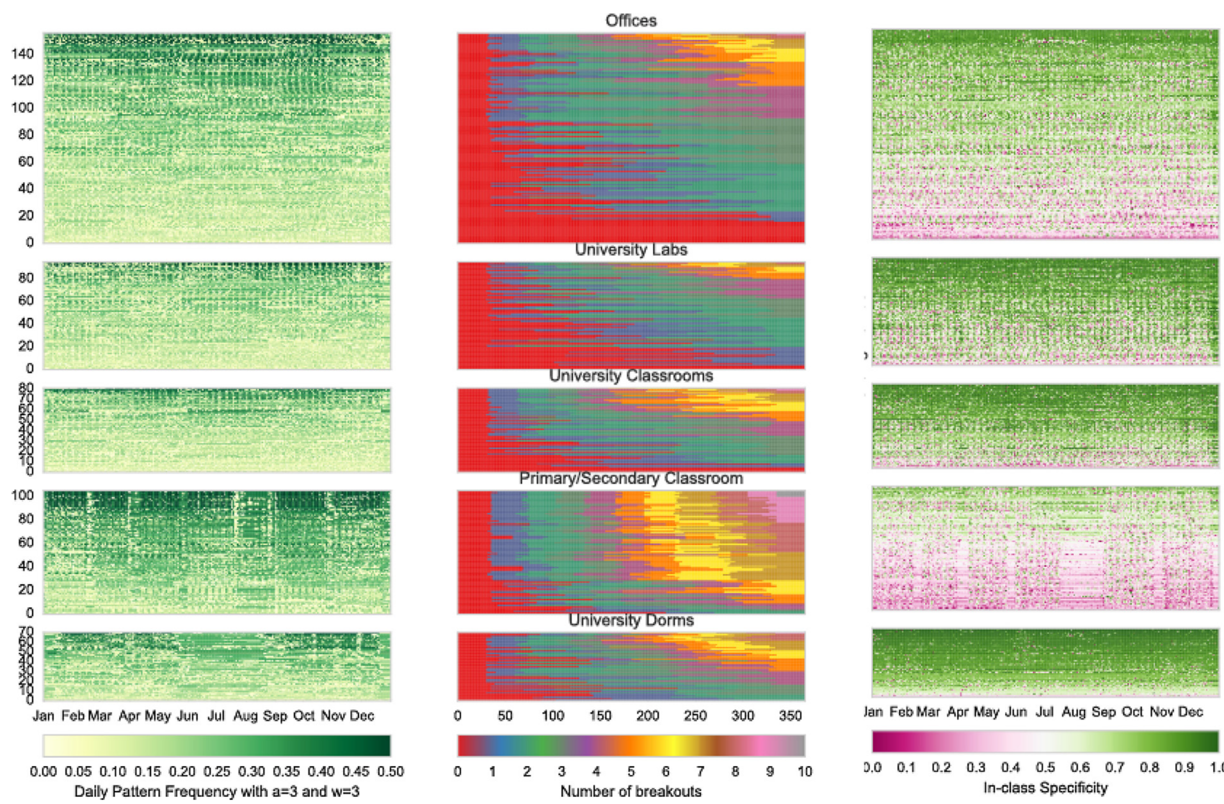

Fig. 9. Heat map of a selection of pattern-based temporal features: daily pattern frequency from the *DayFilter* process (left), breakout detection for long term volatility (center), and in-class specificity using the *SAX-VSM* process (right).

fied strategy in which categories are chosen randomly based on the percentage of each class occurring in the training set. Based on the analysis, university dormitories and primary/secondary classrooms are the best-characterized use types overall with a precision of 92% and 96% respectively and accuracies of 74% and 75%. The office category is easily confused with university classrooms and laboratories. This situation is not surprising as many of these facilities are quite similar and uses within these categories often overlap.

Previously, an example of how to characterize building use type was illustrated using a random forest model and various feature importance techniques. In this subsection, a discussion is presented of how this sort of characterization can be useful in a practical setting. In the case study interviews, the topic of benchmarking of buildings was discussed. One of the issues presented to the operations teams was the concept of not having a complete understanding of the way the buildings on their campus were being used. For example, several of the campuses have a spreadsheet outlines various metadata about the facilities on campus. This worksheet, in many cases, includes the *primary use type* of the building. It was found that this primary use type designation is often loosely based on information from when the building was constructed or through informal site survey. In other situations, the building has an accurate breakdown of all the sub-spaces in the building and approximately for what the spaces are being used. In these discussions, the idea was presented that building use type characterization could be used to determine automatically whether the

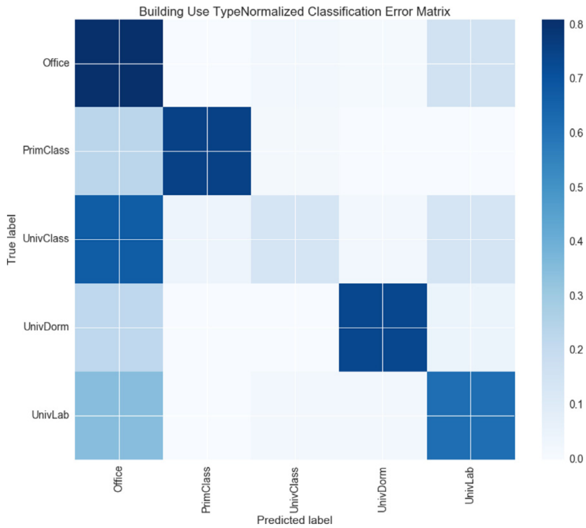

Fig. 10. Classification error matrix for prediction of building use type using a random forest model.

labels within these spreadsheets aligned with the patterns of use characterization using the temporal feature extraction. This proposal was met some positive feedback, albeit there was a hesitation to confirm fully that this process would be entirely necessary if labor were directed to do the same task.

Many of the case study subjects then were shown a series of graphics designed to tell the story of building use type characterization in an automated way. Fig. 11 is the first graphic shown to the subjects. Each of the variables visualized in this figure has been scaled within their ranges, which causes the most extreme values to occur at the minimum and maximum of the y-axis. This figure illustrates several of the most easily understood temporal features and how they break down across the various building use types. This graphic was created using the data for a particular case study; therefore more separation between the classes exist than in the prediction of classes found in the previous section. Discussions using this graphic first centered around the first feature: *Daily Magnitude per Area*. It was intuitive to most participants that a university laboratory has more and primary/secondary schools have less consumption per area than the other use types. It is more surprising, however, that certain building use types are characterized well by other features, such as a number of breakouts with primary/secondary schools and daily and weekly specificity with university dormitories. Outlier buildings for each of the primary use types can be found for all of the variables; this occurrence is natural in the construction industry, and these have not been filtered.

6.2. Characterization of building performance class

The second objective targeted in this study is the ability for temporal features to characterize whether a building performs well or not within it use-type class. Consumption is the metric being measured; therefore it's not the goal of this analysis to predict the performance of a building, its to determine which temporal characteristics are correlated with good or poor performance. This effort is related to the process of benchmarking buildings. Using the insight gained through characterization of building use type, it is possible to inform whether a building's behavior matches its peers. Once a building is part of a peer group, its necessary to understand how well that building performs within that group. In this section, the case study buildings are divided according to which percentile each fits within in its in-class performance. The buildings are divided according to percentiles, with those in the lowest 33% are classified as "Low," the 33–66% percentile are "Intermediate," and the top 33% are classified as "High." As in the previous section, these classifications and a subset of temporal features are implemented into a random forest model to understand how well the features are at characterizing the different classes. Since this objective is related to consumption, all input features with known correlations to consumption were removed from the training set. These include the prominent features of consumption per area, but also include many of the statistical metrics such as maximum and minimum values. Most of the daily ratio input features remain in the analysis as they are not directly correlated with total consumption. Fig. 12 illustrates the results of the model in an error matrix. It can be seen that *high* and *low* consuming buildings are well characterized. The *intermediate* buildings have higher error rates and are often mis-

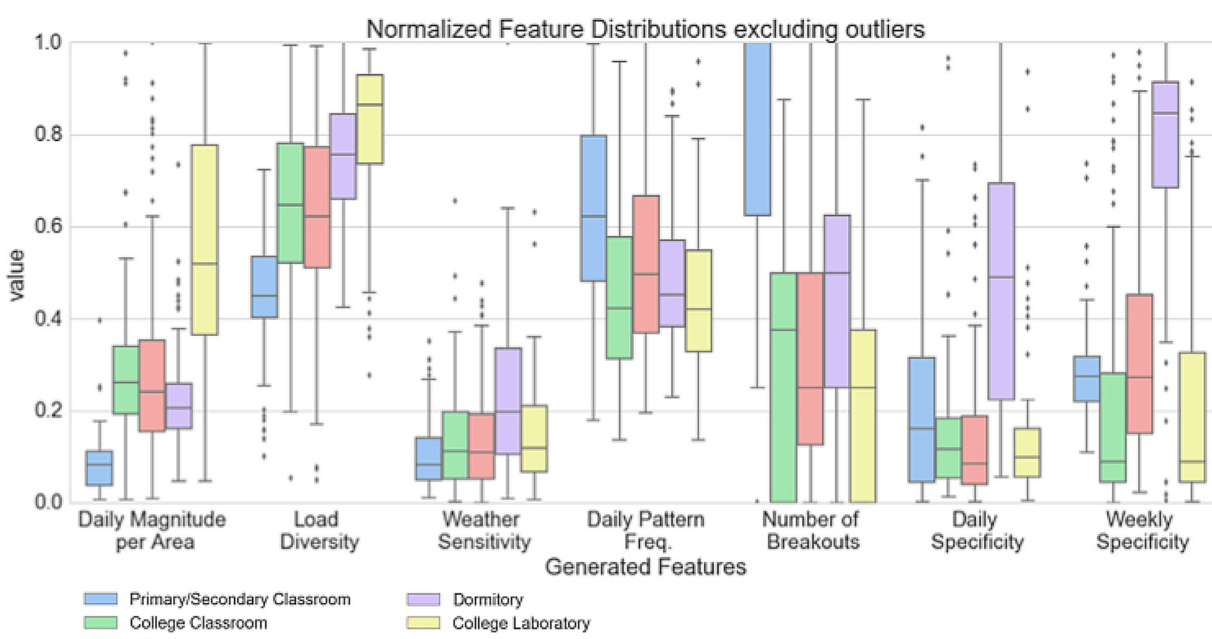

Fig. 11. Simplified breakdowns of general features according to building use type that were presented to case study subjects.

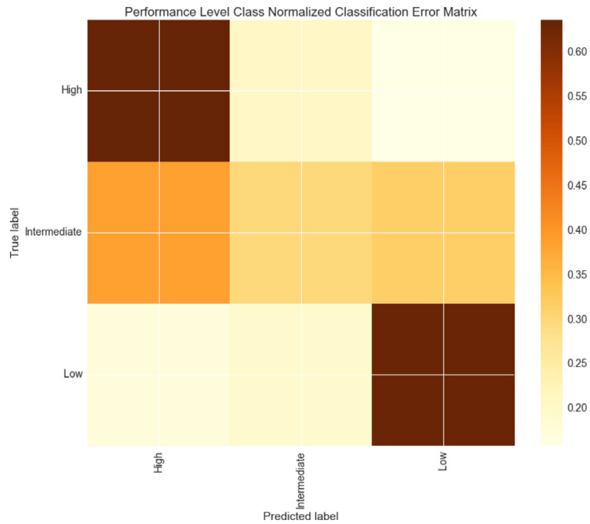

Fig. 12. Classification error matrix for prediction of performance class using a random forest model.

classified with the other two classes. The overall accuracy of the model for classification is 62.3% as compared to a baseline of 38%.

In a situation similar to the discussion about building use type, participants in the case studies were guided through the process of analysis using a subset of features from buildings on their campus. Fig. 13 illustrates a graphic that was shown to the groups. In this case, the buildings are divided into two classes: *Good* and *Bad*. These categories are based on whether the building falls in the upper or lower 50% within its class. The first observation by the case study participants is that the load diversity, or the daily maximum versus minimum, is a reliable indicator of the performance class. This fact is not surprising as this metric indicates the magnitude of the base load consumption as compared to the peak. Other relatively strong differentiators, in this case, are cooling energy, seasonal changes, and weekly specificity. The discussions related to this graphic centered around the potential for the temporal features to inform *why* a building is performing well or not.

Fig. 14 illustrates another graphic related to building consumption classes that were discussed with case study participants. This graphic is an overview of the distributions of the simplified set of features for a certain campus as compared to the entire set of case study buildings. This graphic shows where the buildings on this campus stand as compared to their peers. In this case, the buildings are on the higher end of the normalized consumption, which could likely be because they're also almost all in the most top 20% of buildings for heating energy consumption. The buildings also have a relatively high load diversity, thus the base loads for this campus are likely higher than average and interventions could be designed to reduce this unoccupied load. Many of the case study participants saw this insight as useful as it *supplements* the information from benchmarking.

6.3. Characterization of operational strategies

The final characterization objective for the case studies is the ability for the temporal features to classify buildings from the same campus, and thus buildings that are being operated in similar ways. This characterization takes into account the similarity in occupancy schedules, patterns of use, and other factors related to how a building performs. Like the performance classes, this type of classification is more important in understanding the features that contribute to the differentiation, rather than the classification itself. Seven campuses were selected from the 507 buildings to create seven *groups* of buildings to characterize the difference between their operating behavior. Features were removed for this objective that are indicators of weather sensitivity as these would be related to the location of the buildings, and thus, the campus that they're located. Fig. 15 illustrates the results from the random forest model trained on these data. The accuracy of this model is 80.5% as compared to a baseline of 16.9%. The model is excellent at predicting some of the groups, such as groups 1–4, which more deficient in others, such as 5–7. The high accuracy of this prediction is surprising and lends itself to the ability of the temporal features and the random forest model to predict the operational normalities of these buildings.

7. Conclusion

This paper was undertaken with objectives related to the characterization of building behavior using temporal feature extraction and variable importance screening. The primary goal of the effort is to automate the process of predicting various types of meta-data.

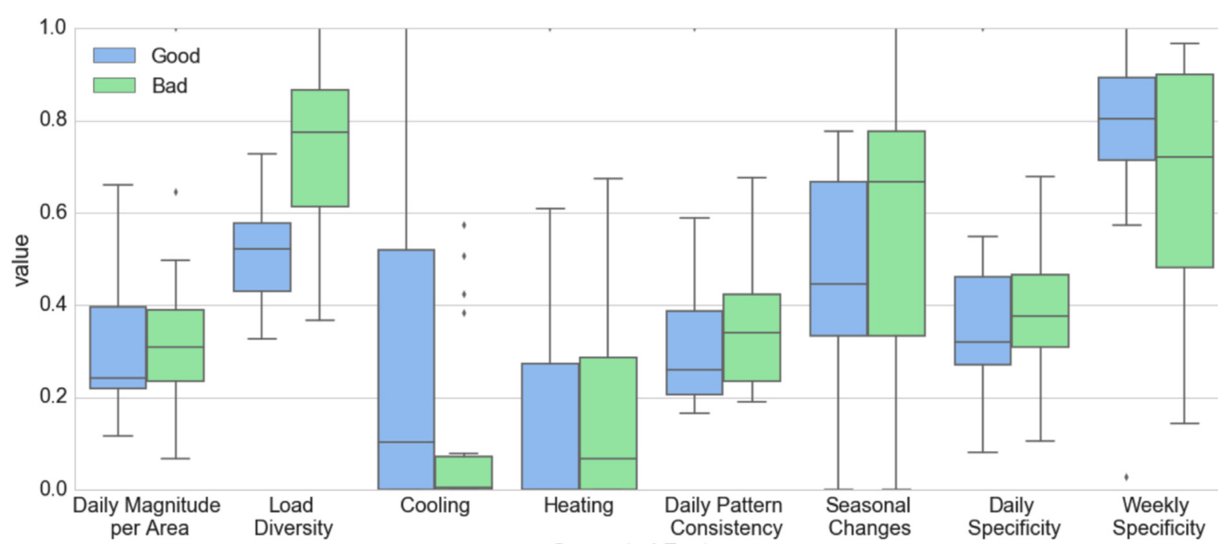

Fig. 13. Simplified breakdowns of general features according to performance level that were presented to case study subjects.

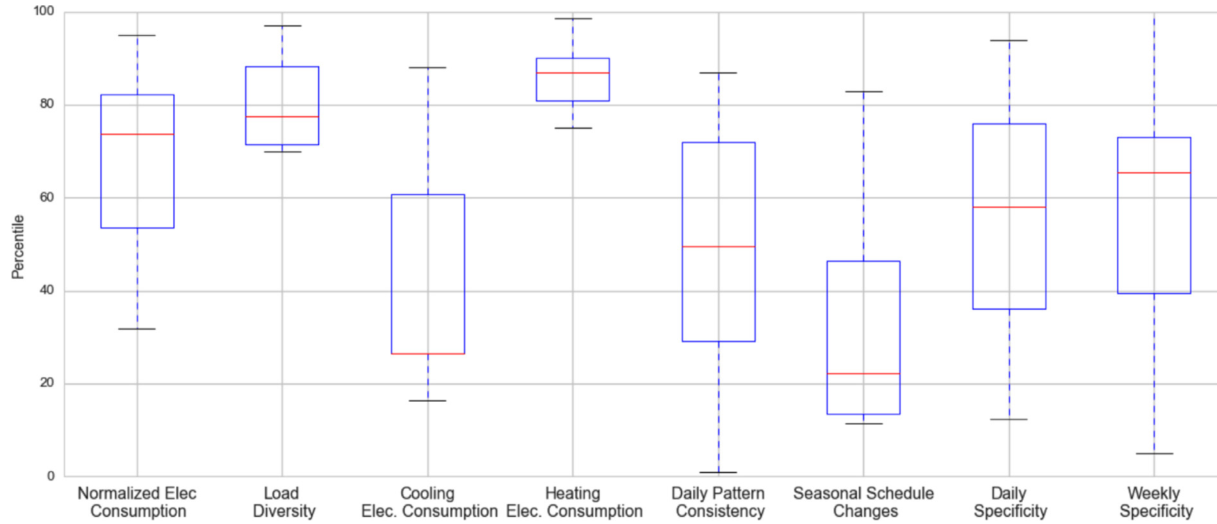

Fig. 14. Feature distributions of a single campus as compare to all other case study buildings.

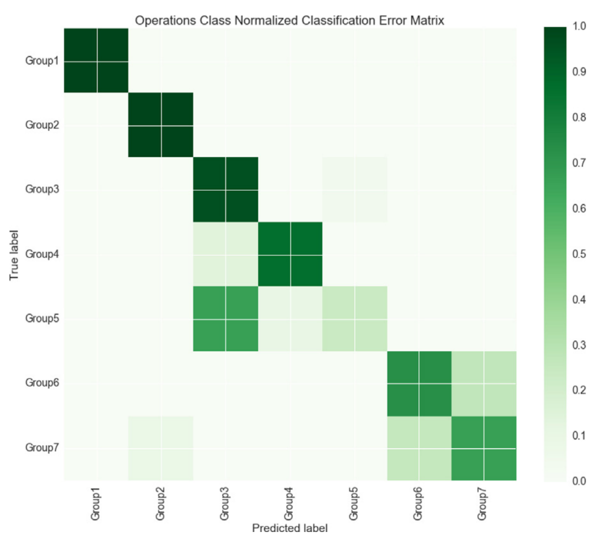

Fig. 15. Classification error matrix for prediction of operations group type using a random forest model.

A framework of analysis was developed to address and test this effort. This process was implemented on two sets of case study buildings and the key quantitative conclusions include:

- The framework can characterize primary building use type with a general accuracy of 67.8% as compared to a baseline model of 22.2% based on five use type classes. Temporal features enable a three-fold increase in building use prediction. Pattern-based features are the most common category in the top ten in the characterization of use-type, thus are important differentiators as compared to more traditional features. Features from the *stl* decomposition process were found to be important as well due to the ability to distinguish differences in normalized weekly patterns.
- Building performance class overall accuracy of the model for classification is 62.3% as compared to a baseline of 38%. The top indicator of high versus low building in-class performance was temporal features pattern specificity. Once again, pattern-based temporal features were found to be significant in distinguishing between different types of behavior.
- For operations class, the accuracy of this model is 80.5% as compared to a baseline of 16.9%, a four-fold increase. Daily scheduling of buildings was captured using the *DayFilter* features, accounting for half of the entire input features.

7.1. Open data and reproducible research

The source code and analytics workflow of this paper can be found in a series of Jupyter notebooks found in a GitHub repository.[8] The data set that is utilized for the analysis is the Building Data Genome Project.[9] These analysis files can be downloaded, and much of the work replicated.

Acknowledgements

The authors would like to thank all of the building operations and maintenance professionals from around the world that assisted in the gathering of the data utilized. This study was funded by a Fellowship from the Institute of Technology in Architecture (ITA) at the ETH Zürich.

References

[1] E. Mills, Building commissioning: a golden opportunity for reducing energy costs and greenhouse gas emissions in the United States, Energy Effic. 4 (2011) 145–173.
[2] M. Effinger, H. Friedman, D. Moser, Building Performance Tracking in Large Commercial Buildings: Tools and Strategies – Subtask 4.2 Research Report: Investigate Energy Performance Tracking Strategies in the Market, Technical Report, 2010.
[3] J. Ulickey, T. Fackler, E. Koeppel, J. Soper, Building Performance Tracking in Large Commercial Buildings: Tools and Strategies – Subtast 4.3 Characterization of Fault Detection and Diagnostic (FDD) and Advanced Energy Information System (EIS) Tools, Technical Report, 2010.
[4] E. Greensfelder, H. Friedman, E. Crowe, Building Performance Tracking in Large Commercial Buildings: Tools and Strategies – Subtask 4.4 Research Report: Characterization of Building Performance Metrics Tracking Methodologies, Technical Report, 2010.
[5] The White House, FACT SHEET: Cities, Utilities, and Businesses Commit to Unlocking Access to Energy Data for Building Owners and Improving Energy Efficiency, 2016.
[6] C. Miller, Z. Nagy, A. Schlueter, A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings, Renew. Sustain. Energy Rev. (2017).
[7] A. Albert, M. Maasoumy, Predictive segmentation of energy consumers, Appl. Energy 177 (2016) 435–448.
[8] A. Albert, R. Rajagopal, R. Sevlian, Segmenting consumers using smart meter data, in: Proceedings of the Third ACM Workshop on Embedded Sensing

[8] https://github.com/buds-lab/temporal-features-for-nonres-buildings-library.
[9] http://www.buildingdatagenome.org/.

Systems for Energy-Efficiency in Buildings, BuildSys '11, ACM, New York, NY, USA, 2011, pp. 49–50.
[9] S. Borgeson, Targeted Efficiency: Using Customer Meter Data to Improve Efficiency Program Outcomes, PhD, University of California, Berkeley, Berkeley, CA, USA, 2013.
[10] J. Kwac, R. Rajagopal, Demand response targeting using big data analytics, in: 2013 IEEE International Conference on Big Data, IEEE, 2013, pp. 683–690.
[11] T. Rasanen, M. Kolehmainen, Feature-based clustering for electricity use time series data, in: Adaptive and Natural Computing Algorithms: 9th International Conference, ICANNGA 2009, Springer, Kuopio, Finland, 2009, pp. 401–412.
[12] F. Iglesias, W. Kastner, Analysis of similarity measures in times series clustering for the discovery of building energy patterns, Energies 6 (2013) 579–597.
[13] S. Petcharat, S. Chungpaibulpatana, P. Rakkwamsuk, Assessment of potential energy saving using cluster analysis: a case study of lighting systems in buildings, Energy Build. 52 (2012) 145–152.
[14] A. Lavin, D. Klabjan, Clustering time-series energy data from smart meters, Energy Effic. 8 (2014) 681–689.
[15] G. Chicco, I.S. Ilie, Support vector clustering of electrical load pattern data, IEEE Trans. Power Syst. 24 (2009) 1619–1628.
[16] S.M. Bidoki, N. Mahmoudi-Kohan, M.H. Sadreddini, M. Zolghadri Jahromi, M.P. Moghaddam, Evaluating different clustering techniques for electricity customer classification, in: Transmission and Distribution Conference and Exposition, 2010 IEEE PES, IEEE, New Orleans, LA, USA, 2010, pp. 1–5.
[17] I. Panapakidis, M. Alexiadis, G. Papagiannis, Evaluation of the performance of clustering algorithms for a high voltage industrial consumer, Eng. Appl. Artif. Intell. 38 (2015) 1–13.
[18] S.P. Pieri, M. IoannisTzouvadakis, Santamouris, Identifying energy consumption patterns in the Attica hotel sector using cluster analysis techniques with the aim of reducing hotels' CO_2 footprint, Energy Build. 94 (2015) 252–262.
[19] H.X. Zhao, F. Magoules, A review on the prediction of building energy consumption, Renew. Sustain. Energy Rev. 16 (2012) 3586–3592.
[20] S.V. Verdu, M.O. Garcia, C. Senabre, A.G. Marin, F.J.G. Franco, Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps, IEEE Trans. Power Syst. 21 (2006) 1672–1682.
[21] A.R. Florita, L.J. Brackney, T.P. Otanicar, J. Robertson, Classification of commercial building electrical demand profiles for energy storage applications, in: Proceedings of ASME 2012 6th International Conference on Energy Sustainability & 10th Fuel Cell Science, Engineering and Technology Conference (ESFuelCell2012), San Diego, CA, USA, 2016.
[22] T.G. Nikolaou, D.S. Kolokotsa, G.S. Stavrakakis, I.D. Skias, On the application of clustering techniques for office buildings' energy and thermal comfort classification, IEEE Trans. Smart Grid 3 (2012) 2196–2210.
[23] J. Ploennigs, B. Chen, P. Palmes, R. Lloyd, e2-Diagnoser: a system for monitoring, forecasting and diagnosing energy usage, in: 2014 IEEE International Conference on Data Mining Workshop (ICDMW), IEEE, Shenzhen, China, 2014, pp. 1231–1234.
[24] A. Shahzadeh, A. Khosravi, S. Nahavandi, A. Shahzadeh, A. Khosravi, S. Nahavandi, Improving load forecast accuracy by clustering consumers using smart meter data, in: Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 2015, pp. 1–7.
[25] A. Reinhardt, S. Koessler, PowerSAX: fast motif matching in distributed power meter data using symbolic representations, in: Proceedings of 9th IEEE International Workshop on Practical Issues in Building Sensor Network Applications (SenseApp 2014), IEEE, Edmonton, Canada, 2014, pp. 531–538.
[26] Z. Liu, H. Li, K. Liu, H. Yu, K. Cheng, Design of high-performance water-in-glass evacuated tube solar water heaters by a high-throughput screening based on machine learning: a combined modeling and experimental study, Sol. Energy 142 (2017) 61–67.
[27] Y. Chen, T. Hong, M.A. Piette, Y. Chen, T. Hong, M.A. Piette, City-scale building retrofit analysis: a case study using CityBES, in: Building Simulation 2017, San Francisco, CA, USA, 2017.
[28] C. Miller, Screening Meter Data: Characterization of Temporal Energy Data from Large Groups of Non-Residential Buildings, Ph.D. thesis), ETH Zürich, Zurich, Switzerland, 2017.
[29] C. Miller, F. Meggers, The Building Data Genome Project: an open, public data set from non-residential building electrical meters, Energy Proc. 122 (2017) 439–444.
[30] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[31] T. Hastie, R. Tibshirani, J.H. Friedman, The elements of statistical learning: data mining, inference, and prediction, in: Springer Series in Statistics, 2nd ed., Springer, New York, NY, 2009.
[32] G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Understanding variable importances in forests of randomized trees, in: Advances in Neural Information Processing Systems, 2013, pp. 431–439.
[33] S. Borgeson, J. Kwac, visdom: R package for energy data analytics. R package version 0.9, 2015.
[34] T. Mitsa, Temporal Data Mining, Chapman and Hall/CRC, 2010.
[35] C. Miller, A. Schlueter, Forensically discovering simulation feedback knowledge from a campus energy information system, in: Proceedings of the 2015 Symposium on Simulation for Architecture and Urban Design (SimAUD 2015), SCS, Washington DC, USA, 2015, pp. 33–40.
[36] M.F. Fels, PRISM: an introduction, Energy Build. 9 (1986) 5–18.
[37] J.W. Taylor, L.M. De Menezes, P.E. McSharry, A comparison of univariate methods for forecasting electricity demand up to a day ahead, Int. J. Forecast. 22 (2006) 1–16.
[38] P. Price, Methods for Analyzing Electric Load Shape and its Variability, Lawrence Berkeley National Laboratory, 2010.
[39] J.L. Mathieu, P.N. Price, S. Kiliccote, M.A. Piette, Quantifying changes in building electricity use, with application to demand response, IEEE Trans. Smart Grid 2 (2011) 507–518.
[40] J.K. Kissock, C. Eger, Measuring industrial energy savings, Appl. Energy 85 (2008) 347–361.
[41] R.B. Cleveland, W.S. Cleveland, J.E. McRae, I. Terpenning, STL: a seasonal-trend decomposition procedure based on loess, J. Off. Stat. 6 (1990) 3–73.
[42] P. Patel, E. Keogh, J. Lin, S. Lonardi, Mining motifs in massive time series databases, in: Proceedings of 2002 IEEE International Conference on Data Mining (ICDM 2002), Maebashi City, Japan, 2002, pp. 370–377.
[43] E.J. Keogh, J. Lin, A. Fu, Hot sax: efficiently finding the most unusual time series subsequence, in: Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05), IEEE, Houston, TX, USA, 2005.
[44] C. Miller, Z. Nagy, A. Schlueter, Automated daily pattern filtering of measured building performance data, Autom. Construct. 49 (Part A) (2015) 1–17.
[45] J. Lin, E.J. Keogh, S. Lonardi, B. Chiu, A symbolic representation of time series, with implications for streaming algorithms, in: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD '03), ACM, San Diego, CA, USA, 2003, pp. 2–11.
[46] P. Senin, S. Malinchik, SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space Model, in: 2013 IEEE 13th International Conference on Data Mining, Institute of Electrical & Electronics Engineers (IEEE), 2013.
[47] N.A. James, A. Kejariwal, D.S. Matteson, N.A. James, A. Kejariwal, D.S. Matteson, Leveraging Cloud Data to Mitigate User Experience from Breaking Bad, 2014 arXiv:1411.7955.