ARTICLE IN PRESS

Applied Energy xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Applied Energy

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

Machine learning approaches for estimating commercial building energy consumption

Caleb Robinson[a], Bistra Dilkina[a,*], Jeffrey Hubbs[e], Wenwen Zhang[b], Subhrajit Guhathakurta[b], Marilyn A. Brown[c], Ram M. Pendyala[d]

[a] School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, United States
[b] School of City and Regional Planning, Georgia Institute of Technology, Atlanta, GA 30332, United States
[c] School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332, United States
[d] School of Sustainable Engineering and the Built Environment, Arizona State University, 660 S. College Avenue, Tempe, AZ, United States
[e] Georgia Institute of Technology, Atlanta, GA 30332, United States

HIGHLIGHTS

- Machine learning models were used to estimate commercial building energy consumption.
- CBECS was used to train a US-wide model with five commonly available features.
- Validation of the model on city-specific building data was performed for New York City.
- The gradient boosting model performs best compared to Linear, SVM, and other methods.
- Availability of more building features results in more accurate models.

ARTICLE INFO

Keywords:
Commercial building energy consumption
Modeling
Machine learning
CBECS

2010 MSC:
00-01
99-00

ABSTRACT

Building energy consumption makes up 40% of the total energy consumption in the United States. Given that energy consumption in buildings is influenced by aspects of urban form such as density and floor-area-ratios (FAR), understanding the distribution of energy intensities is critical for city planners. This paper presents a novel technique for estimating commercial building energy consumption from a small number of building features by training machine learning models on national data from the Commercial Buildings Energy Consumption Survey (CBECS). Our results show that gradient boosting regression models perform the best at predicting commercial building energy consumption, and can make predictions that are on average within a factor of 2 from the true energy consumption values (with an r^2 score of 0.82). We validate our models using the New York City Local Law 84 energy consumption dataset, then apply them to the city of Atlanta to create aggregate energy consumption estimates. In general, the models developed only depend on five commonly accessible building and climate features, and can therefore be applied to diverse metropolitan areas in the United States and to other countries through replication of our methodology.

1. Introduction

There is substantial evidence to suggest that different configurations of the built environment are closely associated with variations in energy consumption and climate altering greenhouse gas emissions [1–5]. While the relationship between urban form and energy use in transportation has been well studied, we know far less about the impact of urban form on residential and commercial energy demands [6–9]. A 2009 study commissioned by the National Academy concluded that increasing development densities leads to modest savings in energy use in transportation, and by extension, a reduction in greenhouse gas emissions [10]. Yet, if our interest is in building energy efficient communities, a more comprehensive set of attributes of the built environment need to be examined to determine whether increasing development densities actually lead to energy savings. The estimation of building energy consumption at the scale of small urban areas is difficult without building level data and few studies have attempted to provide energy footprints for residential and commercial buildings at

* Corresponding author.
E-mail addresses: dcrobins@gatech.edu (C. Robinson), bdilkina@cc.gatech.edu (B. Dilkina), jeffrey.hubbs@gmail.com (J. Hubbs), wzhang300@gatech.edu (W. Zhang), subhro.guha@coa.gatech.edu (S. Guhathakurta), Marilyn.Brown@pubpolicy.gatech.edu (M.A. Brown), ram.pendyala@asu.edu (R.M. Pendyala).

http://dx.doi.org/10.1016/j.apenergy.2017.09.060
Received 26 May 2017; Received in revised form 4 August 2017; Accepted 10 September 2017

neighborhood scales. This paper fills some of that gap by providing a generic technique for estimating commercial building energy from publicly available data in the U.S.

According to the 2015 annual energy consumption data released by the U.S. Energy Information Administration (EIA), residential and commercial buildings consumed 39 quadrillion Btus., representing 40% of total energy consumption in the United States [11]. Similarly, according to the European Commission [12], building energy consumption accounts for 40% of the total energy consumption in the EU. Globally, the building sector accounted for approximately 32% of energy consumption in 2010 [13]. Thus, advanced economies spend a particularly large percentage of their energy in buildings compared to developing countries. While the EIA releases highly detailed annual energy consumption estimates by sector for the U.S. as a whole, it is useful for local policy makers to have small area or neighborhood level estimates of energy consumption. Without access to fine scale data on energy use, urban planners will not be able to benchmark the effects of environmental or climate related policies affecting different sections of the urban region or make confident predictions about the outcomes of proposed policies. Machine learning models can also help city and regional planners predict the energy burdens that could result from alternative urban growth patterns and global warming scenarios. Spatial energy consumption information at a granular scale is therefore crucial to fulfilling sustainability goals.

One way of estimating building energy consumption, in the absence of actual sensor data, is to create physical building models with a "template" of representative buildings, then run thermodynamic simulations to estimate the energy demands [14]. These "engineering" models of building energy consumption are computationally expensive and cannot capture the wide variety of different buildings present in cities, as modeling each type of building requires very detailed input data, which is costly to collect. Statistical models can be used to fill the gaps where resources are too limited to use physical models, or the scale of the study area makes physical modeling impractical.

We aim to model commercial building energy consumption at the building level using machine learning models. This statistical approach avoids expensive physical modeling efforts, and is able to provide reasonable estimates that can be validated against existing building level energy consumption databases. Specifically, we train machine learning models on the 2012 Commercial Building Energy Consumption Survey microdata [15], then validate this approach using the Local Law 84 (LL84) dataset from New York City. We show how our models can be used to create comprehensive metropolitan wide commercial energy consumption maps by applying them to 73,388 commercial buildings in Atlanta, GA. These maps will help city planners better understand the relationships between urban form and energy consumption, and plan for the future. Our models purposefully only rely on a limited set of building level features, namely: square footage, principal building activity, number of floors, and heating and cooling degree days, so that they can be applied to *any* metropolitan area in the United States. Furthermore, to facilitate the wider adoption of our methods in other metropolitan areas throughout the US, we have released the code and trained models used in this paper in a public GitHub repository[1]. The code and instructions provided in the GitHub repository can be used to reproduce the modeling and validation results from this paper, and to apply trained models in new settings. In general, the machine learning modeling approach for broad commercial building energy consumption prediction presented in this work is a novel step toward better understanding the energy consumption landscape in the United States.

[1] The code and trained models are available at: https://github.com/SEI-ENERGY/Commercial-Energy/

2. Related work

Methods for predicting building energy consumption can be categorized into three groups: engineering methods (white-box models), statistical methods (black-box models), and hybrid approaches (grey-box models) [14,16,17]. Engineering methods physically model building energy consumption by simulating the laws of thermodynamics using extensive building level data. This method cannot be applied precisely at the urban scale, due to its large data and computational demands, however it is used to estimate the energy consumption of a small typology of buildings which are then aggregated over entire urban areas [18,16,17]. Statistical methods for estimating building energy consumption aim to directly regress energy consumption values on associated building and climate variables. In general, machine learning methods (such as the methods used in this study) fall into this category, although Zhao and Magouls have separated machine learning based studies from linear regression model based studies in their review [14]. Hybrid methods involve a combination of both engineering and statistical models, and use the output from engineering models as an input to statistical models. The purpose of these models is to offset some of the constraints involved with physical modeling (such as the inability to model every building in a district) with the flexibility of statistical approaches [19].

A commonality between the engineering, statistical, and hybrid methods is that they are all limited by the availability of relevant data. Indeed, availability of data is crucial for any statistical modeling approach, but our method lowers the bar for data requirement as discussed later. Mathew et al. discuss big data applications of the US Department of Energy's building performance database (BDP) [20]. The BDP has data for both residential and commercial buildings on a larger scale than either the Commercial Building Energy Consumption Survey (CBECS) or the Residential Building Energy Consumption Survey (RECS), however can only be used in benchmarking applications. Access to fine grained data, such as that collected by BDP, will be crucial for development of more accurate and relevant statistical based models [21].

Linear statistical models have been used in studies for predicting energy consumption at both the building and zone level resolutions. Boulaire et al. use robust linear models to model energy consumption at the zone level in NSW, Australia [22]. Kuusela et al. use a lognormal modeling framework to model electricity consumption from aggregate building features at the zone level of a Finnish city [23]. Kontokosta use robust linear models to estimate building energy consumption of residential and commercial buildings using 2011 New York City's Local Law 84 (LL84) dataset [24]. While these models are easily interpretable, machine learning models are better suited for modeling the complex relationships between building level characteristics and energy consumption since such models have fewer constraints about the statistical relationships among variables.

Previous studies have shown that machine learning models outperform linear models in modeling building energy consumption. Tso et al. use linear regression models, decision tree models, and neural networks to model residential electricity consumption at the building level in Hong Kong [25]. The study splits the dataset across the summer and winter seasons and trains models separately for each season.Similarly, Fan, Xiao, and Wang use an ensemble of machine learning models to predict the next-day building energy consumption of the "International Commerce Center" in Hong Kong with good results [26]. Wei et al. use two linear models and four non-parametric machine learning models to estimate gas and electricity consumption at a zone level in London [27]. Similarly, Yalcintas et al. train an artificial neural network and multiple linear regression models to predict the energy use intensity values (kWh per square meter) with the 1999 CBECS data [28]. This study only uses one category of building from the CBECS dataset, and categorizes the target values to convert the problem into an easier classification problem. These three studies all find that machine