Energy Policy 139 (2020) 111327

Contents lists available at ScienceDirect

Energy Policy

journal homepage: http://www.elsevier.com/locate/enpol

Examining the feasibility of using open data to benchmark building energy usage in cities: A data science and policy perspective

Jonathan Roth [a], Benjamin Lim [a], Rishee K. Jain [a,*], Dian Grueneich [a,b]

[a] Urban Informatics Lab, Department of Civil & Environmental Engineering, Stanford University, United States
[b] Precourt Institute for Energy, Stanford University, United States

ARTICLE INFO

Keywords:
Building energy performance
Benchmarking
Variable selection
Open data
Energy efficiency
Disclosure policy

ABSTRACT

Buildings are by far the largest source of urban energy consumption. In an effort to reduce energy use, cities are mandating that buildings undergo energy benchmarking—the process of measuring building energy performance in order to identify buildings that are inefficient. In this paper, we examine the feasibility of using city-specific, public open data sources in two benchmarking models and compare the results to the same models when using the Commercial Building Energy Consumption Survey (CBECS) dataset, the basis for Energy Star. The two benchmarking models use datasets containing building characteristics and annual energy use from ten major cities. To examine the difference in performance between linear and non-linear models, we use random forest and lasso regression. Results demonstrate that benchmarking models using open data outperform models based solely on the CBECS dataset. Additionally, our results indicate that *building area*, *property type*, *conditioned area*, and *water usage* are the most important variables for cities to collect. Having demonstrated the benefits of using open data, we recommend two changes to current benchmarking practices: (1) new guidelines that support a data-driven benchmarking framework relying on open data and a transparent modeling process and (2) supporting policies that publicize benchmarking results and incentivize energy savings.

1. Introduction

By 2050, the global urban population is forecasted to double in size to a total of 6.4 billion people, increasing the total urban energy demand from about 240 EJ to an estimated 730 EJ (Creutzig et al., 2015). In cities alone, buildings currently consume up to 75% of the total primary energy use and account for 50–75% of carbon emissions, making them the largest energy consuming and pollution emitting sector (Chen et al., 2017; U.S Energy Information Administration, 2016). Such heft also means that savings in this sector have far-reaching effects. Numerous studies have identified an energy efficiency gap, observing a large discrepancy between optimal and actual implementation of energy efficient technologies, whereby consumers and firms fail to make positive net present value energy saving investments (Gillingham and Palmer, 2014; Backlund et al., 2012). Explanations for this gap include market barriers, imperfect information, hidden costs, regulatory failures, and behavioral negligence (Gillingham and Palmer, 2014). With the existing level of policy support, however, two-thirds of economically viable building energy efficiency potential will go untapped by 2035 (Kerr et al., 2017; Capturing the Multiple, 2014).

Recognizing the need for urgent action, many cities are now striving to meet new global initiatives, like the United Nations Sustainable Development Goals (SDGs) and the Paris Climate Agreement, where reducing energy consumption is a primary objective. In August 2018, 19 major cities worldwide pledged to make all new buildings from 2030 on carbon-neutral and to retrofit others to the same standard by 2050 (19 global cities Commit, 2018). Additionally, the Carbon Disclosure Project (CDP) counts 620 cities and 122 regions that have reported climate actions to CDP, while the UN-run Non-State Actor Zone for Climate Action (NAZCA) lists pledges from 2,500 cities and 209 regions (The Economist, 2018). Initiatives by subnational governments to address global energy issues have proliferated, with a large focus placed on building energy use, the largest energy consuming sector.

Reducing the amount of energy buildings use has numerous environmental, economic, and social benefits. Given the high demand from buildings, a slight increase in efficiency can substantially decrease city-wide energy use, translating to increasing fractions of electricity generated from renewables and satisfying tangential government and utility clean power goals. Decreased energy demand can also create large economic benefits by delaying the need to upgrade electric grid

* Corresponding author.
E-mail address: rishee.jain@stanford.edu (R.K. Jain).

https://doi.org/10.1016/j.enpol.2020.111327
Received 4 February 2019; Received in revised form 30 January 2020; Accepted 3 February 2020
Available online 26 February 2020

infrastructure, which has an estimated depreciated value of roughly $1.5 to $2 trillion in the United States alone (Stone and Ozimek, 2010). Reduced energy production emissions, from carbon to particulate matter, due to greater building energy efficiency, can result in significant environmental and health impacts, thereby improving the quality of life for citizens in cities (Markandya and Wilkinson, 2007). The energy efficiency market has produced 2.25 million jobs in the United States and is the fastest growing job sector in energy, accounting for half of the entire energy industry's growth (Energy Efficiency Jobs in, 2018). A report by C40 Cities finds that policies promoting energy efficiency and decarbonizing public transport and power generation could create an additional 13.7 million new jobs in cities and prevent 1.3 million premature pollution-related deaths by 2030 (Gonzales-Zuñiga et al., 2018). According to the International Energy Agency (IEA), implementing widespread global policies to reduce energy waste could result in global energy savings of more than $500 billion per year, yielding twice as much global economic value from the energy used today (Energy efficiency 2018, 2018).

Cities have responded to the benefits of energy efficiency by implementing new energy efficiency policies and programs. In particular, cities are enacting building energy benchmarking policies to quantify building performance, evaluate building energy use patterns, and identify inefficient buildings. Building energy benchmarking refers to the reporting of building energy usage to a governmental entity, where individual building energy performance is calculated—based on the building type and its characteristics—and compared to similar buildings. Reported building energy usage encompasses electricity, natural gas, and other fuel consumption in order to measure total on-site energy use. These policies help cities target poor performers with programs, engender energy efficiency through competition, and highlight energy savings potential. As of 2019, 34 cities and states in the US have such benchmarking ordinances in place, mandating that buildings over a certain size must report and disclose their energy usage (Map U.S., 2017). Over 10 billion square feet of floor space in major real estate markets have been affected by these policies since 2008, when Washington DC became the first U.S. city to implement benchmarking and disclosure legislation (Beddingfield et al., 2018). Fig. 1 shows the increasing penetration of these energy benchmarking and disclosure policies since 2004 and how the total area of Energy Star certified buildings has increased with it (a score of 75 or higher makes the building eligible for Energy Star certification)[1].

Work has also begun to demonstrate the benefits of benchmarking programs. A 2012 U.S. Environmental Protection Agency (EPA) report found that energy consumption decreased by 7% over a four year period for a portfolio of 35,000 benchmarked buildings using the EnergyStar system (Benchmarking and Energy S, 2012). In 2016, results from San Francisco showed an overall drop in site Energy Use Intensity (EUI) of 18.9% from 2009 to 2015 and a 33% improvement in the average carbon footprint for 467 facilities (Kozuch, 2016). In 2017, New York City—who began benchmarking in 2011—issued a study reporting 6% EUI savings after three years and a 14% reduction after four years for 3,710 buildings (Meng et al., 2017). A separate report examining New York City benchmarking estimated cumulative energy savings exceeding $267 million (Seiden et al., 2015). However, these savings were not consistent across the entire building stock; most energy reductions were from office buildings (Papadopoulos et al., 2018).

While such reports show substantial promise for city energy benchmarking programs, numerous challenges exist that are preventing cities from fully realizing the savings potential of such programs. First and foremost, there is a lack of understanding of how new data-driven energy benchmarking models that utilize *open data*[2] compare to the current modeling practice (i.e., Energy Star) (U.S. EPA, 2014). As a result, cities are unable to ascertain the value of these newly emerging methods. Second, there is a lack of consensus among policymakers and practitioners as to what data fields should be collected by cities as part of their benchmarking programs (Hsu, 2014a). For example, San Francisco collects interior building information (e.g., number of units and number of bathrooms), while Boston does not (see Table 1 and Appendix A). Lastly, deriving actionable insights from energy benchmarking data is a non-trivial task and previous work has been limited in its analysis of what policies and programs cities can adopt to enhance the efficacy of their energy benchmarking programs. Given such challenges, this paper aims to help answer three key questions related to city energy benchmarking:

1. Can city wide energy benchmarking be conducted using only *open data* and how do such models compare to the current practice?
2. What benchmarking data fields are important for cities to collect?
3. What policies and programs can cities adopt to enhance the efficacy of their energy benchmarking programs?

The remainder of this paper is structured as follows: Section 2 provides a background on the current practice of energy benchmarking, highlights new emerging data-driven methods, and reviews previous analyses of benchmarking data; Section 3 details the methodology employed in our analysis; Section 4 describes the data sets utilized in this analysis; Section 5 discusses the results; Section 6 provides policy recommendation; and Section 7 highlights key conclusions and implications.

2. Background

2.1. Current benchmarking practices and limitations

Every city in the United States that has mandated building energy benchmarking requires the use of the Environmental Protection Agency's (EPA) Energy Star Portfolio Manager software by those building owners and operators subject to the benchmarking requirements (Mims et al., 2017). Though benchmarking in its current state has resulted in energy savings (Meng et al., 2017), there are several acknowledged shortcomings with Energy Star's current methodology. First, Energy Star requires facility managers to manually input energy data and other building characteristics to generate a score (U.S.E.P. Agency, 2014). This process can take hours, lacks transparency; often leads to faulty, biased, or missing data; and may even require hiring of consultants (Palmer and Walls, 2017). In New York City, benchmarking was estimated to cost approximately $500-$1500 per building in 2013 due to the fees charged by these energy consultants (Hsu, 2014a). Second, Energy Star's models are based on a national database of building data and therefore has a limited capability to capture local conditions. The national Energy Star database is collected through the Commercial Building Energy Consumption Survey (CBECS) run by the US Energy Information Administration (EIA); the last survey data was collected in 2012. The next data collection period is for 2018 but EIA will not release that update until late 2020. The lack of more real-time information and the lags in data release can hinder more responsive policy-making. As a result, energy inefficiencies in buildings may persist for months or even years before they are noticed. Third, this dataset has a limited number of fields restricted to a specific building's energy use. It fails to consider local effects (e.g., weather, building codes, urban heat island effect), overlooks emerging data sources (e.g., advanced metering infrastructure

[1] Several government entities have updated their benchmarking ordinances while others have let their ordinances lapse, therefore accounting for the slight discrepancy in the total number of current ordinances in effect.

[2] We define *open data* in this paper as energy and building data that is publicly available.

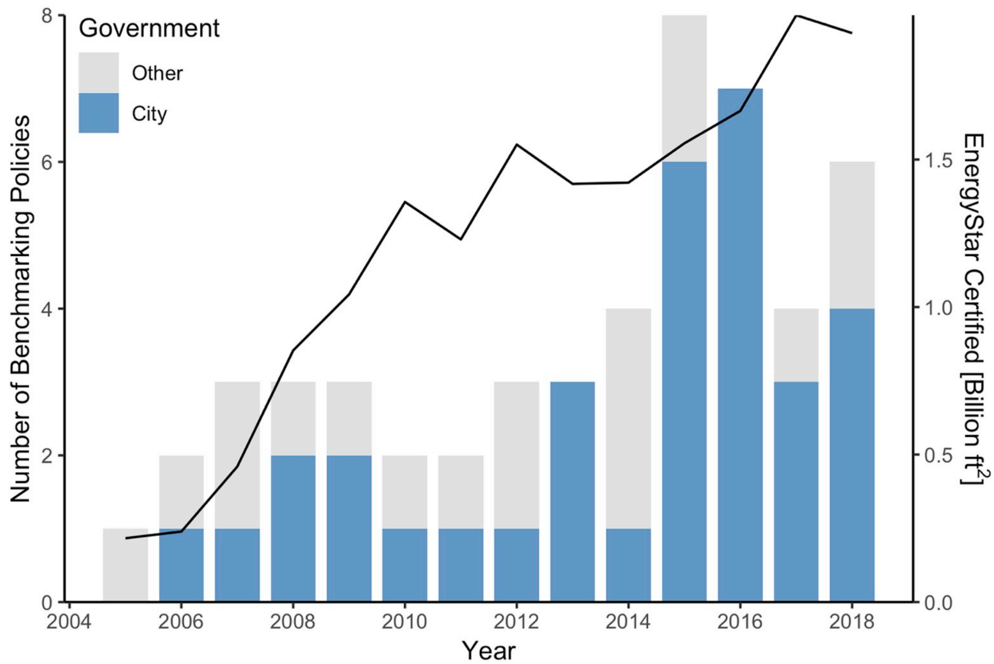

Fig. 1. Trends of energy benchmarking and disclosure policies adoption in the United States since 2004 at the city and other government levels. The gross area of Energy Star certified buildings (line) has risen in tandem with these policies. See Appendix B for the data sources for this figure.

data, remote sensing, public open databases), is limited in building diversity, and is expensive to gather. Fourth, Energy Star's statistical models are based on ordinary least squares regression (OLS) to produce scores. The model is sensitive to outliers, provides a poor understanding of buildings far from the average, and assumes a constant relationship between energy consumption and building characteristics (Chung, 2011). Finally, the interrelated nature of these issues causes them to exacerbate one another.

Since the original use of OLS, several other statistical models have been adapted to suit building energy benchmarking purposes. Frontier models, adapted from the econometrics field, produce benchmarked scores by measuring the difference from current building energy use to a constructed frontier, representing a theoretical maximum established from the input data (Kavousian and Rajagopal, 2014). Data envelopment analysis (DEA) is the most popular nonparametric model, but suffers from a large outlier-effect, the inability to evaluate data not in the training set, and the "curse of dimensionality" - a propensity to overestimate the number of efficient buildings when many features are included (Zhou et al., 2008). Stochastic frontier analysis (SFA) is the most popular parametric model; unlike traditional statistical techniques, it attempts to differentiate random error from inefficiency. These two error terms are still based on the normal distribution, thereby giving SFA similar issues to OLS in benchmarking applications (i.e., a high sensitivity to outliers and an assumed constant relationship between variables) (Buck and Young, 2007). A body of work has also been dedicated to applying machine learning techniques, like artificial neural nets and clustering (Yalcintas and AytunOzturk, 2007; Gao and Malkawi, 2014). However, these techniques require copious data and are difficult to interpret. Given the varied types, uses, and climate zones of buildings, models must maintain interpretability so that building managers and owners have the opportunity to assess proposed sources of inefficiencies and weigh the tradeoffs among efficiency, costs, and utility of their buildings.

2.2. Key variables for energy benchmarking

Due to the limitations of Energy Star, several papers have explored how to identify key variables in predicting building energy consumption (Hsu, 2014a, 2014b, 2015; Ma and Cheng, 2016). In one study, Hsu proposed the use of regularization to perform variable selection and improve prediction accuracy on commercial and multi-family buildings in New York City (Hsu, 2015). In another study, Ma and Cheng performed a similar analysis but utilized other methods (i.e., random forest, Lasso, SVM) to identify influential variables on energy usage of New York City multi-family buildings (Ma and Cheng, 2016). Both studies found several non-building variables to be pertinent predictors of energy usage in New York City. While valuable in identifying key non-building factors that drive energy usage dynamics in buildings, this previous work is limited in its ability to help inform data collection procedures and policies for municipal energy benchmarking programs. Hsu also examined building energy benchmarking more directly by comparing model accuracies using open data and energy audit information, again for New York City (Hsu, 2014a). This study found released benchmarking data to be more useful than engineering audits in explaining the observed energy performance of existing buildings. Further, the study found that only a small subset of variables is needed to provide accurate benchmarking models. Previous works have demonstrated that energy benchmarking can be accomplished using a small number of variables but it still remains unclear if solely using open data sources can provide the needed information. Moreover, a comprehensive look at multiple cities is needed to understand if this trend is generalizable to more regions than just New York City, and to identify the key variables needed for each city.

2.3. Emerging open data-driven benchmarking frameworks

In response to the research outlined above, numerous recent works have focused on incorporating open data into energy benchmarking and predictive urban energy models (Kontokosta and Tull, 2017; Howard et al., 2012). Two recent papers (Yang et al., 2018; Papadopoulos and Kontokosta, 2019) constructed new benchmarking models based on open data from New York City, selecting this city since it was one of the first to both mandate building energy benchmarking for a portion of its buildings and publicly release the collected benchmarking data. Utilizing and combining various open data sources, the first study sought to curb the limitations of SFA by preprocessing the data through a recursive

Table 1
Summary of higher-level variables included in the datasets for the ten examined cities and the CBECS dataset. See Appendix A for a detailed breakdown of each variable name and how features for each city were classified. Every dataset summarized provides site EUI but only a fraction break total energy consumption down into electricity, natural gas, and district steam use. Our modeling therefore utilizes site EUI.

Cities/Datasets	Energy Variables							Other Variables											Number of Buildings
	Source EUI	Site EUI	EnergyStar Score	Electricity Use	GHG Emissions	Natural Gas Use	District Steam Use	Total Areas	Partial Areas	Values	Type	Zoning	Interior Building Information	Building Systems	Water Use	Building Dimensions	Year Built	Other	
New York City	X	X	X		X		X	X	X	X	X		X			X	X	X	14,260
Chicago	X	X	X	X	X	X		X			X					X	X		2,688
Boston		X	X	X	X	X		X			X	X			X		X	X	1,599
San Francisco	X	X	X	X	X			X		X	X	X	X			X	X	X	759
Philadelphia	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	1,477
Seattle	X	X	X	X	X	X	X	X	X		X	X				X	X		3,237
Minneapolis	X	X	X		X			X	X	X	X	X	X	X	X	X	X	X	410
Washington, DC	X	X	X	X	X	X		X			X	X		X	X		X	X	1,455
Los Angeles	X	X	X	X	X			X			X				X				451
London		X			X			X						X				X	23,474
CBECS		X		X		X	X	X			X		X	X	X	X	X	X	6,720
Occurrences in Cities	**8**	**10**	**9**	**7**	**10**	**5**	**2**	**10**	**4**	**4**	**9**	**6**	**4**	**4**	**5**	**6**	**8**	**7**	

partitioning process (using a classification and regression tree) to reduce variance in the data (Yang et al., 2018). Results showed that this open data-driven framework produced more robust results than the conventional energy use intensity (EUI) or Energy Star models. The second study built a model using a variant of gradient tree boosting and showed that the Energy Star models had worse predictive power compared to the proposed open data-driven framework (Papadopoulos and Kontokosta, 2019). Another study used openly available smart meter data from California schools, combined with open data on building characteristics and granular weather data, to produce daily benchmarked scores, thereby decreasing feedback lag for facility managers that want to more immediately understand building performance (Roth and Jain, 2018). These previous works demonstrate that more robust benchmarking models can be constructed using open data sources where higher levels of insight can be gathered when coupled with utility data.

3. Methodology

The goal of our methodology is two-fold. First, we seek to determine if higher model accuracy using open datasets can be achieved when compared to models built upon the CBECS dataset used in Energy Star models. Second, we aim to understand what variables are important for cities to collect as part of their building energy benchmarking programs. We include multiple cities to validate the hypothesis that certain features may be universally important in benchmarking. We use two different statistical models—multivariate linear regression with lasso regularization and random forest—both of which help us examine model fit and understand variable importance. These models were selected for two reasons. First, given that a variety of benchmarking models exist (as discussed in the previous section) we wanted to explore the differences in variable importance between linear and non-linear models. Results from the lasso model can be generalized to other linear models used for benchmarking, such as stochastic frontier analysis (fitted through the maximum likelihood estimate) or quantile regression (fitted using linear optimization techniques) (Filippini and Hunt, 2012; Roth and Rajagopal, 2018). The random forest model captures non-linear effects and results directly correspond to other hierarchical clustering techniques, like classification and regression trees, which have also been utilized in previous works on energy benchmarking (Gao and Malkawi, 2014; Yang et al., 2018). Second, linear regression with lasso regularization is a standard method to perform variable selection in both statistics and machine learning; it avoids issues present in other variable selection techniques, such as stepwise regression, by not relying on a greedy algorithm, therefore avoiding a local optimal, and its ability to efficiently handle datasets with large numbers of features. Though random forest is not typically used directly to select variables, the structure of the model allows for a ranking of variable importance in a natural way that is uncommon for other non-linear models. Further, the random forest algorithm has fewer hyperparameters compared to other non-linear models and is relatively computationally efficient. Perhaps most importantly, both lasso and random forest models have been utilized in previous work in the energy benchmarking field (Ma and Cheng, 2016).

Both the lasso and random forest models are implemented independently on each of the examined datasets. Total building energy use (kBtu) for each building in each city is determined by multiplying the site energy use intensity (EUI) by the square footage area of the building. Before modeling, this variable was log-transformed to account for the wide range and positive skew of energy consumption values, resulting in a log-normal distribution.

3.1. Multivariate linear regression with lasso regularization

Multivariate linear regression is one of the most widely used statistical techniques to model the relationship between a dependent variable and multiple explanatory variables. In order to perform variable selection and regularization, the model fit is altered from the typical ordinary least squares (OLS) method, by adding a penalization term to the cost function, such as the lasso (L1-norm penalty), which sets coefficients to zero (Tibshirani, 1996). The lasso penalization adds an L1-norm term to the cost function with a hyperparameter λ that can be adjusted to modify the amount of penalty added, as seen in equation (1). Here, $\widehat{\beta}^{lasso}$ are the coefficients that are being estimated, y is a vector of building energy use, and X is a matrix with p building characteristics. When λ is set to zero, no penalty occurs resulting in a normal linear regression model that includes every independent variable. As the λ hyperparameter increases, certain coefficients are forced to zero, effectively choosing a simpler model that does not include those coefficients. In order to pick the optimal λ that results in the best fit model (in this case the lowest prediction error), we use 5-fold cross validation over a sequence of λ hyperparameters and select the model with the lowest mean cross validation error. This technique also combats over-fitting by validating the fit of the model using unseen data.

$$\widehat{\beta}^{lasso} = \underset{\beta \in R^p}{\operatorname{argmin}} \parallel y - X\beta\|_2^2 + \lambda \parallel \beta\|_1 \tag{1}$$

In our study, we use the R-package "glmnet" to construct the lasso models and determine the λ (i.e., hyperparameter) that leads to the lowest cross-validation errors. To consolidate the results from each model and report the variable importance for the heterogeneous datasets, we construct a unique feature importance metric. Since the lasso model is searching for the optimal λ that results in the lowest cross-validation score, the lasso parameter is first set to such a high level that no features are included in the model. Then the parameter is slowly relaxed, letting more features contribute to the model, until every feature is added. We observe the order that features are added to the model, as the λ is relaxed, and stop counting features once the λ reaches its optimal level; at this point, only a subset of variables is included in the final model, where we record the order that each selected feature is added as the λ is decreased in value. We then take the inverse of the ordering of the selected features as the variable importance and normalize so that the sum adds to one. This procedure allows us to compare the importance of features between cities.

3.2. Random forest

Random forest is an ensemble learning technique developed to balance the Bias-Variance tradeoff that many models face. Rather than producing a single model which may suffer from high variance leading to overfitting or a simplistic model which suffers from high bias leading to underfitting, random forest produces many regression trees and averages their results. The many constructed regression trees are simple and highly interpretable (weak learners) and can be used to determine important features relevant for an energy benchmarking model (Breiman, 2001). By themselves, regression trees typically have poor predictive performance; however, random forest aggregates the results across many independent trees to improve the predictive performance by constructing each tree on values from a random vector (of observations and features) sampled independently and with the same distribution for all trees in the forest. This bootstrapping decorrelates the individual trees and reduces the variance by averaging the results. Since each tree is built using about $1 - e^{-1} \approx 2/3$ of the data, the error of the model can be calculated using the remaining unseen data using the Out-Of-Bag (OOB) estimate. This OOB estimate acts as a type of cross-validation which can occur in parallel with the training step and helps ensure that the model is not being overfit by validating the model fit on unseen data (Svetnik et al., 2003).

In our study, we use the R-package "randomForest", which implements Breiman's random forest algorithm for regression (Liaw and Wiener, 2002). We utilized the default hyperparameters for the model, generating 500 trees and using one-third of the features for each sub-model. We choose the default hyperparameters since this model is known to have high performance without tuning (Svetnik et al., 2003). The model calculates the feature importance by substituting a vector of noise for each feature and measuring the amount the error increases. The resulting increase in error can be interpreted as the amount of decreased error that particular feature produces. In order to compare variables across cities, we normalize the resulting feature importance for each city independently by dividing each feature's reduction in MSE by the summed reduction in MSE for every feature in that given city. Similar to the lasso model results, this procedure provides values for features that sum to one for each city, where the larger the number, the higher the importance.

4. Data

This paper utilizes open data released from benchmarking ordinances from ten different cities—9 in the United States and 1 in Europe. The cities are: New York City, Chicago, Boston, San Francisco, Philadelphia, Seattle, Minneapolis, Washington DC, Los Angeles, and London. The data all comes from calendar year 2016, except for Minneapolis which is from 2015 and London from 2010. Tax assessor databases with additional building characteristics were also collected where available (New York City, Boston, San Francisco, Philadelphia, and Washington DC) and appended to the energy benchmarking data based on geographic coordinates. These sources provide additional building data, such as assessed property value and number of units, that may be useful for benchmarking models. Both the energy benchmarking and tax assessor datasets are released annually and come formatted as tabular data that can be easily downloaded as.csv files. See Appendix B for the names and sources of all the datasets used in this study. Merging the tax assessor and energy data tables based on raw addresses proved to be cumbersome and inaccurate, while merging based on geocoded coordinates provided better results. Geocoding was accomplished by utilizing the ArcGIS REST API. Given the wide range of features names and the subtle difference between what these features are measuring, we have classified city-specific features into ten higher-level variable names. We utilize the term *variables* in the rest of this paper to indicate this higher-level abstraction. The variable abstractions are used to better organize the features present in each city dataset and capture the underlying characteristics of the buildings that may be important. Table 1 shows the types of higher-level variables included for each city. See Appendix A for a detailed breakdown of which features are in each higher-level variable.

We then cleaned the datasets by removing irrelevant building-identifying features (such as address) and other features that had missing values for greater than 40% of the buildings in the dataset. This threshold was selected because it is important to remove features with great amounts of missing data before imputations; otherwise the imputing process will heavily bias the results of the models. We also eliminated any building with missing site EUI since we did not want to impute values for the dependent variable that we are trying to model. For the remaining missing values, we imputed the data by generating multiple imputations by Gibbs sampling using classification and regression trees (done using the MICE package in R) (Doove et al., 2014). This method employs multivariate imputations by chained equations to impute missing values based on the observed values for a given individual data point and the relations observed in the data for the remaining data points, assuming the observed variables are included in the imputation model. Table 1 includes the size of each of the datasets after cleaning. The number of buildings for each city ranges from 23,474 buildings in London to 410 buildings in Minneapolis. This spread can affect how the models perform, because some models are better able to handle less data than others. The size of the buildings in the final datasets also have a wide range. The released data for each city depends on local energy disclosure mandates, which range from a minimum building size of 20,000 to 50,000 square feet. Several datasets, like New York City, include some buildings with smaller footprints due to either the inclusion of government buildings or building owners opting into the program. Many cities also include massive buildings that can be over 2 million square feet, like the Empire State building in New York City or the John Hancock Center in Chicago. In other words, the datasets are not homogenous and exhibit a wide range of building numbers and sizes.

In addition to the collected city datasets, we also used the CBECS dataset (Commercial Buildings Energy, 2016) which is collected by the Energy Information Agency (EIA) about every seven years. The EIA administers an extensive survey that is sent to building operators throughout the United States and then collects, cleans, and imputes the data, creating one of the most comprehensive datasets on buildings available. The CBECS dataset contains detailed information for 6,720 buildings, including dozens of building characteristic features, ranging from floor to ceiling height ratio to type of primary heating equipment installed. Unlike the other city datasets that we collected, this dataset covers a much larger geographical area and provides much more detailed information about building systems, which are critical components for energy consumption. We selected this dataset because it is used to construct the Energy Star models and provides an upper-bound of features that could be collected for city buildings.

5. Results and discussion

5.1. Open data-driven benchmarking vs. current practice

We analyzed how our two models performed on open data sets for each of the ten cities and compared it to the performance on the CBECS dataset, which is utilized by Energy Star and represents the current most widely adopted practice. Results of the mean square error (MSE) are provided in Fig. 2. The MSE is the average squared difference between the estimated values from the model and the actual values. The lower the MSE, the better the model performs and the more explanatory power it has. For every city, both the lasso and random forest models selected only a subset of the inputted features—which is discussed in more detail in section 5.3—indicating that each dataset included a number of superfluous features. For CBECS, we modeled energy consumption using the whole dataset and independently for each of the nine census regions in CBECS—the most granular geographical area defined. Using the entire CBECS dataset, we got errors of 0.59 for both the lasso and random forest models, while the average MSE for each of the nine regions was 0.901 for lasso and 0.840 for the random forest model. In Fig. 2, the model results for the CBECS census region encompassing each specified city is displayed under the results for the models built on the open datasets. The 10 cities examined are located in 5 of the 9 census regions from CBECS. The error for the lasso and random forest models on the CBECS dataset is higher for nine out of the ten cities examined. Given that CBECS is the dataset EPA uses to construct the widely used Energy Star models and with the number of features in the dataset, it is surprising to see this dataset results in the worst fit model. However, other studies have shown similar poor fitting results when constructing models using the CBECS dataset (Robinson et al., 2017).

In order to understand the robustness of our results, we also constructed models on each building type independently in the CBECS dataset (much like Energy Star's current practice) and the results were even worse. The average MSE across the twenty building types was 0.950 for the lasso model and 0.794 for random forest, showing that modeling the building types independently does not lead to improved results. Overall, the results indicate that data-driven benchmarking models built on open data can outperform the current practice (i.e., Energy Star) in explaining the variability and energy dynamics of an

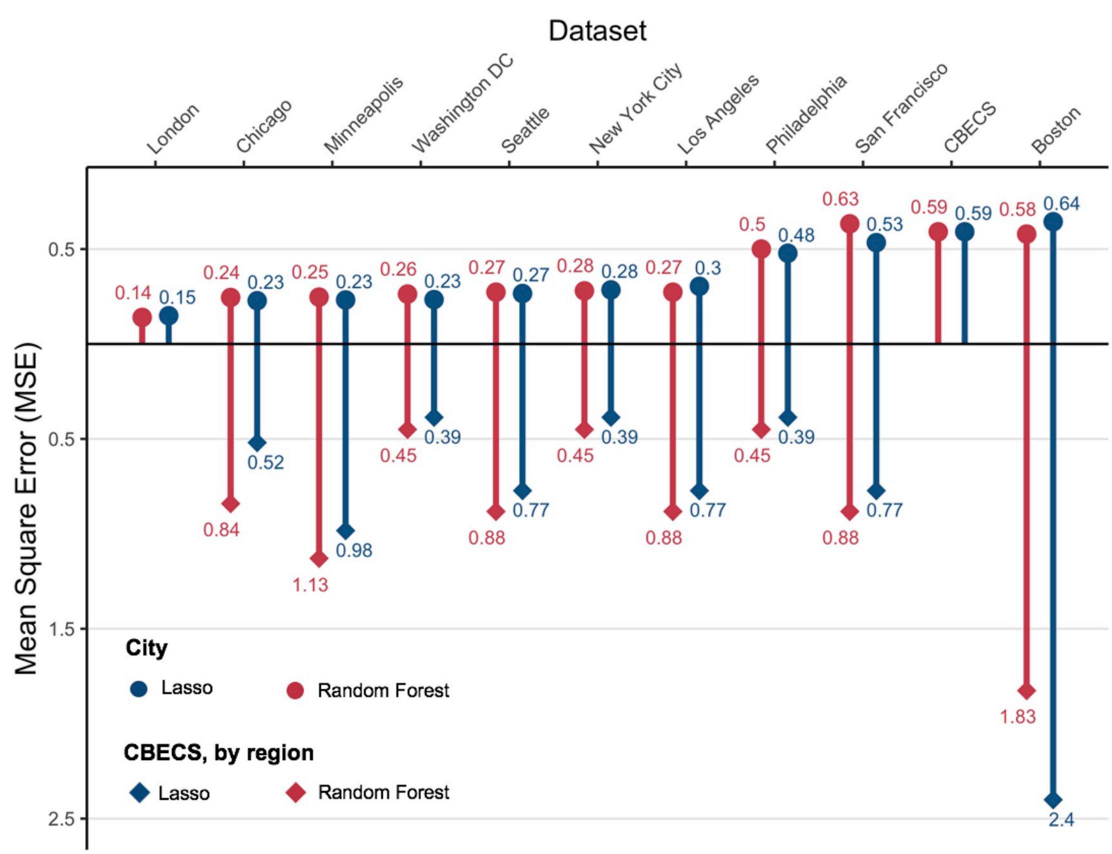

Fig. 2. Summary mean squared error (MSE) statistics for the lasso and random forest model implemented on each dataset. In addition to modeling the energy consumption for every building in the CBECS dataset at once—with errors of 0.59 for both the lasso and random forest models—we also divided the CBECS dataset into census regions and modeled each of these independently. The "CBECS, by region" results are shown on the bottom half of the figure and are represented by diamonds. Results based on the open datasets were better than the models constructed on the CBECS dataset for nearly every city in our dataset.

urban building stock.

We postulate that there are several possible explanations for the inferior results when using the CBECS dataset. First, the dataset fails to account for local effects that can influence building energy consumption. Although CBECS includes several location and weather related features, attempting to compare buildings that span the entire country (or even a large region) with varying climates, architectural designs, building codes, urban densities, etc. is a difficult task. Further, the urban heat island effect alone can cause temperatures in cities to rise by several degrees resulting in much higher cooling and electricity consumption demands (Li et al., 2018). Second, the CBECS dataset may not have enough buildings to account for numerous types of geographical variations in buildings. Although several of the cities have small datasets, the geographical variation in a city is much smaller than that over a country or region. Third, with the large number of features collected in the CBECS dataset, many missing data entries are known to be imputed by the collectors and thus further bias the dataset. Overall, these underlying issues of the CBECS dataset underscore that open data-driven benchmarking frameworks are viable alternatives to Energy Star and have the potential to enable the creation of more nuanced and detailed benchmarking models.

The results also show that the random forest model performs similarly to the lasso model. The models for Boston and San Francisco were the only two cities that had poor performance similar to the CBECS results. Both these cities had more missing data points than the other cities, resulting in the removal of more data entries and imputing a greater portion of the dataset. However, building energy benchmarking requires a model that is interpretable and separate building inefficiency from modeling error. Otherwise, a model that only achieves a low error may be adequately predicting energy consumption habits of both high and low efficiency buildings, but may be unable to tell them apart. Reducing the error of benchmarking models should not be the ultimate goal, but rather should be considered in tandem with the objective of constructing a model that best captures energy performance and efficiency opportunities.

5.2. Data field importance

5.2.1. Lasso model results

Fig. 3 shows the summary results for the data field importance of all building variables for the lasso regression model. Certain variables do not show up in this figure, not because they are not important, but because each city contains a different subset of variables as summarized in Table 1. The order that the results are presented in correspond to the MSE value as summarized in Fig. 2. London had the lowest MSE, so it is displayed first in Fig. 3 while Boston had the highest MSE and is displayed last. The lasso regression model exhibited very high errors for every dataset when the λ parameter was set to zero (providing the same results as an ordinary least squares regression), confirming the need to eliminate certain features from the dataset. In order to compare variable importance between cities we normalized the resulting values so that they added to one. For example, in Chicago we observed that three variables are selected (*Total Areas, Type, and Building Dimensions*). Since *Total Areas* were selected first, we took the inverse of one (equal to 1.0) as its importance; *Type* was selected second so we took the inverse of two (equal to 0.5) as its importance while *Building Dimensions* was selected third so we took the inverse of three (equal to 0.333) as its importance. Each importance value was then divided by the summed values (i.e., equal to 1.833).

Fig. 3 shows that a different subset of variables are important for each model, indicating that accounting for local effects is essential considering that nearly every city was able to produce a better model than the CBECS dataset. However, there are several consistently important variables shown across the cities. For example, *Total Areas* is the most important variable for all cities. The CBECS dataset shows this variable to be relatively less important than the other examined cities which may be due to the heterogeneity in the locations of the buildings in the CBECS dataset. *Type* is also quite important for every city, followed by *Water Use*. These three variables (*Total Areas*, *Type*, and *Water Use*) should be prioritized in any data collection efforts in order to increase model performance and create more robust benchmarking models. Fortunately, these variables, except for *Water Use*, are already often collected by city tax bureaus. Some variables are less important to collect. *Interior Building Information*, such as number of bathrooms,

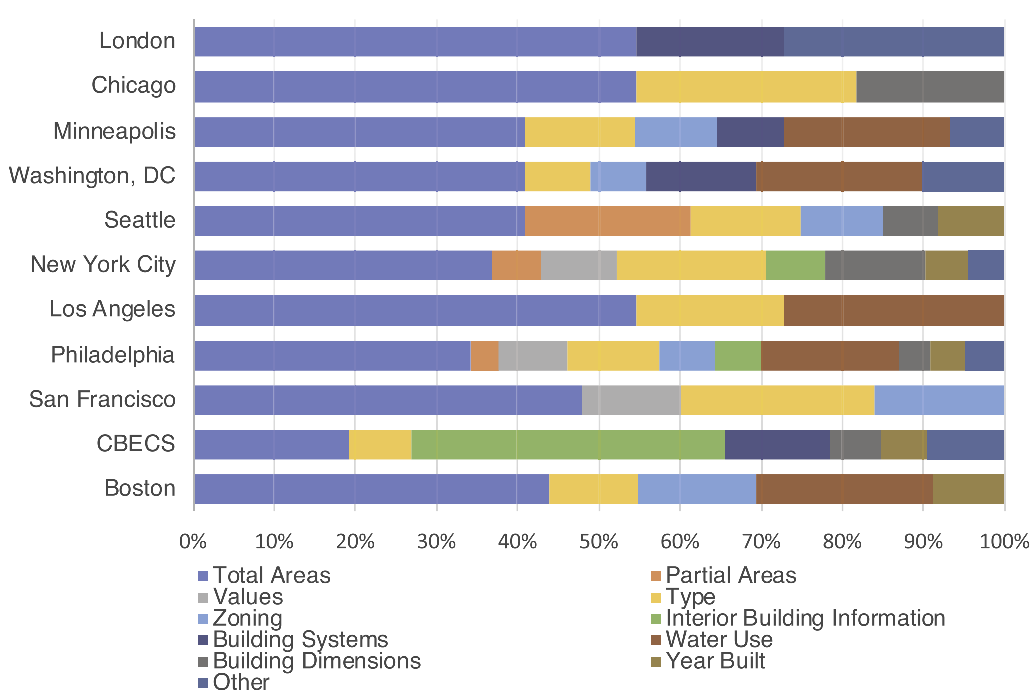

Fig. 3. Summary results for the variable importance of all building variables for each dataset for the lasso regression model. Each variable is represented by a different color. The order of the bars correspond to the MSE for the lasso models as summarized in Fig. 2.

bedrooms, escalators, units, and permits, do not contribute to increasing model performance. Additionally, the variable *Year Built* is only important in less than half the cities despite being present in nearly every dataset, as shown in Table 1. Only the CBECS dataset shows this variable as important; however, as discussed above, this dataset resulted in one of the worst performing models. We postulate that this variable is not very important for benchmarking analysis because buildings often undergo many retrofits throughout their lifespan which can bring them up-to-code and/or significantly alter the equipment and amenities that drive energy usage.

5.2.2. Random forest results

Fig. 4 shows the data field importance for each city examined for the random forest model. Similar to the lasso model results, each city has a unique set of important variables. Again, the differing variables deemed important may be partly explained by the local contexts of each city which is being overlooked by the Energy Star models; however, there are several variables that do seem to be important across every city. *Total Areas* is again one of the most important variables for the random forest model, followed by *Type* and *Water Use*. Likewise, *Year Built* does not seem to be very important, indicating that cities do not need to prioritize collecting this variable as part of their energy benchmarking process. Notably, the *Partial Areas* variable is also observed to have a larger role for fitting the model when *Water Use* is unavailable. We postulate this result occurs because water and energy use scale with human based services, which is captured in the *Partial Areas* variable. Specifically, variables indicating the amount of unconditioned space—such as garage area, storage area, etc.—are likely to have less energy and water consumption compared to a similarly sized hotel or apartment space. This variable may act as an important substitute for collection if water consumption values are difficult to obtain. Further, rather than collect a detailed breakdown of room types in buildings (i.e., *interior building information*), it is much more important to collect information on building type and how much of the building is conditioned (i.e., heating and cooling) versus unconditioned (represented through partial areas); this information provides a better understanding of how much energy is being consumed by building systems to maintain environmental conditions in these spaces. Further, CBECS contained numerous variables that fall into this category, giving this category a high bias in this dataset.

5.3. Discussion of results

Overall, results demonstrate the feasibility of utilizing open data to construct robust data-driven urban energy benchmarking models and that only a few key variables—*Total Area*, *Type*, *Partial Areas (or conditioned area)*, and *Water Use*—are needed for such models. Moreover, by employing open data approaches cities can reduce or even eliminate the current system of relying on Energy Star whose use involves manual efforts, leading to high expenses and inaccurate inputs. The manual data entry process could further worsen the presented results of the benchmarking process, but this effect is difficult to measure because this data is not available to the public. Employing two different methodologies, two other recent studies also found benchmarking results using New York City's open data to be more robust than results obtained from Energy Star, corroborating our findings (Yang et al., 2018; Papadopoulos and Kontokosta, 2019). Cites are already collecting most variables that proved to be valuable—like building area and property type—as part of their tax collection services. Including these valuable data variables into benchmarking programs for a large number of buildings represents a data integration problem, that has been studied by previous work, and is therefore addressable (Chen et al., 2019).

The analysis in this section surrounding important variables to collect is highly dependent on the type of data used. The CBECS datasets included many more features under the *Interior Building Information variable* and *Building Systems variable* than the city datasets. These two variable classes, which are known to affect building energy consumption, were deemed as important by the lasso regression and random forest models, respectively, but both models for this dataset showed worse results. It may be the case that certain features in these variable classes are indeed important for modeling building energy consumption, but due to the vast geographical coverage of the CBECS dataset, their effect has been diminished such that the overall fit of the models is still poor. These building system related features are difficult to collect—as exemplified by the extensive and costly survey needed to construct the CBECS dataset—especially at the large scale required for building energy benchmarking, which requires data from hundreds of buildings.

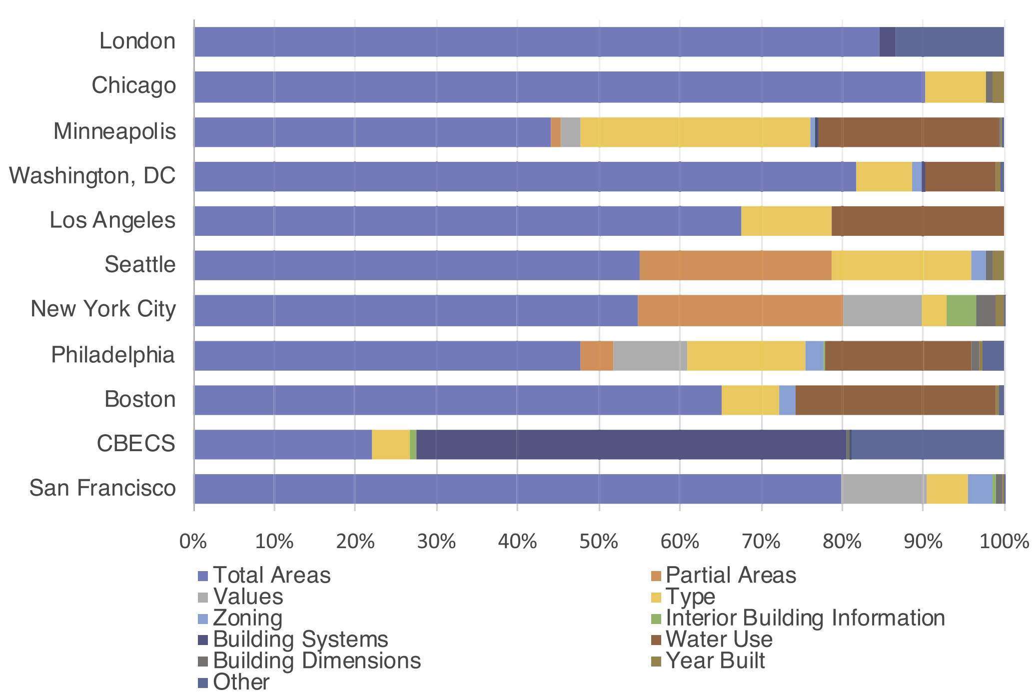

Fig. 4. Summary results for the variable importance for each dataset for the random forest model. Each variable is represented by a different color. The order of the bars correspond to the MSE for the random forest models as summarized in Fig. 2.

Emerging data sources, however, hold promise for further automation of data collection, especially for variables that are currently difficult to obtain. Such variables that proved to be useful in the CBECS dataset—like work hours (contained in the *Interior Building Information* variable: See Appendix A)—could theoretically be obtained from services such as Google Maps that keeps data on store hours. Furthermore, improvements in remote sensing could also lead to variables, such as area and height, that could be collected automatically rather than using information from tax assessor databases.

6. Conclusion & policy implications

Our research described above supports an expanded approach to benchmarking—one that relies upon existing, open data sets already collected by cities and modeling methodologies that can be automated to avoid the expenses and inaccuracies of Energy Star. Our research above identifies which data cities should collect and how to analyze this data in order to improve the effectiveness of energy benchmarking policies. However, this research information alone does not guarantee changes in benchmarking approaches, improve market efficiency, or ultimately, enhance energy savings. The information gained from the analysis in section 5 must be incorporated into policy structures that heighten the usefulness of benchmarking and drive progress on multiple fronts.

In this section we address two aspects of how to enhance the effectiveness of benchmarking in achieving energy savings through development of better benchmarking policy approaches.

1. **The adoption and use of an open data-driven benchmarking framework.** With the numerous stakeholders affected by building energy benchmarking and its broad scope, all-around transparency will reduce the friction of implementation and enable quicker progress. Our research shows there are three critical areas for improvement in this area: use of open data for benchmarking; use of a transparent modeling and scoring methodology; and use of a public rating system.
2. **The adoption of complementary policies and programs to increase building energy efficiency that are linked to benchmarking results.** Examples of complementary policies include mandatory audits, retrofits and/or retro-commissioning; provision of incentives such as utility rebates, loans and/or technical assistance; and linkage of benchmarking programs to city/utility energy efficiency and carbon reduction goals. Integration of varied policies such as requirements on improving building performances, incentives and technical assistance, and linkage to goals will solidify energy benchmarking's role in society.

In other words, without proper and transparent measuring and use of complementary building policies, management of building energy will not improve. This section explores how policy mechanisms can further enhance the efficacy of data-driven energy benchmarking programs.

6.1. Open data-driven benchmarking framework

Transparency enables easy access to information that is necessary for markets to work and stakeholders to act. For example, cars have MPG ratings, food products have nutritional labels, and appliances have energy use tags for comparison; each of these examples provides instant direction to people so that they can assess costs and information. For buildings, transparency allows interested parties to assess energy performance, efficiency potential, and historical trends. Building performance information is a potential driver for change and a necessary tool to measure actual outcomes and hold parties responsible for performance (Cohen and Bordass, 2015). Measuring performance is particularly pressing with building energy consumption since a significant gap still remains between designed and actual performance (Menezes et al., 2012). This mismatch is most striking for LEED certified buildings which have been shown to have only a weak correlation with reduced energy consumption (Newsham, 2009). Increasing accountability and transparency will aid the construction industry, property markets, and engineers by properly awarding those parties that build to design specifications. Increasing feedback on energy performance enables better policy-making, quicker identification of inefficiencies, and an improved ability for the market to understand which interventions are effective. Below we describe three levels of transparency, each with its own benefits, that should be incorporated into benchmarking policies and programs.

1) Use of Open Data. Energy benchmarking policies, and the underlying models they are built on, are nearly impossible to make transparent and reproducible without public access to the data used in the models. A first step is releasing data publicly, to ensure accurate use of the data and confidence in the benchmarking results. If governments can work with utilities to automatically access energy use data for buildings, the entire benchmarking system could be automated by using other publicly available open data, such as those highlighted in this study. The GreenButton initiative is one potential solution to accessing building energy data which eliminates the need for error prone, self-reported data (Sayogo and Pardo, 2013). A second benefit of open data access is that it will spur more effective collaboration across the science-policy boundary, reducing parallel efforts and duplicated work (Pfenninger et al., 2017). Such collaboration can save a tremendous amount of time and money, freeing up researchers to explore new questions and giving officials more time to build effective policies. In large datasets used in government decision-making, traceability and referencing can become major problems because civil servants are generally not trained as scientific researchers. The burden of this work can be reduced with open data that is shared more widely across academia, government, and the public while improving quality assurance (Strachan et al., 1038). A third benefit, described in our research above, is that the data sets most relevant to accurate benchmarking may already be publicly available (e.g., tax rolls) and thus their automated use for benchmarking may be relatively inexpensive and easy. Cities with a more systemized data collection process could reduce the errors caused from the current manual data entry process required by Energy Star.

Fortunately, several cities have already exhibited methods worth emulating for releasing data to the public. For example, the City Energy Project has partnered with 20 cities and counties across the U.S. to share data with the public to help motivate action and has published their best practices on their website. In this study, we leveraged open datasets that come in a tabular and organized format (i.e., .csv files) and which are released on an annual basis, therefore making this analysis easier than if the data were in a much more difficult to extract format, such as pdf files. Accessibility of data, from data portals to file types, and periodic release dates makes data more valuable because researchers and the public can analyze trends over time and reduce the friction for getting the data prepared for analysis. Open data portals from various city governmental organizations websites (e.g., data.boston.gov) are one-stop-shops that centralize and make accessible all the datasets that are released from one city; without these portals, knowing where to get data becomes a hurdle. Individual libraries are getting involved with opening data by curating, hosting, and maintaining these open data portals, promoting the idea that the data is part of the public domain with no restrictions to access (How libraries can make, 2018). Conceptually, libraries are predisposed for this role because they are often the most trusted local institutions whose responsibilities align with the goals surrounding open data; they are advocates for literacy, engagement, and equitable access to public information. Typically, these open data portals include datasets on building permits, crime incidents, city employee earnings and property assessments, among others. In the context of this energy benchmarking study, all of our variables are taken from open data portals run by city governments. However, we note that data availability and data quality vary significantly from city to city and remains a challenge for the municipal open data movement.

2) **Transparent Model Construction and Scoring Methodology.** Details behind current benchmarking models are also largely private, which prohibits researchers from improving the methodologies and can cause the public to distrust the results. Although the CBECS data set is publicly available, the feature selection process for the Energy Star model is private, thus making it difficult to verify the validity of the model. Additionally, the hidden nature of the modeling process and unreported model fit (i.e., modeling error) creates distrust and does not allow for public scrutiny or improvements to the methodology, thereby limiting identification of specific sources of energy waste in buildings. Providing interpretability is key for policy-makers, facility managers, and building owners to better target sources of inefficiency. With the numerous types and uses of buildings, understanding the drivers of energy consumption in each building is essential since the notion of building energy efficiency is subjective; a building with no windows may be very energy efficient but is a dismal environment for tenants. It would be unfair to penalize a gym for having a spa, an energy intensive amenity, and even more senseless to suggest that the owners remove it since this is a central service to the business. However, owners and building operators should be aware if their spa is more inefficient than others, with potential for overall building energy savings. With an open benchmarking framework, building owners and operators, as well as public officials, can track building performance and identify specific areas for improvement through remote or on-site audits.

3) **Public Rating System.** Benchmarking building performance without distributing the scoring results is unlikely to lead to any savings. Numerous stakeholders have a vested interest in building energy performance. Table 2 summarizes those stakeholders, their functions and needs, and how an open benchmarking framework can benefit them. Reliable assessment of building performance and rating information can help stakeholders to save money, comply with laws, reduce lending and insurance costs, increase visibility, and improve accountability. Increased information to consumers can positively affect the behavior of stakeholders and the efficiency of markets, which has led economists to support policies that increase information availability (ZheJin and Leslie, 2003); markets cannot value what is not measured. Requiring a scorecard to be placed in a visible location in buildings, analogous to health inspection ratings for restaurants, can catalyze and encourage market transformation by adding a "green premium" for high performers. Such visible ratings for restaurants led to improved food-safety practices and garnered high program approval ratings, suggesting that a similar system for energy building benchmarking can drive savings through increased recognition (ZheJin and Leslie, 2003; Wong et al., 2015). With regard to building energy efficiency, a study examining 1100 leasing transactions in the Netherlands found that buildings designated as inefficient had rental levels that were 6.5% lower compared to efficient, otherwise similar buildings (Kok and Jennen, 2012). Currently, New York City and Chicago both require buildings to post their benchmarking rating in a prominent location on the property and to share this information at the time of sale or lease. We recommend implementing similar mandates of publicly displayed benchmarking scores so that clear and reliable information can aid market forces. Additionally, providing access to all scores through a computerized database maintained by the relevant governmental entity in charge of the benchmarking program would improve access to benchmarking information and results. These changes aim to improve visibility and transparency of the reported information, make energy use information easily accessible to residents, and encourage adoption of energy saving practices.

Table 2
Summary of key stakeholders, their needs, and the aspects of the open benchmarking framework that they will benefit from.

Stakeholders	Function/Needs	Open Benchmarking Framework: Benefits by Category		
		Data	Modeling	Rating
Landlords	Reduce operating expenditures			X
Tenants	Compare buildings to assess utility costs			X
Investors	Deploy capital efficiently		X	X
Lenders	Evaluate eligibility of lendee		X	X
Policymakers/ Regulators	Track compliance of buildings; target incentive payments; assist in setting efficiency goals			X
Insurers	Properly assess risk	X	X	X
Facility Managers	Receive feedback on performance and identify savings opportunities	X	X	X
Energy Service Companies (ESCOs)	Reduce customer acquisition cost		X	X
Utilities	Target inefficient buildings	X	X	X
Lawyers	Consider impacts on real estate, construction, and building industries		X	X
Engineers	Develop better building design performance models and software	X	X	X
Statisticians/ Researchers	Convert collected data into actionable insights	X	X	X

6.2. Complementary policies to benchmarking

Achieving high savings requires policies that go beyond adoption of a benchmarking ordinance. Policies that target the increased visibility of benchmarking results in accessible and effective ways can help markets better value the energy performance of buildings. We have identified three types of complementary policies that can be based on benchmarking results to further drive energy savings.

1) **Mandatory Building Audit, Retrofit, and/or Retro-commissioning Requirements**. A first category of complementary policies is the requirement for building owners or operators to improve building performance, starting with an audit and then proceeding to retrofits and/or retro-commissioning (RCx). New York City, for example, requires benchmarked buildings to undergo an energy audit and RCx process every 10 years to ensure building systems are operating as intended. Austin, TX has a similar program that requires audits of single-family homes prior to sale and multifamily homes every 10 years. Several other cities have comparable programs which are summarized by Palmer and Walls (2017). Though costlier than just benchmarking, RCx can lead to substantial savings and returns. The Department of Energy (DOE) estimates most buildings save between 10 and 30% of energy associated with RCx where savings persist for 3–5 years (Katipamula et al., 2012). A study from Lawrence Berkeley National Lab identifies RCx as arguably the single most cost-effective strategy for reducing energy costs and GHG emissions (Mills and Mathew, 2009). Other potential requirements include mandated audits for poor performing buildings at regular intervals. Energy audits of over 800 buildings in San Francisco revealed \$60.6 million in opportunities for cost-effective energy efficiency investments, with a net present value of \$170 million (Hart, 2015). Buildings often undergo retrofits regularly, though these retrofits only focus on energy efficiency a fraction of the time. Previous research has shown that energy efficiency opportunities need to be identified in the early stages of renovation when building owners are thinking about ways to improve their building (Pettifor et al., 2015). Requiring both benchmarking and periodic energy audits can bring this information to the attention of owners and building operators, thereby leading to higher likelihood of implementing efficiency retrofits.

2) **Financial Incentives and Technical Assistance**. Utility customer funded energy efficiency programs in the U.S. make available

billions of dollars and technical assistance for building improvements. Yet, because building benchmarking programs are run by local officials, unless the local government is also a municipal utility, the benchmarking program may not be structured to provide building owners and operators with seamless access to these utility-run programs. Utility regulators, such as the state public utilities commission, could require utilities to work with local governments in the utility's service area to target incentives and technical assistance to the lowest performing buildings in benchmarking programs. Likewise, local governments themselves could reach out to their local utility to develop a voluntary assistance program. And, local governments, even if not a municipal utility, could develop assistance programs independently. For example, the New York City Retrofit Accelerator is a program that sends report cards to building owners with information about relevant utility programs. Other incentives could allow buildings that receive scores below a certain threshold be eligible for "green loans" offered by the city, with reduced interest rates, or other types of project financing. Cities could also provide tax exemptions, expedited permitting reviews, reduced fees, density bonuses, or administration variances to allow for additional yard setbacks, landscape buffers, and driveways. Equity benefits could be explored through the use of absolute (e.g., buildings that achieve a score below a predefined threshold) and/or relative (e.g., buildings that have achieved a certain level of improvement) cut-offs for incentives or other assistance.

3) Linkage to Efficiency and/or Climate Goals. Growing concerns over energy demand and the climate has spurred numerous cities to join initiatives such as the City Energy Project, a joint project with the NRDC (Natural Resources Defense Council) and IMT (Institute for Market Transformation) and the DOE Better Buildings Challenge. Similar initiatives led by the UN have encouraged cities worldwide to strive to reduce energy use, by promoting energy efficiency in buildings. Likewise, many states have set efficiency goals for their regulated utilities. Building energy benchmarking can encourage buildings to lower their energy consumption and simultaneously inform progress on both local and global sustainability goals.

6.3. Conclusions

Building energy benchmarking is increasingly seen as a low-cost, impactful practice that can help cities reduce the energy demand of their buildings. Cities and other stakeholders are striving to better understand how buildings use energy and which ones are highly inefficient so that policy-makers and others can make informed policies and investments.

The emergence of new data sources has made data-driven benchmarking more powerful. Current benchmarking practices rely on the opaque Energy Star methodology that is expensive to operate and requires manual data entry which can be time-consuming and error prone. This paper demonstrates that exploiting open datasets can deliver benchmarking models that are equally or more accurate than those developed using the CBECS dataset, the basis for Energy Star. Although the CBECS dataset provides a comprehensive understanding of buildings across the United States, given its large feature set and geographical coverage, the dataset falls short for the purposes of energy benchmarking.

Our results offer several major benefits. By using open data sources, cities can achieve similar or more accurate results than the Energy Star system, increase transparency of the process, and utilize models that are more interpretable in determining sources of energy waste. Furthermore, the open data used in this analysis allows for every building in a city to be benchmarked contingent on the collection of annual building energy use data. The open building energy benchmarking framework is not restricted to the larger cities examined but can be applied to any city or location that collects the several key variables identified in this paper. Eventually data from each city could be collected and anonymized to create a new national database. A larger compiled database of buildings would greatly expand the reach of benchmarking programs and facilitate their automation, thereby greatly reducing the required capital and time requirements of the current benchmarking process.

Benchmarking by itself will not result in energy savings. However, it can pave the path to construct better policies for improved market efficiency. Given the quantitative nature of benchmarking, we recommend policies that increase all around transparency in how these models are constructed, from data inputs to modeling, in order to increase collaboration across the science-policy boundary and to make the results more accurate and trustworthy. An open data-driven benchmarking framework can allow for faster improvements to modeling and increased trust from parties involved. Increased reporting of benchmarking scores—such as requiring buildings to place their scores in prominent locations—will provide necessary information to various stakeholders that have an interest in the operations, expenses, and energy use of buildings. The effectiveness of benchmarking can also be enhanced by complementary policies that employ requirements, incentives, and goal-based strategies. Governments are starting to realize the potential benefits of new open data sources that can be used to transition cities to a more sustainable energy future.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Jonathan Roth: Conceptualization, Methodology, Data curation, Formal analysis, Software, Writing - original draft. **Benjamin Lim:** Methodology, Data curation, Formal analysis, Software, Writing - original draft. **Rishee K. Jain:** Conceptualization, Methodology, Writing - original draft, Supervision, Project administration. **Dian Grueneich:** Conceptualization, Writing - original draft, Project administration.

Acknowledgments

This material is based upon the work supported in part by the Stanford School of Engineering under a Terman Faculty Fellowship, the Precourt Institute for Energy and the National Science Foundation under Grant Nos. 1461549, 1642315, 1941695. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.enpol.2020.111327.

References

Backlund, S., Thollander, P., Palm, J., Ottosson, M., 2012. Extending the energy efficiency gap. Energy Pol. 51, 392–396. https://doi.org/10.1016/J.ENPOL.2012.08.042.

Beddingfield, E., Hart, Z., Hughes, J.. How cities are using building energy data to drive efficiency. https://www.imt.org/wp-content/uploads/2018/03/PuttingDatatoWork_SummaryReport.pdf. (Accessed 26 June 2018).

Benchmarking and energy savings. Washington DC. www.energystar.gov/–. (Accessed 25 October 2018).

Breiman, L., 2001. Random Forest, Mach. Learn 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Buck, J., Young, D., 2007. The potential for energy efficiency gains in the Canadian commercial building sector: a stochastic frontier study. Energy 32, 1769–1780. https://doi.org/10.1016/j.energy.2006.11.008.

Capturing the multiple benefits of energy efficiency. Washington DC. https://www.iea.org/publications/freepublications/publication/Multiple_Benefits_of_Energy_Efficiency.pdf–. (Accessed 7 October 2018).

Chen, Y., Hong, T., Piette, M.A., 2017. Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis. Appl. Energy 205, 323–335. https://doi.org/10.1016/j.apenergy.2017.07.128.
Chen, Y., Hong, T., Luo, X., Hooper, B., 2019. Development of city buildings dataset for urban building energy modeling. Energy Build. 183, 252–265. https://doi.org/10.1016/j.enbuild.2018.11.008.
Chung, W., 2011. Review of building energy-use performance benchmarking methodologies. Appl. Energy 88, 1470–1479. https://doi.org/10.1016/j.apenergy.2010.11.022.
Cohen, R., Bordass, B., 2015. Mandating transparency about building energy performance in use. Build. Res. Inf. 43, 534–552. https://doi.org/10.1080/09613218.2015.1017416.
Commercial Buildings Energy Consumption Survey (CBECS), 2016. User's guide to the 2012 CBECS public use microdata file. Washington DC. www.eia.gov. (Accessed 17 October 2018).
Creutzig, F., Baiocchi, G., Bierkandt, R., Pichler, P.-P., Seto, K.C., 2015. Global typology of urban energy use and potentials for an urbanization mitigation wedge. Proc. Natl. Acad. Sci. U.S.A. 112, 6283–6288. https://doi.org/10.1073/pnas.1315545112.
Doove, L.L., Van Buuren, S., Dusseldorp, E., 2014. Recursive partitioning for missing data imputation in the presence of interaction effects. Comput. Stat. Data Anal. 72, 92–104. https://doi.org/10.1016/J.CSDA.2013.10.025.
Energy efficiency 2018, 2018. analysis and outlooks to 2040. www.iea.org/t&c/. (Accessed 25 October 2018).
Energy efficiency jobs in America 2.25 million Americans work in energy efficiency #Eejobsinamerica. https://www.e2.org/wp-content/uploads/2018/09/2018-EE-Jobs-Report.pdf, 2018–. (Accessed 1 October 2018).
Filippini, M., Hunt, L.C., 2012. US residential energy demand and energy efficiency: a stochastic demand frontier approach. Energy Econ. 34, 1484–1491. https://doi.org/10.1016/j.eneco.2012.06.013.
Gao, X., Malkawi, A., 2014. A new methodology for building energy performance benchmarking: an approach based on intelligent clustering algorithm. Energy Build. 84, 607–616. https://doi.org/10.1016/j.enbuild.2014.08.030.
Gillingham, K., Palmer, K., 2014. Bridging the energy efficiency gap: policy insights from economic theory and Empirical evidence. Rev. Environ. Econ. Pol. 8, 18–38. https://doi.org/10.1093/reep/ret021.
19 global cities Commit to make new buildings "Net-zero carbon" by 2030, C40 cities. https://www.c40.org/press_releases/global-cities-commit-to-make-new-buildings-net-zero-carbon-by-2030, 2018–. (Accessed 20 November 2018).
Gonzales-Zuñiga, S., Nascimento, L., Höhne, N., Fekete, H., Sterl, S., Hans, F., Warembourg, A., Anica, A., Van Breevoort, P., 2018. Climate Opportunity: More Jobs; Better Health; Liveable Cities Quantifying the Benefits of Climate Change Mitigation Measures in Buildings, Transport and Energy Supply/03 Climate Opportunity: More Jobs; Better Health. Liveable Cities. http://ambitiontoaction.net/. (Accessed 2 October 2018).
Hart, Z., 2015. The Benefits of Benchmarking Building Performance. https://www.imt.org/wp-content/uploads/2018/02/PCC_Benefits_of_Benchmarking.pdf. (Accessed 5 October 2018).
How libraries can make public data more accessible - CityLab. https://www.citylab.com/life/2019/02/libraries-public-information-city-data-digital-archive/581905/. (Accessed 28 October 2019).
Howard, B., Parshall, L., Thompson, J., Hammer, S., Dickinson, J., Modi, V., 2012. Spatial distribution of urban building energy consumption by end use. Energy Build. 45, 141–151. https://doi.org/10.1016/J.ENBUILD.2011.10.061.
Hsu, D., 2014a. How much information disclosure of building energy performance is necessary? Energy Pol. 64, 263–272. https://doi.org/10.1016/J.ENPOL.2013.08.094.
Hsu, D., 2014b. Improving energy benchmarking with self-reported data. Build. Res. Inf. 42, 641–656. https://doi.org/10.1080/09613218.2014.887612.
Hsu, D., 2015. Identifying key variables and interactions in statistical models of building energy consumption using regularization. Energy 83, 144–155. https://doi.org/10.1016/J.ENERGY.2015.02.008.
Katipamula, S., Wang, W., Huang, Y., Liu, G., 2012. Energy Savings Modeling of Standard Commercial Building Re-tuning Measures: Large Office Buildings Efficient Cooling, Precooling, and Demand Response by MPC View Project. https://doi.org/10.2172/1048616.
Kavousian, A., Rajagopal, R., 2014. Data-driven benchmarking of building energy efficiency utilizing statistical frontier models. J. Comput. Civ. Eng. 28, 79–88. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000327.
Kerr, N., Gouldson, A., Barrett, J., 2017. The rationale for energy efficiency policy: Assessing the recognition of the multiple benefits of energy efficiency retrofit policy. Energy Pol. 106, 212–221. https://doi.org/10.1016/J.ENPOL.2017.03.053.
Kok, N., Jennen, M., 2012. The impact of energy labels and accessibility on office rents. Energy Pol. 46, 489–497. https://doi.org/10.1016/J.ENPOL.2012.04.015.
Kontokosta, C.E., Tull, C., 2017. A data-driven predictive model of city-scale energy use in buildings. Appl. Energy 197, 303–317. https://doi.org/10.1016/J.APENERGY.2017.04.005.
Kozuch, M., 2016. 2015 energy benchmarking report - san Francisco municipal buildings. San Francisco. https://sfwater.org/modules/showdocument.aspx?documentid=10201. (Accessed 7 October 2018).
Li, G., Zhang, X., Mirzaei, P.A., Zhang, J., Zhao, Z., 2018. Urban heat island effect of a typical valley city in China: Responds to the global warming and rapid urbanization. Sustain. Cities Soc. 38, 736–745. https://doi.org/10.1016/J.SCS.2018.01.033.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. http://www.stat.berkeley.edu/. (Accessed 18 October 2018).
Ma, J., Cheng, J.C.P., 2016. Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forest. Appl. Energy 183, 193–201. https://doi.org/10.1016/J.APENERGY.2016.08.096.
Map: U.S., 2017. Building benchmarking and transparency policies. Inst. Mark. Transform. http://www.imt.org/resources/detail/map-u.s.-building-benchmarking-policies. (Accessed 10 May 2017).
Markandya, A., Wilkinson, P., 2007. Electricity generation and health. Lancet 370, 979–990. https://doi.org/10.1016/S0140-6736(07)61253-7.
Menezes, A.C., Cripps, A., Bouchlaghem, D., Buswell, R., 2012. Predicted vs. actual energy performance of non-domestic buildings: using post-occupancy evaluation data to reduce the performance gap. Appl. Energy 97, 355–364. https://doi.org/10.1016/J.APENERGY.2011.11.075.
Meng, T., Hsu, D., Han, A., 2017. Estimating energy savings from benchmarking policies in New York City. Energy 133, 415–423. https://doi.org/10.1016/j.energy.2017.05.148.
Mills, E., Mathew, P., 2009. Monitoring-based commissioning: benchmarking analysis of 24 UC/CSU/IOU projects. https://cloudfront.escholarship.org/dist/prd/content/qt06t5w9mg/qt06t5w9mg.pdf. (Accessed 23 July 2018).
Mims, N., Schiller, S., Stuart, E., 2017. Evaluation of U.S. Building Energy Benchmarking and Transparency Programs: Attributes, Impacts, and Best Practices. LBNL. https://escholarship.org/uc/item/3170j728. (Accessed 19 June 2018).
Newsham, G.R., 2009. Do LEED-certified buildings save energy? Not really. Energy Build. 41, 1386–1390. https://doi.org/10.1016/j.enbuild.2009.08.006.
Palmer, K., Walls, M., 2017. Using information to close the energy efficiency gap: a review of benchmarking and disclosure ordinances. Energy Effic 10, 673–691. https://doi.org/10.1007/s12053-016-9480-5.
Papadopoulos, S., Kontokosta, C.E., 2019. Grading buildings on energy performance using city benchmarking data. Appl. Energy 233–234, 244–253. https://doi.org/10.1016/J.APENERGY.2018.10.053.
Papadopoulos, S., Bonczak, B., Kontokosta, C.E., 2018. Pattern recognition in building energy performance over time using energy benchmarking data. Appl. Energy 221, 576–586. https://doi.org/10.1016/J.APENERGY.2018.03.079.
Pettifor, H., Wilson, C., Chryssochoidis, G., 2015. The appeal of the green deal: Empirical evidence for the influence of energy efficiency policy on renovating homeowners. Energy Pol. 79, 161–176. https://doi.org/10.1016/J.ENPOL.2015.01.015.
Pfenninger, S., DeCarolis, J., Hirth, L., Quoilin, S., Staffell, I., 2017. The importance of open data and software: is energy research lagging behind? Energy Pol. 101, 211–215. https://doi.org/10.1016/J.ENPOL.2016.11.046.
Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M.A., Pendyala, R.M., 2017. Machine learning approaches for estimating commercial building energy consumption. Appl. Energy 208, 889–904. https://doi.org/10.1016/J.APENERGY.2017.09.060.
Roth, J., Jain, R.K., 2018. Data-driven, multi-metric, and time-varying (DMT) building energy Benchmarking using smart meter data. Lect. Notes Comput. Sci. 568–593. https://doi.org/10.1007/978-3-319-91635-4_30, 10863 LNCS.
Roth, J., Rajagopal, R., 2018. Benchmarking building energy efficiency using quantile regression. Energy 152, 866–876. https://doi.org/10.1016/J.ENERGY.2018.02.108.
Sayogo, D.S., Pardo, T.A., 2013. Understanding smart data disclosure policy success: the case of green button. www.greenbuttondata.org. (Accessed 17 October 2018).
Seiden, K., Luboff, J., Chwastyk, D., Merchant, E., Russell, R., Cooper, S., Szlachetka, A., Winter Associates, S., Rode, M.. New York city benchmarking and transparency policy impact evaluation report. New York City Benchmarking snd Transparency Policy Impact Evaluation....pdf. https://www.energy.gov/sites/prod/files/2015/05/f22/DOE. (Accessed 2 October 2018).
Stone, M., Ozimek, J.F., 2010. The challenge of new marketing issues. J. Database Mark. Cust. Strategy Manag. 17, 188–200. https://doi.org/10.1057/dbm.2010.20.
N. Strachan, B. Fais, H. Daly, Reinventing the Energy Modelling–Policy Interface, (n.d.). doi:10.1038/NENERGY.2016.12.
Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. https://doi.org/10.1021/ci034160g.
The Economist, 2018. California leads subnational efforts to curb climate change. Econ. Times. https://www.economist.com/international/2018/09/15/california-leads-subnational-efforts-to-curb-climate-change. (Accessed 20 November 2018).
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Source J. R. Stat. Soc. Ser. B. 58, 267–288. https://www.jstor.org/stable/pdf/2346178.pdf?refreqid=excelsior%253Abcf4da962c8bef86ee7c6e8f8a90ac08. (Accessed 18 October 2018).
U.S.E.P. Agency, 2014. ENERGY STAR score technical reference, pp. 1–14. https://portfoliomanager.energystar.gov/pdf/reference/ENERGYSTARScore.pdf. (Accessed 7 September 2017).
U.S Energy Information Administration, 2016. Annual Energy Outlook 2016 with Projections to 2040, Washington DC doi:EIA-0383(2016).
Wong, M.R., McKelvey, W., Ito, K., Schiff, C., Jacobson, J.B., Kass, D., 2015. Impact of a letter-grade program on restaurant sanitary conditions and diner behavior in New York City. Am. J. Publ. Health 105, e81–e87. https://doi.org/10.2105/AJPH.2014.302404.
Yalcintas, M., Aytun Ozturk, U., 2007. An energy benchmarking model based on artificial neural network method utilizing US Commercial Buildings Energy Consumption Survey (CBECS) database. Int. J. Energy Res. 31, 412–421. https://doi.org/10.1002/er.1232.

Yang, Z., Roth, J., Jain, R.K., 2018. DUE-B: data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis. Energy Build. 163, 58–69. https://doi.org/10.1016/j.enbuild.2017.12.040.

Zhe Jin, G., Leslie, P., 2003. The effect of information on product quality: evidence from restaurant hygiene grade. Q. J. Econ. 118, 409–451. https://www.jstor.org/stable/pdf/25053911.pdf?refreqid=excelsior%253A99c043a7457d90109e66c4700d454cd8. (Accessed 2 October 2018).

Zhou, P., Ang, B.W., Poh, K.L., 2008. A survey of data envelopment analysis in energy and environmental studies. Eur. J. Oper. Res. 189, 1–18. https://doi.org/10.1016/j.ejor.2007.04.042.