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.