The Building Data Genome Project: A Collection of Public Datasets for Non-Residential Building Electrical Meter Characterization

Clayton Miller
Institute of Technology in Architecture, ETH Zürich
Address
Zürich, Switzerland
miller.clayton@arch.ethz.ch

Forrest Meggers
Dept. of Architecture, Princeton University
Address
Princeton, NJ, USA
fmeggers@princeton.edu

Zolan Nagy
Affiliation
Address
City, State
nagy@arch.ethz.ch

Arno Schlueter
Institute of Technology in Architecture, ETH Zürich
Address
Zürich, Switzerland
schlueter@arch.ethz.ch

ABSTRACT

As of 2015, there are over 60 million smart meters installed in the United States. These data are at the forefront of *big data* analytics for the building and construction industry. Despite the massive number of meters, only a few public data sources of hourly non-residential meter data exist for the purpose of testing clustering, classification or prediction algorithms. This paper describes the collection, cleaning, and compilation of several such data sets found publicly online, 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 from around the world. The intent of this collection is to serve as an initial repository of open, non-residential data sources that can be built upon by other researchers. Various characterization techniques such as temporal feature extraction, clustering, classification, and prediction are implemented on the data sets to illustrate the usefulness of such a repository.

CCS Concepts

•**Computer systems organization** → **Embedded systems;** *Redundancy;* Robotics; •**Networks** → Network reliability;

Keywords

Open Data, Non-Residential Building Meter Data, Benchmark Data Set

BUILDSYS '16 Stanford, CA, USA
© 2016 ACM. ISBN XX...$X
DOI: XX

1. INTRODUCTION

2. DATA SOURCES

2.1 Data Cleaning and Standardization

2.2 Data Repository

3. TEMPORAL FEATURE EXTRACTION

3.1

4. CLASSIFICATION

5. PREDICTION

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 [1]. Over the years, various other techniques have been developed using techniques such as neural networks, ARIMA models, and more complex regression [**?**]. However, simplified techniques 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 expected assumptions. For example, if actual measurements and predicted consumption match well, the underlying behavior of a energy-consuming systems in the building has been captured adequately. If not, there is uncharacterized phenomenon that will need to be captured with a different type of model or feature.

A contemporary, simplified load prediction technique is selected to create temporal features that capture whether 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 [**?**, **?**]. 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[1].

According to the literature, the model operates as follows [**?**]. 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 simply an empirical indication of when occupancy-related systems are detected to be operating. Seperate 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 $j = 1...6$, is assigned to each interval. Within the model, the outdoor air temperature at time, t, occuring 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 Equation 1.

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

Prediction of unoccupied mode occurs using a single temperature parameter, β_u. Unnoccupied load, L_u, is calculated with Equation 2.

$$L_0(t_i, T(t_i) = \alpha_i + \beta_u T_{c,j}(t_i) \tag{2}$$

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 specific hour, t, is calculated using Equation 3.

$$R_t = \frac{L_{t,a} - L_{t,p}}{max_{L_a}} \tag{3}$$

An example of the TOWT model implemented on one of the case study buildings is seen in Figure **??**. Two primary characteristics are captured from model residual analysis. The first is the building's primary 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 obvious 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 these phenomenon can inform whether the building is of a certain use-type or in a certain juristiction. The second characteristic captured are periods of under prediction when the building is consuming more electricity than expected. These data inform whether a building is being consistently utilized consistently, or whether there is volatility in its normal operating schedule from week-to-week. Figure **??** illustrates an overview of implementation on all the buildings.

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

5.1 LoadS

6. CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] M. F. Fels. PRISM: An introduction. *Energy and Buildings*, 9(1):5–18, Feb. 1986.

APPENDIX