



Data-Driven Benchmarking of Building Energy Performance at the City Scale

Zheng Yang

Dept. of Civil & Environmental Eng.
Stanford University

473 Via Ortega, Stanford, CA 94305

zheng.yang@stanford.edu

Jonathan Roth

Dept. of Civil & Environmental Eng.
Stanford University

473 Via Ortega, Stanford, CA 94305

jmroth@stanford.edu

Rishee K. Jain

Dept. of Civil & Environmental Eng.
Stanford University

473 Via Ortega, Stanford, CA 94305

rishee.jain@stanford.edu

ABSTRACT

Cities across the country (20 to date) are rapidly passing laws to mandate the collection and disclosure of energy usage data with the hopes that such data could be utilized to benchmark building energy performance and provide a basis for designing and deploying efficiency measures. However, numerous municipalities are struggling to translate such data into actionable insights on which buildings are the best candidates for energy efficiency interventions. Current studies are limited in their ability to compare energy efficiency of buildings at the city scale and maintain interpretability necessary to result in effective decision making that can facilitate municipal policy and program design. In this paper, we present an integrated data-driven method based on recursive partitioning and stochastic frontier analysis to benchmark energy performance of building stock across an entire city. We implement the proposed method on a dataset of over 10,000 buildings in New York City. The preliminary results begin to quantify the potential for building energy efficiency and give further support to city officials for examining effects of possible interventions. We aim to establish a robust methodology that provides a standard way to benchmark building energy performance at the city scale, which can be easily applied to other cities across the country.

Categories and Subject Descriptors

J.1 [Computer Application]: Administrative data processing

Keywords

Energy efficiency; buildings; benchmarking; city scale, recursive partitioning; stochastic frontier; data analytics

1. INTRODUCTION

In the United States, over 40% of energy use is attributed to buildings [1]. Their sustainability and energy conservation are becoming increasingly important, as the majority of the building stock has been identified as inefficient [1]. Understanding the energy performance of buildings has significant potential to encourage building managers and city officials to discover how much energy could be saved if appropriate actions (e.g. changes to building envelope, systems and energy-use behaviors) are enacted to catch up to energy-efficient peers. The advancement of information techniques has made data acquisition of energy use and building characteristics (physical properties such as floor area) increasingly available, even at the city scale. To date, twenty municipal city governments across the country have already passed

legislations mandating the collection and disclosure of energy usage data for a significant portion of their building stock [2]. These large-scale datasets could improve the reliability and accuracy of benchmarking as energy efficient/inefficient buildings can be better identified and targeted by policies and programs. However, a major challenge is how to translate such new energy usage and building characteristics data streams into actionable insights that can aid decision-making and policy design at the municipal level. Moreover, assessing the energy performance of buildings at the city scale also introduces technical challenges since buildings across an entire city vary significantly in the magnitude of energy usage and other building characteristics. Methods for benchmarking building energy performance at the city scale requires balancing interpretability, such that it is useful to policy-makers, and robust as well as scalable, such that the variance of urban building data is adequately addressed.

In this paper, we first apply a recursive partitioning method using a classification and regression tree (CART) to partition buildings into groups to reduce the variance of total energy use based on building characteristics. Stochastic frontier analysis (SFA) is implemented independently within each group for benchmarking building unit energy use (energy use intensity-EUI), where the impact of random errors on energy performance can be successfully separated from that of inefficiency sources. The combination of recursive partitioning and stochastic frontier analysis allows us to preserve the interpretability of results while addressing the variance challenge inherent in the city scale data. Using the coupled datasets of actual building energy use from New York City's Local Law 84 (LL84) and building characteristics from New York City Primary Land Use Tax Lot Output (PLUTO) for the 2014 calendar year, our integrated method is applied to quantitatively assess the extent to which energy efficiency of any given building could be improved.

2. RELATED WORK

Extensive studies have developed benchmarking methods aimed at comparing the energy performance of buildings. This body of work can be divided into multiple categories based on how building characteristics are linked with energy use. These categories include simple indicators (e.g., EUI measured in kBtu/ft²) [3], point-based ratings (e.g., LEED rating) [4], simulation models (e.g., EnergyPlus model) [5], statistical methods (e.g., linear regressions and artificial neural networks) [6,7], and clustering models (e.g., k-means clustering) [8]. Primarily, these methods do not able to provide a potential maximum level of energy efficiency as the benchmark but only compare buildings to each other. Each method also comes with its own set of limitations. For example, simulation models are often inaccurate even after calibration, simple indicators do not account for differences of characteristics among buildings, and point based systems do not have the ability to quantify the extent of energy saving potentials. Statistical approaches require a sufficient amount

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

UrbanGIS'16, October 31, 2016, San Francisco, CA, USA.

Copyright 2016 ACM 978-1-4503-4583-5/16/10 ...\$15.00.

DOI: <http://dx.doi.org/10.1145/3007540.3007541>

of data to achieve significance and are often hampered by the lack of interpretability for categorical variables and the inseparability of random error from inefficiencies. *Stochastic Frontier Analysis (SFA)* calculates the divergence between EUI of the frontier formed by energy-efficient buildings and that of a particular building for benchmarking and could differentiate inefficiency from random errors [9]. Currently, SFA has only been applied to one building type (e.g. commercial) and its applicability to buildings with mixed uses (e.g., buildings with both commercial and residential use) has not yet been demonstrated. Buildings at city scale have significant variance in consumption and use types, making it unreliable to apply the same frontier to evaluate all other buildings across the entire city. Therefore, the objective of this paper is to develop an integrated data-driven benchmarking method that can be applied to data at a city scale and maintain interpretability to facilitate municipal policy making and building manager actions. Our proposed method aims to (1) recursively partition buildings into groups based on total energy use through the classification of building characteristics; and (2) quantify the efficiency level of each building in terms of EUI compared to the energy-efficient frontier in each group for identifying improvement potentials.

3. METHODOLOGY AND DATA

Data is first collected and cleansed before buildings are grouped through non-parametric recursive partitioning, where the feature space is recursively partitioned into rectangular and disjoint groups containing similar observations with reduced variance [10]. The classification and regression tree (CART) algorithm is selected for implementing the partitioning due to its straightforward interpretability and scalability to large datasets. In this paper, the different building characteristics are applied to recursively partition the buildings into groups with similar total energy use. Starting at the root node, CART applies ANOVA (analysis of variance) method to build the regression tree using binary partitions. Building characteristics are examined and a node is selected that could result in the best separation with maximum reduction of variance (mean square error) in total energy use (equivalent to maximizing the inter-cluster sum of square in analysis of variance). At each node of the tree, buildings are partitioned into two groups. The left group contains the buildings that satisfy the characteristic cut-off condition while the right node receives the rest of the buildings. There are two steps to build the CART: growing and pruning. First, the tree is grown with a prior complexity (cp) and the *minsplit* for each node. The cp determines the minimum reduction of mean square error that must be obtained from each partition, and controls the balance between tree size and partitioning quality. Higher cp means more complicated trees with larger number of nodes, potentially resulting in overfitting. The effect of each partition to reduce the mean square error is evaluated as Eq.(1) and Eq.(2):

$$e_j = \frac{1}{N} \sum_{x_i \in R_i} (y_i - \bar{y}_i)^2 \quad (1) \quad e_s = e_j - e_{jl} - e_{jr} \quad (2)$$

Where y is the total energy use and e_j is the mean square error of N buildings. e_{jl} and e_{jr} are the mean square errors for the left and right sub-branch of node j , respectively. If e_s is greater than cp , the partition is accepted, otherwise j is set as the leaf. The *minsplit* determines the minimum number of buildings for each node before attempting a partition. This is done recursively until each node can no longer be partitioned. The tree's performance is evaluated by two measures: *relative error* (equivalent to $1-R^2$), the ratio between the overall mean square error after running the regression tree, to that of a tree with a single root, and *xerror* (x -val *relative error*), the average mean square error of 10-fold cross validation. Second, the resulting tree is pruned back using the one-standard error (1-SE)

rule. The determination of cp , and thus the final number of partitions, are found by summing the absolute minimum *xerror* with its standard error, and then finding the corresponding *xerror* with the fewest number of partitions below this summation. The cp associated with this *xerror* is used for pruning the final tree. The CART is robust to the irrelevant building characteristics and outliers. Both categorical and numeric building characteristics are considered equally. It is able to handle missing values in building characteristics by tracking the hierarchical surrogates (alternative characteristics for partitioning when original data value is missing). The partitioning results are well structured and highly interpretable, allowing municipal officials to easily identify how groups are constructed from the building characteristics.

Within each partitioned group (leaf) of the CART, stochastic frontier analysis (SFA) is implemented to determine the frontier, which represents the potential level of efficient EUI [11]. It is assumed that the frontier is determined by building characteristics, including functional features (e.g., resident use percentage), physical properties (e.g., number of floors), and use patterns (e.g., year built). The difference (error term) in EUI between any particular building and the frontier results from: 1) variations in response to random errors due to system malfunctions (e.g., system out of service), occupancy dynamics (e.g., sick leave), weather changes (e.g., thunder shower), measurement errors (e.g., meter reading error), and input omissions (e.g., missing numbers in building area). These errors are exogenous and unavoidable; 2) sources of inefficiency including poor operations (e.g. lights always on), energy inefficient appliances, rigid system controls, etc. Therefore, the EUI of a given building can be formulated as Eq.(3).

$$y_i = \beta x_i + (v_i + u_i) \quad (3)$$

Where y_i is the EUI of the i -th building and x represents the vector of building characteristics of the i -th building. β is the vector of parameters defining the shape and location of the frontier. v is the stochastic portion of the frontier, accounting for the impact of random errors, and is assumed to be normally distributed as $v \sim iidN^+(0, \sigma_v^2)$. u represents the extent to which the EUI of a specific building exceeds the potential efficiency level. It is non-negative and assumed to have truncated half normal distribution as $u \sim iidN^+(0, \sigma_u^2)$, independent of v . Buildings on the frontier are considered energy efficient ($u=0$) and are used as the benchmark. Cobb-Douglas function is applied to build the relationship between building characteristics and EUI for easy interpretation:

$$\ln(y_i) = \beta_0 + \beta_1 \ln(x_{i1}) + \beta_2 \ln(x_{i2}) + \dots + \beta_n \ln(x_{in}) + (v_i + u_i) \quad (4)$$

If one building characteristic has a value of zero, $\ln(x)$ is replaced by x . Parameters of the stochastic cost frontier are obtained based on maximum likelihood estimates. There are three steps to calculate the coefficient vector β and composite error ($v+u$): 1) obtain the unbiased β from ordinary least square (OLS) estimates; 2) conduct a grid search to calculate γ ($\gamma = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2}$), and update σ ($\sigma^2 = \sigma_u^2 + \sigma_v^2$) according to the corrected ordinary least squares formula [11]; 3) execute the iterative process based on the Davidon-Fletcher-Powell method to get the final maximum likelihood estimates [11]. The level of energy efficiency of a given building, denoted as the *efficiency estimate*, is measured as the expectation of $\exp\{u\}$ according to the conditional distribution of u given ($v+u$). It ranges between one and infinity, where one indicates fully energy-efficient while infinity indicates completely energy-inefficient.

We applied the proposed method to data from New York City. The main dataset comes from New York City's Local Law 84 (LL84), the first major building energy disclosure mandate in the United States. Buildings that are larger than 50,000 square feet and tax lots

with more than 100,000 square feet of combined building area are required to disclose yearly energy usage in compliant with LL84. The data used in this study is for the 2014 calendar year and is coupled with a comprehensive dataset of basic information of building characteristics from New York City's Primary Land Use Tax Lot Output (PLUTO). Initially there were 13,912 buildings in the coupled dataset. Data cleaning was performed to remove the buildings containing missing EUI or total energy use, combine duplicate entries, and revise contradictory and erroneous building characteristics (e.g., year altered is earlier than year built). In total 10,494 buildings were partitioned by the CART. Within each group, buildings with specialized primary uses were then discarded before applying SFA. Since buildings with large residual variations can significantly hamper the performance of SFA, those buildings were considered as outliers and also eliminated.

4. RESULTS AND DISCUSSIONS

An initial test was conducted to run SFA for all buildings before running CART. The log likelihood test result ($\gamma \rightarrow 0$) indicates there is no inefficiency among buildings (either $\sigma_u^2 \rightarrow 0$ or $\sigma_v^2 \rightarrow \infty$) thus no "globally" energy-efficient buildings can be found using a single frontier for benchmarking, demonstrating the necessity to run CART as the first step for partitioning buildings into groups. R was used for the recursive partitioning. Total building energy use was log-transformed to reduce the wide ranges among buildings and to alleviate the positive skew of energy use amounts, in part caused by converting EUI (kBtu per square foot) to total energy use (kBtu). The transformed total energy use has an approximately normal distribution. The CART model was then run to partition the buildings into groups based on building characteristics (*minsplit* = 10) for reducing variance in energy use. Figure 1 shows the *xerror* (*x-val relative error*) of each tree produced when an additional partition is added to CART. All *xerror* values, based on 10-fold cross-validation, were scaled so that the first node has an *xerror* value of 1. The *1-SE* rule (red horizontal line in Figure 1) was then applied to prune the tree back with fewer partitions. The resulting *cp* was selected as the threshold capacity (*cp* = 0.00311 and *error* = 0.27952). Figure 1 also shows the diminishing returns from each sequential partition as the reduction in *xerror* becomes smaller as the tree grows larger.

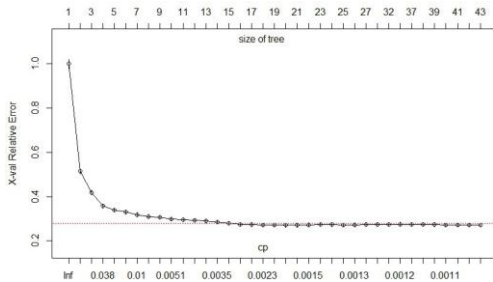


Figure 1. *Xerror* for trees with different *cps* and tree sizes

The final tree is shown in Figure 2, where 16 groups were formed on the basis of 3 building characteristics¹: *DOE_Prop*, *BldgClas*, and *AssessTo*. Figure 2 provides a visualization that is quick and easy to interpret. The *DOE_Prop* appeared in the tree repeatedly, indicating its importance in the partitioning. Most characteristics did not appear in the final tree, such as geographic information, as CART found them to be insignificant in terms of capturing and

reducing the variance in total energy use. The buildings in different groups are with significantly different means and variances of total energy use, empirically demonstrating the significance to partition buildings into separated groups before benchmarking analysis.

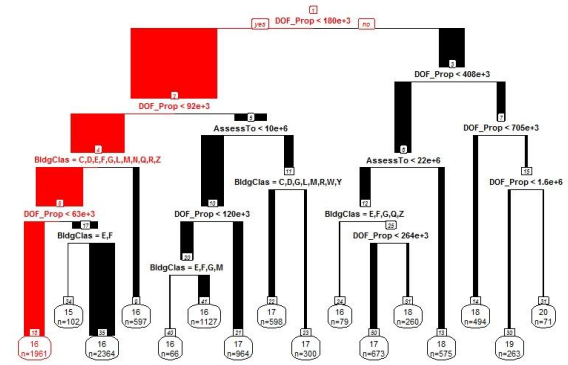


Figure 2. Classification and regression tree (CART)

Preliminary study for SFA was run for one group (red-highlighted in Figure 2). It is the second largest group produced from the partitioning and consists of five types of buildings including walk-up apartments, elevator apartments, loft buildings, asylums and homes, and condominiums. The frontier was constructed resulting in a likelihood ratio ($= -842.13910$) and β parameters that are statistically significant at 1% level, showing its validity and reliability. γ ($\gamma = 0.85569$) is close to the upper bound, indicating that the majority of variations in EUI are from inefficiency sources with a relatively small portion being caused by random errors. This proves the assumption that using building functional features, physical properties, and use patterns is appropriate for defining the shape and location of the cost frontier. The *efficiency estimates* of individual buildings were then calculated based on the prediction of unobservable *u*. A histogram was created to show the approximate distribution of *efficiency estimates*, as displayed in Figure 3. The *efficiency estimates* ranged from 1.03503 to 9.33080, revealing that all buildings are energy inefficient to some extent. The median ($= 1.43657$) is shown as the red dotted line. Although 91% of the *efficiency estimates* are smaller than 2, the histogram has a long right tail indicating there are a considerable number of highly inefficient buildings with a wide range of estimates.

A kernel density estimation (KDE) was then applied, using a Gaussian function as the smoothing kernel, to investigate the underlying distribution (density) of energy *efficiency estimates* (shown in Figure 3 as the red filled in curve). It depicts the probability of obtaining an energy estimate between two values, by taking the integral across the bound, and is further truncated by the frontier, as the *efficiency estimate* cannot be less than 1. The density of *efficiency estimates* has an approximately truncated normal distribution, which is important when analyzing the probability of *efficiency estimates* in finer ranges for better decision-making. To explore energy performance of buildings relative to a threshold, the cumulative distribution function (CDF) was also calculated (black solid line in Figure 3). It has an "S" shape and is also truncated by the frontier at *efficiency estimate* = 1. The generation of the density function and cumulative distribution function further enables the comparisons of overall building energy performance among different groups and even across cities.

¹ DOE-Prop: gross square footage of the property; BldgClas: major use of building; AssessTo: assessed total value of the property.

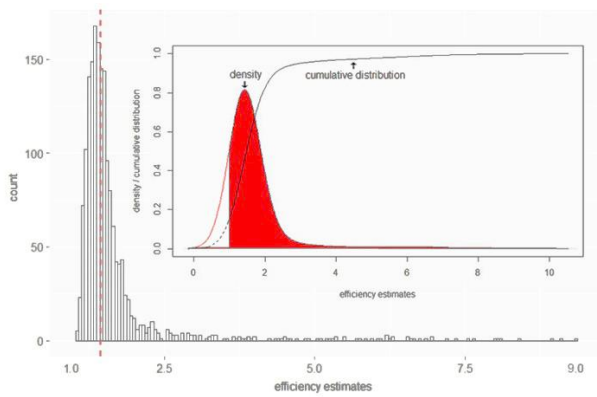


Figure 3. Histogram, density, and CDF of efficiency estimates

Buildings that need interventions and improvements can then be identified according to the estimate of u . The ten relatively most energy-efficient and ten relatively least energy-efficient buildings of this group were geocoded and visualized in Figure 4 below.

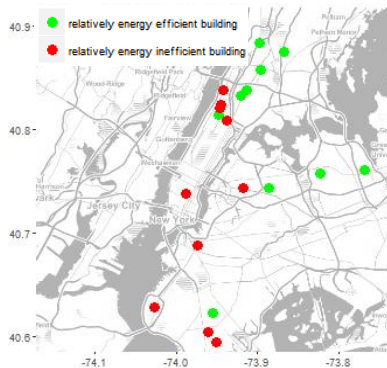


Figure 4. Buildings identified for targeted energy auditing

5. CONCLUSIONS AND ONGOING WORK

This paper proposes a novel data-driven method that integrates recursive partitioning and stochastic frontier analysis for benchmarking building energy efficiency at the city scale. The proposed method first applies a classification and regression tree to recursively partition buildings into groups. A highly interpretable tree is built based on building characteristics, separating the dataset into similar type buildings and lowering total energy use variance. Within each group, a stochastic frontier analysis is used to define the cost frontier, representing the potential levels of efficient EUI attainable, which is then used for benchmarking. The impact of random errors can be successfully separated from that of inefficiency sources. The level of energy efficiency of any particular building is measured by the *efficiency estimate*, which is calculated as the expectation of inefficiency according to the analysis of its conditional distribution given composite error term. Using the coupled datasets from LL84 and PLUTO, the integrated method was utilized in a preliminary study to benchmark the energy performance of buildings across New York City.

The proposed method is capable of evaluating the relationships between building energy use and high-dimensional building characteristics across a whole city and systematically considering the impacts of random errors and inefficiency sources. However, there are some limitations that are being solved in the ongoing work. First, restricted by the data availability, the proposed method is not able to quantitatively allocate inefficiency to individual sources. Detailed data will be collected and made available to better explore

the drivers of inefficiencies. The changes and patterns of benchmarking results will also be analyzed when more granular building information becomes available. Second, real-world validation of the proposed method is necessary to empirically prove its superiority to other leading benchmarking methods. Buildings identified as energy efficient will be tested to ensure that they are actually top performers. Several inefficient buildings will be selected for onsite energy audits to validate the reliability of the calculated *efficiency estimates*. Third, the use and implementation of the method will be introduced to both city officials and buildings managers to examine whether they can precisely interpret the benchmarking results and how they take actions based on them. We plan on conducting a survey for parties of interest to assess whether the proposed method addresses the main drawbacks of commonly applied benchmarking methods in industry and academia (i.e., robustness and interpretability). Overall, the research aims to facilitate data-driven development of municipal policy and programs aimed at improving citywide building energy performance. The proposed solution can be used as a tool to make more informed decisions about where resources should be used by highlighting energy inefficient buildings.

6. ACKNOWLEDGMENTS

The material presented in this manuscript is based in part upon work supported by the US National Science Foundation under Grant No. 1461549.

7. REFERENCES

- [1] Department of Energy. Better Buildings. Updated: 2014. Available at: <https://betterbuildingsolutioncenter.energy.gov>
- [2] Institute for Market Transformation. Map: U.S. building benchmarking and transparency policies. Updated: 2015. Available at: <http://www.imt.org/resources/detail/map-u.s.-building-benchmarking-policies>.
- [3] Filippin C. 2000. Benchmarking the energy efficiency and greenhouse gases emissions of school buildings in central Argentina. *Build. Environ.* 35, 5, 407-414.
- [4] Sartor D., Piette M.A., Tschudi W., and Fok S. 2000. Strategies for energy benchmarking in cleanrooms and laboratory-type facilities. *ACEEE Summer Study on Energy Efficiency in Buildings* (Pacific Grove, United States, August 20-25, 2000). ACEEE, Washington, D.C.
- [5] Lee W.L., Burnett J. 2008. Benchmarking energy use assessment of HK-BEAM, BREEAM and LEED. *Build. Environ.* 43, 11, 1882-1891.
- [6] Wu X., Priyadarsini R., Eang L.S. 2010. Benchmarking energy use and greenhouse gas emissions in Singapore's hotel industry. *Energy Policy*. 38, 8, 4520-4527.
- [7] Yalcintas M. 2006. An energy benchmarking model based on artificial neural network method with a case example for tropical climates. *Int.J.Energy Res.* 30, 14, 1158-1174.
- [8] Gao X. 2013. A new methodology for building energy benchmarking: An approach based on clustering concept and statistical models. Doctoral Thesis. University of Pennsylvania.
- [9] Buck J. and Young D. 2007. The potential for energy efficiency gains in the Canadian commercial building sector: a stochastic frontier study. *Energy*. 32, 9, 1769-80.
- [10] Breiman L., Friedman J., Stone C.J., Olshen R.A. 1984. Classification and regression trees. CRC press.
- [11] Coelli T. 1995. Estimators and hypothesis tests for a stochastic frontier function: a Monte Carlo analysis. *Journal of productivity analysis*. 6, 3, 247-268.