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

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