Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 122 (2017) 439–444

Energy Procedia

www.elsevier.com/locate/procedia

CISBAT 2017 International Conference Future Buildings & Districts Energy Efficiency from Nano to Urban Scale, CISBAT 2017 6-8 September 2017, Lausanne, Switzerland

Smart Buildings (Predictive & Neuro-Fuzzy Control)

The Building Data Genome Project: An open, public data set from non-residential building electrical meters

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

[a]*Building and Urban Data Science (BUDS) Lab, Dept. of Building, National University of Singapore (NUS), 4 Architecture Drive, Singapore 117566, Singapore*

[b]*Future Cities Laboratory (FCL), Singapore-ETH Centre (SEC), 1 Create Way, Singapore 138602, Singapore*

[c]*Cooling and Heating for Architecturally Optimized Systems (CHAOS) Lab, Andlinger Center for Energy and Environment, Dept. of Architecture, Princeton University, Princeton, NJ, 08544, USA*

Abstract

As of 2015, there are over 60 million smart meters installed in the United States; these meters are at the forefront of big data analytics in the building industry. However, only a few public data sources of hourly non-residential meter data exist for the purpose of testing algorithms. This paper describes the collection, cleaning, and compilation of several such data sets found publicly on-line, in addition to several collected by the authors. There are 507 whole building electrical meters in this collection, and a majority are from buildings on university campuses. This group serves as a primary repository of open, non-residential data sources that can be built upon by other researchers. An overview of the data sources, subset selection criteria, and details of access to the repository are included. Future uses include the application of new, proposed prediction and classification models to compare performance to previously generated techniques.

Peer-review under responsibility of the scientific committee of the CISBAT 2017 International Conference – Future Buildings & Districts – Energy Efficiency from Nano to Urban Scale

Keywords: Open Data, Non-Residential Building Meter Data, Benchmark Data Set, Big Data, Machine Learning

1. Introduction

The past three decades have produced an explosion of building performance research. Thousands of publications have been written to describe techniques, algorithms, and workflows designed to reduce energy consumption in buildings and improve indoor environmental conditions. The largest challenge in the utilization of this research is the

* Corresponding author. Tel.: +65 65163421
E-mail address: clayton@nus.edu.sg

1876-6102
Peer-review under responsibility of the scientific committee of the CISBAT 2017 International Conference – Future Buildings & Districts – Energy Efficiency from Nano to Urban Scale
10.1016/j.egypro.2017.07.400

ability to compare different approaches against each other in an objective or quantified way. This situation draws parallels from the fields of generalized time-series data mining. In this domain, a popular paper by Keogh et. al states that: "Much of this work has very little utility because the contribution made"..."offer an amount of *improvement* that would have been completely dwarfed by the variance that would have been observed by testing on many real world datasets, or the variance that would have been observed by changing minor (unstated) implementation details."[1] In the temporal data mining research community, the challenge was addressed through the development of open, benchmarking data sets available for researchers to implement their proposed algorithms for an objective comparison to previously developed techniques. One such example is the UCR Time Series Classification Archive [2]. This archive contains numerous time-series data sets which new algorithms can be tested on to understand performance differences from the status quo.

This paper describes the development of an open, building performance data set from a large group of non-residential buildings. This project, known as the *Building Data Genome* project seeks to create such a benchmarking data set created from raw data from real buildings. This project started through a series of interviews with building operations teams at various university campuses around the world. The target of these interactions was to collect at least one year of hourly data from whole building electrical meters, resulting in at least 8760 measurements per building. Several of these data sets were obtained through a series of site visits and interviews. These interactions are detailed by giving an in-depth overview of these case studies by discussing the current performance data acquisition systems and the standard methods of utilizing those data for tracking activities. A key goal of the collection of these data was that they would be a basis for an open, shareable data repository for building performance research. This goal was discussed with the case study participants. Several other raw data sets were collected from open data sources on the Internet and were included in this study, albeit often with less metadata available. This project builds upon previous projects with released open data [3,4] and is similar to open, labeled data sets from the load disaggregation research domains [5,6].

2. Data sources

2.1. Site visits for case studies

Throughout the course of two years, from February 2014 to April 2016, several site visits were conducted to interview operations staff at seven campuses. The purpose of this effort was two-fold: first, to collect as much raw, temporal data from each site as possible and, second, to discuss the status quo of building energy analysis as performed on their campus. This section discusses these site visits, the types of data that were collected, and a few of the lessons learned from the process. A consistent theme in the site visits was that each campus has been investing in electrical metering and data acquisition systems over the past decade. In every one of the case study interviews, the operations staff discussed the under-utilization of the data being collected. A common phrase was, "We have more meter data than any time before, and we don't know what to do with it." Another typical situation was that a campus had a large electrical metering infrastructure but did not know how to extract raw data for this research project. This scenario occurred on two of the seven campuses after the first interview, and data was still not available even after a follow-up visit of those campuses. Therefore, five of the seven case studies had data available and are discussed in the following subsections.

2.1.1. Case study 1

The first case study is a campus in a continental climate in the Midwest region of the United States. It is a university with 226 buildings spread across two main campuses. Altogether, these buildings have a total floor area over 2.3 million square meters (25 million square feet). An initial interview was conducted with the lead statistician of the facilities management in March 2015. Information was gathered on the building and energy management systems of the campus, and a discussion regarding the typical utilization of the data was conducted. It was found that there are over 480 electrical meters on the campus and that these data were primarily used for billing of the individual academic departments. They have a custom metering data management platform with some capabilities for data export. A second site visit was conducted in June 2015 to facilitate the collection of a sample one year data set. In this site visit, a facilities management professional with experience in SQL databases was able to directly query

the underlying back-end of the energy management system to extract one year of raw data from all of the metering infrastructure on the campus. An accompanying meta-data spreadsheet was discovered that included information on floor area, primary space usage, EnergyStar score, and address. These data were then used for the analysis and feature extraction, and some of the results were compiled and presented to the entire facilities management department of this university in March 2016. This presentation gave an overview of the feature creation techniques and an understanding of how the buildings on their campus compare to other schools.

2.1.2. Case study 2

The second case study is a campus in the Northeast region of the United States. It is also a University, and it has 180 buildings on a single main campus. An initial meeting was organized in April 2015 with the facilities management team. This campus has well-organized building and energy management systems with a strong emphasis on data acquisition and management. The campus has an analytics and automated fault detection software platform that is connected to the underlying controls systems. A follow-up campus visit was conducted in August 2015 to facilitate the download of a raw, example data set from the buildings on campus. At this point, a log-in to a new data management platform was given for the purposes data extraction. Several issues arose from the use of this platform, and ultimately, a database query by the software developers of the system was used to extract the one year of electrical meter data from the campus buildings. Once again, a spreadsheet of meta-data was shared that included information on floor area and primary building use type. A final site visit was conducted in April 2016 to discuss some of the results of the data acquisition and upcoming plans for upgrades. A formal presentation of the results was not able to be given; thus only limited feedback of the implementation progress was collected.

2.1.3. Case study 3

The third case study is a campus in the Midwest region of the United States. Once again, it is a university campus with 25 buildings encompassing 204,000 square meters (2.2 million square feet) of floor space. An initial site survey and discussion of the campus was conducted in March 2015 with the campus lead mechanical and energy engineers. This campus has its electrical meters connected to a campus energy management platform that includes various visualizations and analytic techniques. This platform also can quickly provide raw data download for analysis in this study. This platform resulted in the campus achieving by far the most user-friendly on data collection out of the case study set, including the open, on-line data sources. Raw data in flat files was easily downloaded for all data points at once. The meta-data for this campus was also extracted from this energy management platform, albeit in a more manual method from the user interface. A follow-up visit to this campus was conducted in March 2016 with initial results of characterizing the data according to a subset of the tested features. A significant amount of feedback for this case study was given by the facilities management department regarding the ability for these insights to assist in their decision-making processes.

2.1.4. Case study 4

The fourth case study is an international school campus in tropical Southeast Asia. This campus includes five buildings with approximately 58,000 square meters (625,000 square feet). It was built and opened in 2010 and includes some sustainable design features such as an optimized chilled water plant, solar thermal cooling system, and an innovative, fresh air delivery system. The building management and data acquisition system have been a primary focus of the operations director of the campus for many years. Discussions and interviews with the operations staff have occurred numerous times over the course of the last five years. The operations team of this organization has been an active contributor to the development of the methodology.

2.1.5. Case study 5

The final case study to be outlined is a university campus located in Switzerland. This campus includes 22 building encompassing more than 150,000 square meters (1.6 million square feet). This campus has an energy management system with the ability to extract raw data, albeit only one point at a time. Data from this campus were utilized in a previous research project focused on campus and building-scale co-simulation and modeling. Only email correspondence with the campus facilities managers of this campus was conducted. A significant amount of meta-data was available from the facilities department through a spreadsheet that provided the breakdown of primary uses of the spaces in each building.

Source Name	Description	Website
Cornell University	EMCS Portal	http://portal.emcs.cornell.edu/
University of California - Berkeley	Berkeley Campus Energy Portal	http://berkeley.openbms.org/
Arizona State University	Campus Metabolism	https://cm.asu.edu
Carbon Culture	Community Open Data Platform	https://platform.carbonculture.net
EnerNOC	EnerNOC GreenButton Data	https://open-enernoc-data.s3.amazonaws.com/anon/ index.html
University of Southampton	Open Data Service	http://data.southampton.ac.uk/

Table 1. Open, online data sources

2.2. Online open case studies

Several large data sets were found through a search of openly accessible data on-line. This section gives an overview of these data sources and the methods in which the data was extracted and pre-processed for analysis. Table 1 illustrates these sources, a short description of the platform in which the data was downloaded, and the URL of the platform. As in the site visit case studies, one year of hourly, whole-building electrical meter data were collected from each of these sources for as many buildings as possible.

3. Overview of data collected

Through data collection from the on-site case study interviews and on-line data sources, whole-building electrical meter data from 1238 buildings were collected. Figure 1 illustrates the locations of these building around the world. A majority of the buildings are located in the United States, with the highest concentrations in the northeast region. A wide range of building types is included in the data set, from Education and Government to Agriculture and Heavy Industry.

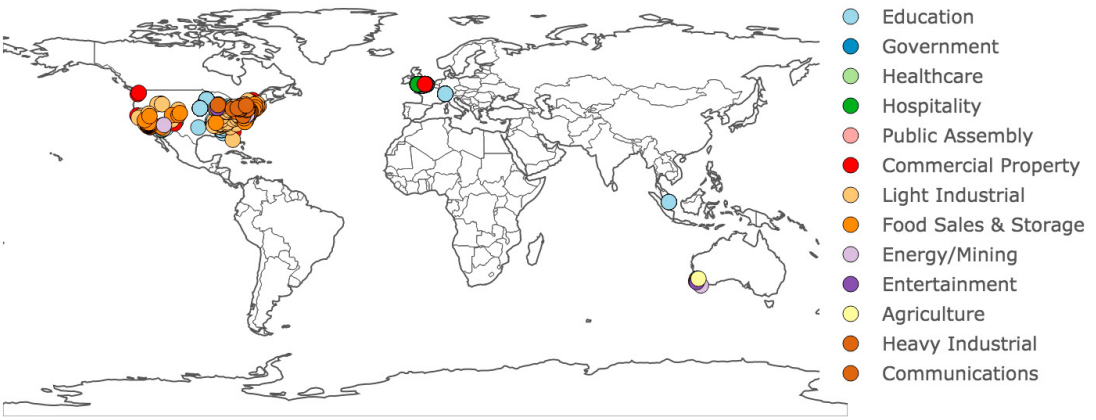

Fig. 1. Locations of 1238 case study buildings collected from across the world

3.1. Selection of case study subset according to most common primary use type

A subset of buildings was chosen based on limiting criteria for inclusion in this study. The primary consideration is that the building is a member of one of the top primary use types: Offices, Primary/Secondary Schools, University Laboratories, University Classrooms, or Dormitories. These categories and the number of buildings in each one are

shown in the lower right bar chart in Figure 2. The buildings are distributed across various time zone regions as seen in the upper left bar chart. The east coast of the United States is the largest group due to the number of campuses and buildings from the EnerNOC data source. All of the buildings from the Carbon Culture data source are located in the United Kingdom. The bar charts in the upper right and lower left illustrated the industries and sub-industries from which the case study buildings are collected. The number of university campuses is strongly evident in both charts.

Two other critical metadata were collected and included as part of the open data set for each sample: gross floor area (in square meters and square feet) and weather (in the form of an hourly weather file). Other types of meta-data are available for some, but not all, samples: heating type, main heating energy source, the number of floors, peak occupancy, energy rating, and year of construction.

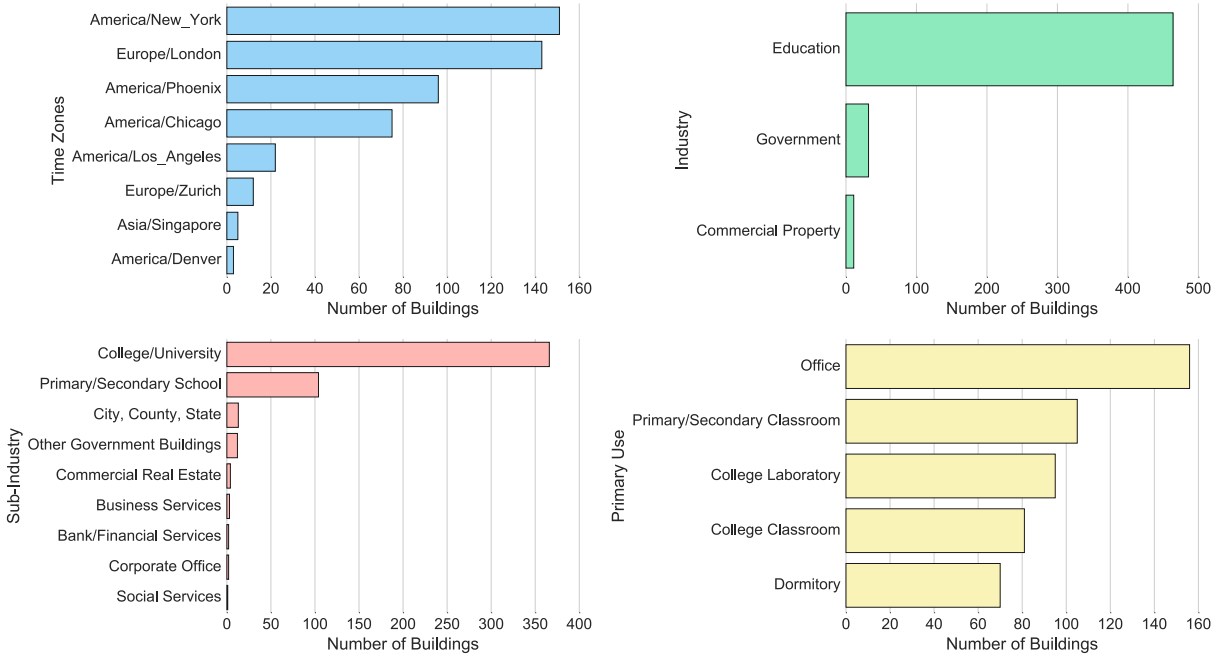

Fig. 2. Distribution of case study buildings among time zone, industry, sub-industry and primary use type

3.2. Data cleaning and organization

The data collected from each of the sources were first re-organized to conform to a standard format for easy dissemination and utilization by third-parties. Each of the meter data sets was cleaned by removing gross outliers and filling small gaps using the LoadShape model and associated Python library. 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[1]. The timestamps of each data set were converted to Coordinated Universal Time (UTC) and combined into one comma-separated value (CSV) file. The columns of this file are each of the 507 buildings with each one including 8760 data points, one for each hour over a certain time range. Some of the data sets with sub-hourly values were resampled by taking the mean of values at each hour. This file has an index time range of over five years as many of the data sources have different start and end dates. Included in the repository is a *meta* data file that contains each of the descriptors of the buildings including starting and ending

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

dates, industry, primary space type, area, timezone, and several other factors related to rated performance or systems specifications. The meta data file includes and anonymized unique identifier that couples the primary use type with a human-like name; e.g., .PrimClass_Eva, Office_Erik, and PrimClass_Ebony.

4. Conclusions

This paper describes the release of an open, shareable set of building performance-related data from 507 buildings. Each building includes hourly, whole-building electrical meter data, and various characteristic meta-data such as gross floor area, primary use type, weather information, and industry. The dataset is available for download and collaboration within a GitHub repository (https://github.com/buds-lab/the-building-data-genome), and the data are analyzed through a set of Jupyter notebooks, mainly within the context of work done as part of a doctoral dissertation [7]. Other researchers are welcome to collaborate, add data, or improve the available data sets.

A potential use of these data is as a basis for testing various algorithms and feature extraction techniques. Significant amounts of essential information can be extracted from temporal data to characterize a commercial building. The harvest of this information can assist in the implementation of conventional analysis techniques, as inputs to classify or benchmark a building, or to predict whether a building is a good candidate for individual energy savings measures. To extract information solely from these sensors, new features can be created from these raw data. These features are designated as temporal as they summarize behavior that is occurring in time-series data. Competitions could be created in which professionals are tasked with creating the best models for forecasting, developing various features, or classifying buildings using the *Building Data Genome Project* as a source for a large generalizable data on non-residential buildings.

5. Acknowledgments

The authors would like to acknowledge the various campus facilities operations teams involved in the collection and dissemination of these data. This research was funded through an Institute of Technology in Architecture (ITA) Fellowship from the ETH Zürich.

References

[1] Keogh, E., Kasetty, S.. On the need for time series data mining benchmarks: A survey and empirical demonstration. Data Mining and Knowledge Discovery 2003;7(4):349–371. URL: http://link.springer.com/article/10.1023/A:1024988512476.

[2] Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A., et al. The UCR Time Series Classification Archive. 2015.

[3] Miller, C., Schlueter, A.. 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). Washington DC, USA: SCS; 2015, p. 33–40. URL: http://dx.doi.org/10.13140/RG.2.1.2286.0964. doi:10.13140/RG.2.1.2286.0964.

[4] Miller, C., Nagy, Z., Schlueter, A.. A seed dataset for a public, temporal data repository for energy informatics research on commercial building performance. In: Proceedings of the 3rd Conf. on Future Energy Business & Energy Informatics. Rotterdam, Netherlands; 2014,doi:10.13140/RG.2.1.4620.8485.

[5] Kolter, J.Z., Johnson, M.J.. REDD: A public data set for energy disaggregation research. In: Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA; vol. 25. Citeseer; 2011, p. 59–62.

[6] Anderson, K., Ocneanu, A., Benitez, D., Carlson, D., Rowe, A., Berges, M.. BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research. In: Proceedings of the 2nd KDD workshop on data mining applications in sustainability (SustKDD). 2012, p. 1–5.

[7] Miller, C.. Screening Meter Data: Characterization of Temporal Energy Data from Large Groups of Non-Residential Buildings. Ph.D. thesis; ETH Zürich; Zurich, Switzerland; 2017.