# REVISING CITY ENERGY PERFORMANCE GRADING: THE INADEQUACY OF CURRENT STANDARDS AND THE PROMISE OF BIG DATA


## Abstract

In recent years, cities in both the U.S. and globally are adopting energy disclosure policies to better understand, and eventually improve, the condition of existing building stocks. Statistical methods have been widely adopted for building energy benchmarking, replacing or supplementing traditional engineering approaches. Although conceptually correct, we find state-of-the-art benchmarking tools to be flawed in various aspects, from data quality to model robustness and generalizability. In this work, we focus on ENERGY STAR® scoring, the prevalent energy benchmarking tool in the United States. Specifically, we show that the ENERGY STAR's models trained on nationwide samples are not able to generalize when applied to city-specific data, leading to estimates with significant uncertainty. We identify the factors resulting in this failure and propose a conceptual ontology for the next generation city energy benchmarking, built on more contextualized statistical learning algorithms and market-specific data sources.



**Authors**

Sokratis Papadopoulos and Constantine E. Kontokosta
*New York University*



**Keywords**

Urban energy benchmarking, energy efficiency labeling, data-driven energy policy

## Introduction

Building energy benchmarking refers to the process of assessing the energy performance of buildings compared to their peers, with the goal of motivating performance improvement (Pérez-Lombard et al., 2009). Both globally and in the US, there is a plethora of local governments that have passed energy benchmarking and disclosure laws aiming to understand urban energy use (Palmer and Walls, 2017). Benchmarking laws require building owners to annually report their energy use, adding transparency to energy-saving opportunities (Kontokosta, 2013). These data, some of the most robust building energy information resources available, are being used to quantify overall building energy efficiency and, increasingly, to study how buildings perform with respect to their peers.

Although benchmarking data have been used in predictive urban energy consumption models (Robinson et al,. 2018; Kontokosta & Tull, 2017), assessment of implemented energy policies (Meng et al., 2017), or energy performance pattern recognition over time (Papadopoulos et al., 2018a), their potential for developing contextualized building energy performance grading schemes has not yet been explored extensively. Building energy performance grading, and particularly its public display that has been recently adopted by New York City (The New York City Council, 2017), can act as a means to supplement existing energy policy and transform the energy efficiency market (Kontokosta, 2015). From the city leadership perspective, decision-makers can identify poor performers and design tailored and more equitable regulations or incentive mechanisms. From the building owners' point of view, energy performance grading would expose them in greater market pressure, increasing the appreciation of energy efficiency in the real estate market.

Unlike other industries, such as restaurant sanitation grading, building energy performance labelling is a complex process influenced by a set of physical, mechanical, behavioral, and meteorological factors, as well as their interactions (Kontokosta, 2015; Li et al., 2014). The current state-of-the-art in building energy performance benchmarking and grading for the U.S. is the Environmental Protection Agency's ENERGY STAR score. Despite ENERGY STAR's wide adoption by the market, its underlying model has been heavily criticized in recent literature (Kontokosta, 2015; Scofield, 2014; Gao & Malkawi, 2014; Hsu, 2014). The critique mainly arises from the model's high level of uncertainty, poor data quality and its specification errors. In this paper, we use energy benchmarking data from New York and Washington D.C. to assess the adequacy of ENERGY STAR as a nationwide standard. We find that the ENERGY STAR model is not statistically significant when applied to city-specific energy data, identify the several reasons behind this failure, and drive a discussion on the development of more fair and more contextualized building energy performance grading systems.

## Critical Assessment of Energy Benchmarking Standard

In this section, we use energy disclosure data from New York and Washington D.C. residential building stock to test ENERGY STAR's generalizability in datasets different than the ones used to train its benchmarking model. The ENERGY STAR regression model for multifamily housing stock, trained on 322 sample buildings across the US, is as follows:

$$\widehat{EUI} = 140.8 + 52.57 * cUnitDensity + 24.45 * cBedroomPerUnit - 18.76 * LowRise + 0.009617 * cHDD + 0.01617 * cCDD$$

Where $\widehat{EUI}$ is the predicted energy use intensity, *UnitDensity* is the number of units per 1,000 square feet, *BedroomPerUnit* is the number of bedrooms per unit, *LowRise* is a dummy variable being 1 if the building is lower than 5 stories tall and 0 otherwise, *HDD* are the total heating degree days and *CDD* the total cooling degree days. Prefix *c* denotes that the values are centered based on the sample's mean value. Based on the model's output, the energy efficiency ratio is defined as $^{\text{actual EUI}}/_{\text{predicted EUI}}$ Finally, based on energy efficiency ratios' distribution, a 0–100 ENERGY STAR score is calculated (ENERGY STAR, 2014).

In Figure 1, we show the ENERGY STAR model's predictions against the actual EUI values for New York and Washington D.C. We notice that the model is not able to explain any of the variance in the city-specific data for both cities (hence the negative $R^2$ values). Practically, such low $R^2$ values make the ENERGY STAR model yield energy performance scores no better than grading buildings solely based on their deviation from the sample's mean EUI.

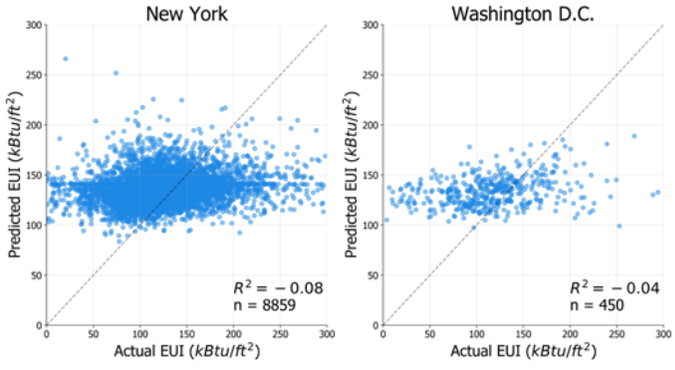


Figure 1: Model-predicted vs. actual EUI values and explained variance.

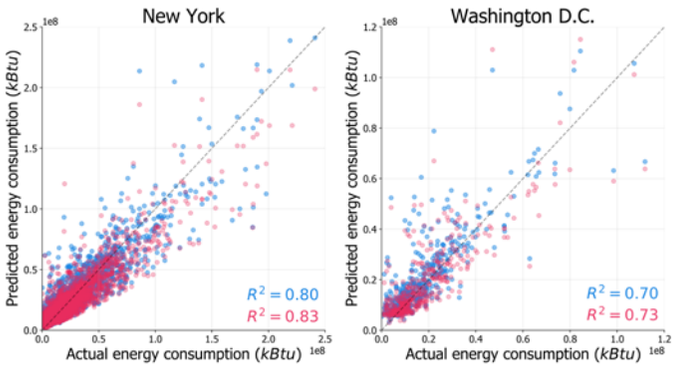


Figure 2: Total energy prediction using ENERGY STAR method (blue) and "naïve" sample mean model (red).

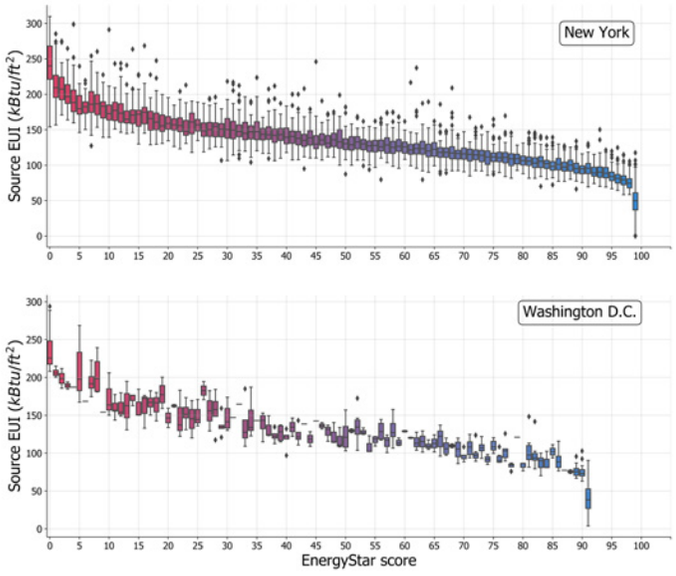


Figure 3: EUI distributions for different ENERGY STAR grades.

Nevertheless, the ENERGY STAR technical reference discusses that when the model's output is multiplied by building area to predict total building energy it explains 92.2% of the variance (ENERGY STAR, 2014). In Figure 2, we show the ENERGY STAR model's output multiplied by each individual building's area against the total annual energy consumption reported (blue-colored scatter points).

The increase in $R^2$ is evident, however this should not be attributed to the model's quality. Gross floor area is a major—if not the most important—driver of total building energy consumption; hence, the higher $R^2$ value. In fact, to validate the above mentioned hypothesis we calculate the dataset's mean EUI, multiply it by each building's gross floor area and report the goodness of fit against its actual energy consumption (red-colored scatterplot). We notice that in both New York and Washington D.C., predicting using a "naïve" model (i.e., assigning the sample's mean EUI to each building) yields higher $R^2$ than the ENERGY STAR method.`

It is imperative that these grades are robust and facilitate fair peer to peer benchmarks, especially in cases where city governments mandate the public display of buildings' energy performance grades. We argue that ENERGY STAR algorithm's inability to generalize is due to various factors such as sample size, data quality, model simplicity, and improper feature selection that we discuss extensively in the following section. Besides model-specific issues, another aspect pertaining to building energy performance grading is the interpretability of the grade itself; in a sense that differences in grade bands should be statistically significant so that there is no ambiguity that an "Excellent" building is performing better than its "Very Good" peer. In the authors' opinion, the 0–100 scale used in ENERGY STAR scoring model is not ideal; not only due to the uncertainty between different scores (based on the findings in Figure 1) but also due to its granularity that makes it harder for interested stakeholders to interpret. In Figure 3 we show the EUI boxplots for the different ENERGY STAR score grades. Although the trend is consistent, with higher scores associated with lower EUI values, we notice that differences in median EUI values are not easily distinguished, especially for mid-tier performing buildings (scores ranging from 30–60).

### The Need For a Paradigm Shift

Driven by the findings in Section 2, here we discuss the main drawbacks and faults of the predominant approach in building energy benchmarking (i.e., ENERGY STAR), and drive the discussion for a needed paradigm shift toward more fair, robust, and contextualized grading schemes. Current methods' limitations lie on two axes: (a) the properties of the benchmarking algorithms, and (b) the nature of the data used to train the benchmarking models. Specifically, we identify five key areas where machine learning and city-specific building energy data can be used to develop novel energy performance indicators (Figure 4).

Starting from the statistical learning algorithms used to benchmark energy performance, ENERGY STAR is built on a linear regression model. Nonetheless, the relationship between energy consumption and building characteristics is oftentimes non-linear (Kontokosta, 2015), making linear models unable to capture it. We argue that more complex machine learning models are more appropriate for building energy performance benchmarking. Artificial neural networks (Melo et al., 2014) and tree-based ensemble learning methods (Papadopoulos et al., 2018b), for instance, have shown promising results, outperforming linear methods. Tree-based ensemble learning algorithms (e.g., Random Forests or Gradient Boosted Regression Trees) in particular, unlike artificial neural networks, account for the issue of model interpretability to a great extent, adding transparency in the benchmarking process.

Data, both in terms of quantity and quality, is another issue related to the ENERGY STAR method. The model utilizes data from a nationwide survey, assuming that the interactions between building characteristics and energy consumption are homogeneous within the entire national building stock. However, this is not a valid assumption contradicting recent research findings; in a 2017 study (Papadopoulos et al., 2017) the authors showed that the relationship between building characteristics, such as age and size, and energy use intensity varies significantly from city to city. Furthermore, the data sample used to train the ENERGY STAR model is relatively small (i.e., 322 residential properties). In Section 2, where we evaluated the ENERGY STAR's generalizability, we applied its model on 8,859 and

450 buildings, in New York and Washington D.C., respectively. In the age of big data and energy disclosure that the amount of information generated and gathered is unprecedented, it is only logical that benchmarking models, whether national or regional, should be trained on richer data samples. Regarding data quality, energy disclosure policies mandate the reporting of a detailed building feature list, from physical properties to occupancy and fuel quality mix. All these pieces of information can (and should) be incorporated in benchmarking models to better explain variations in energy performance. Currently, ENERGY STAR model uses five features, overlooking a multitude of building physical properties as well as the relative use of different fuel types.

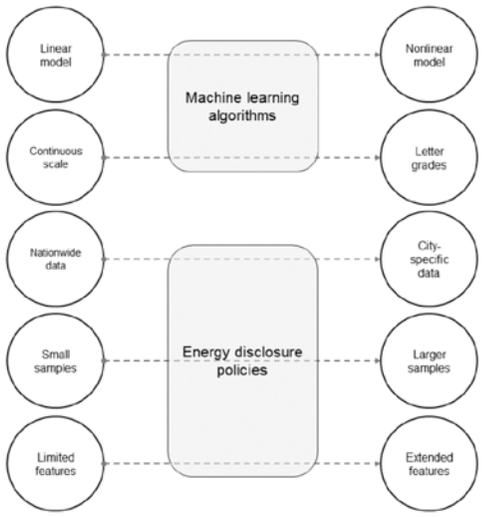


Figure 4: Areas of improvement in energy benchmarking methods.

The last, but not least, significant aspect of energy benchmarking is the communication of the results to a wide and diverse range of stakeholders, such as building owners, potential tenants, and city executives (Kontokosta, 2015). The 1–100 scale used in ENERGY STAR presents a granular numerical rating that might be subject to the model's uncertainty to a great extend (Hsu, 2014) and in various cases distinction between grade bands is unclear (see Figure 3). To address this issue, we argue that a simpler, more intuitive letter-grade scale would be more appropriate to showcase differences in building energy performance levels. Unsupervised learning algorithms can be used to cluster the energy performance ratios and assign each building to its corresponding energy performance band.

As the need for climate change mitigation increases, decision-makers are turning towards more aggressive policy frameworks to reduce buildings' carbon footprint and transform the energy efficiency market. Awareness-raising campaigns are evolving to regulations, and energy disclosure mandates move towards public-facing building energy performance exposure. Transparency, robustness, and fairness should be the core principles of such policies in order to be seamlessly integrated into existing frameworks (Borgstein et al., 2016). We show that the data abundancy from energy disclosure laws, in combination with the appropriate analytical methods can be used to develop more accurate and contextualized building energy grading systems, accounting for the factors driving energy efficiency in individual cities.

## Conclusion

Cities across the US and worldwide are mandating energy disclosure laws to better understand how energy is consumed in the built environment. The abundant data streams from such laws constitute an unprecedented opportunity to: (a) assess the robustness of the state-of-the-art building energy benchmarking and energy performance grading methods, and (b) propose novel approaches to address the limitations of existing methods.

In this work, we used energy disclosure data from New York and Washington D.C. to show that ENERGY STAR, the predominant energy benchmarking model in the US, is not able to generalize when applied to city-specific data making its estimates uncertain and unreliable. Following, we identified the key elements that limit ENERGY STAR's generalizability and proposed a roadmap towards a paradigm shift in building energy performance grading. Our proposition is that a combination of city-specific building energy data and machine learning algorithms would add robustness, transparency, and fairness in the way the market perceives building energy performance.

## References


Borgstein, E. H., Lamberts, R., & Hensen, J. L. M. (2016). Evaluating energy performance in non-domestic buildings: A review. *Energy and Buildings*, *128*, 734–755.

ENERGY STAR (2014, September). Technical Reference ENERGY STAR Score for Multifamily Housing in the United States. Retrieved from https://www.energystar.gov/sites/default/files/tools/Multifamily.pdf

Gao, X., & Malkawi, A. (2014). A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm. *Energy and Buildings*, *84*, 607–616.

Hsu, D. (2014). Improving energy benchmarking with self-reported data. *Building Research & Information*, *42*(5), 641–656.

Kontokosta, C. E. (2013). Energy disclosure, market behavior, and the building data ecosystem. *Annals of the New York Academy of Sciences*, *1295*(1), 34–43.

Kontokosta, C. E. (2015). A market-specific methodology for a commercial building energy performance index. *The Journal of Real Estate Finance and Economics*, *51*(2), 288–316.

Kontokosta, C. E., & Tull, C. (2017). A data-driven predictive model of city-scale energy use in buildings. *Applied Energy*, *197*, 303–317.

Li, C., Hong, T., & Yan, D. (2014). An insight into actual energy use and its drivers in high-performance buildings. *Applied Energy*, *131*, 394–410.

Melo, A. P., Cóstola, D., Lamberts, R., & Hensen, J. L. M. (2014). Development of surrogate models using artificial neural network for building shell energy labelling. *Energy Policy*, *69*, 457–466.

Meng, T., Hsu, D., & Han, A. (2017). Estimating energy savings from benchmarking policies in New York City. *Energy*, *133*, 415–423.

Papadopoulos, S., Bonczak, B., & Kontokosta, C. E. (2017). Spatial and Geographic Patterns of Building Energy Performance: A Cross-City Comparative Analysis of Large-Scale Data. In *International Conference on Sustainable Infrastructure 2017* (pp. 336–348).

Papadopoulos, S., Azar, E., Woon, W. L., & Kontokosta, C. E. (2018). Evaluation of tree-based ensemble learning algorithms for building energy performance estimation. *Journal of Building Performance Simulation*, *11*(3), 322–332.

Papadopoulos, S., Bonczak, B., & Kontokosta, C. E. (2018). Pattern recognition in building energy performance over time using energy benchmarking data. *Applied Energy*, *221*, 576–586.

Pérez-Lombard, L., Ortiz, J., González, R., & Maestre, I. R. (2009). A review of benchmarking, rating and labelling concepts within the framework of building energy certification schemes. *Energy and Buildings*, *41*(3), 272–278.

Palmer, K., & Walls, M. (2017). Using information to close the energy efficiency gap: a review of benchmarking and disclosure ordinances. *Energy Efficiency*, *10*(3), 673–691.

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. *Applied Energy, 208,* 889–904.

Scofield, J. H. (2014). ENERGY STAR building benchmarking scores: good idea, bad science. *Oberlin College study for American Council for an Energy Efficient Economy (ACEEE)*.

The New York City Council (2017, August). A Local Law to amend the administrative code of the city of New York, in relation to energy efficiency scores and grades for certain buildings. Retrieved from http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0