Spatial and Temporal Modeling of Urban Building Energy Consumption Using Machine Learning and Open Data

Jonathan Roth[1]; Aimee Bailey[2]; Sonika Choudhary[3]; and Rishee K. Jain[4]

[1]Urban Informatics Lab, Dept. of Civil and Environmental Eng., Stanford Univ., 473 Via Ortega, Stanford, CA 94305, USA. E-mail: jmroth@stanford.edu
[2]EDF Innovation Lab, 4300 El Camino, Los Altos, CA 94109, USA. E-mail: Sonika.Choudhary@edf-inc.com
[3]EDF Innovation Lab, 4300 El Camino, Los Altos, CA 94109, USA. E-mail: aimee.bailey@edf-inc.com
[4]Urban Informatics Lab, Dept. of Civil and Environmental Eng., Stanford Univ., 473 Via Ortega, Stanford, CA 94305, USA. E-mail: rishee.jain@stanford.edu

ABSTRACT

Understanding the spatial and temporal distribution of energy consumption in cities is critical to facilitate the identification of potential energy saving opportunities and planning of new renewable and integrated district energy systems. Previous work analyzing urban building energy usage has been largely limited to either modeling of individual buildings at granular temporal scales (i.e., hourly or less) or an entire stock of urban buildings at the yearly temporal scale. While such analyses are valuable, their lack of both spatial and temporal granular modeling limits their applicability in planning and design of integrated district energy systems. This paper proposes a new urban building energy model that produces hourly demand profiles for the building stock of New York City (NYC) using only open publicly available data. First, we utilize a machine learning model to predict annual energy consumption of NYC's entire building stock from a subset of buildings that have publicly available annual energy usage data. We validate this part of the model using city-wide electricity data from New York Independent System Operator (NYISO). Results show that random forests have the best building-level prediction accuracy with a mean log squared error of 0.293. Next, we apply a novel optimization algorithm to construct temporal granular hourly profiles using the Department of Energy's commercial and residential simulation building reference sets, and the predicted annual energy values from the random forests model. Results indicate that we are able to achieve an error rate of ~10% (MAPE) in comparison to the overall hourly electricity profile of NYC. Moreover, we found that our iterative approach demonstrates that error rates diminish as buildings are added to the aggregated profile, which underscores the merits of applying our proposed method to model the entire building stock of a city rather than an individual building. In the end, our proposed method takes the first step of large-scale spatial and highly granular temporal characterization of urban building energy usage.

INTRODUCTION

As buildings in cities consume between 30-70% of total primary energy use, many municipalities are focused on better understanding their usage patterns to find ways to reduce or shift their demand (Chen et al. 2017). Increased adoption of renewables are creating a stronger need to not only understanding the spatial patterns of building energy consumption, but also their temporal patterns. With this information, policy makers, engineers, and utilities can identify buildings that may benefit from energy efficiency retrofits, new storage technologies, or access to district energy systems. New publicly available datasets are providing more information about urban morphologies, enabling a better understanding of energy usage and building characteristics. This paper proposes a method for modeling the hourly energy demand of individual buildings across New York City (NYC) using public data and machine learning models. We first construct a machine learning model to predict the annual energy use of each building in NYC using a small sample of buildings with publicly available energy use data. We then map each building in the city to three building archetypes in the U.S. Department of Energy's (DoE) reference building dataset. Finally, we construct hourly loads for each building by fitting a weighted average of these three building types' load curves, adjusting for weather and weekday effects. Lastly, we validate the proposed model using city-wide electricity hourly demand data from the New York Independent System Operator (NYISO) to obtain a sense of error associated with the constructed load profiles.

RELATED WORK

Previous work has analyzed energy usage dynamics in cities, with several focusing on NYC as a case-study. Howard et al. proposed a methodology to model building-level annual energy use intensities by downscaling zip code level energy data using a linear model (Howard et al. 2014). While valuable in understanding general spatial trends of energy usage, a limitation of this work was the lack of validation of individual building loads. Another study (Robinson et al. 2017) built a machine learning model using a subset of building-specific energy usage data but results were validated only at the individual building level. Finally, Kontokosta and Tull proposed a machine learning model that also utilized a subset of publicly available building data and validated it at both the building and zip-code level (Kontokosta and Tull 2017). Results of this work demonstrated that linear regression performed best at the zip-code level while a support vector machine model performed best at the building-level. The model proposed in this paper extends these previous works by constructing an integrated machine learning and optimization method that predicts annual energy loads and then translates these loads into hourly profiles for individual buildings.

PROPOSED MODEL AND RESULTS

The overarching objective of this study is to construct an urban building energy model that calculates the hourly load profiles for every building in New York City (NYC). Figure 1 breaks the analysis down into two primary steps: 1) Constructing annual building-level energy estimates for all buildings in NYC using machine learning techniques using real annual energy consumption data from about 15,000 buildings; 2) Converting annual energy loads into hourly demand profiles using archetypal simulation model outputs and a novel optimization algorithm.

Data Collection and Cleansing

For this study, we focused on NYC due to its size, number of buildings, and availability of publicly available energy data as part of its local building benchmarking initiative (i.e., Local Law 84). For step 1 of our method, we utilized the 2016 *energy* data disclosed as part of Local Law 84 which contains about 15,000 buildings. We opted to use total building *energy* data since the disclosed *electricity* data contained more missing data and erroneous values. Building characteristic data was obtained from the publicly available Primary Land Use Tax Lot Output (PLUTO) dataset. The PLUTO dataset contains building features for every tax lot in NYC and contains over 1 million buildings. Since the data from these two sources were quite messy, an extensive data cleansing process was performed. First, total site energy use was computed for each building in the LL84 dataset by multiplying the site energy use intensity (EUI) by the building area. Outliers were identified by finding all points that were outside four times the interquartile range for site EUI and then removed. Finally, any site EUI below one was also removed. For the PLUTO dataset, a number of features were constructed. The log transformation was applied to ten separate features and appended to the dataset. In addition, fractions of floor space by use type were also calculated and added to the final dataset. A total of 38 features were included in the final dataset. Before modeling, all missing values were imputed using predictive mean matching (note, that less than 1% of data was missing from features that are used in our model) (Landerman et al. 1997). The MICE package in R was used to generate multiple imputations for the incomplete data through Gibbs sampling. The classification and regression tree methodology was used due to its flexibility in handling missing data and ability to find non-linear relationships.

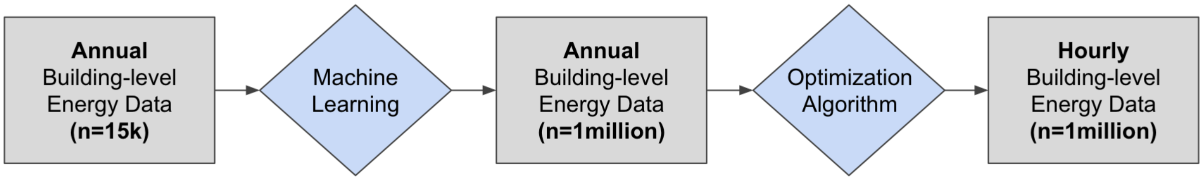
Annual Building-level Energy Data **(n=15k)**
Machine Learning
Annual Building-level Energy Data **(n=1million)**
Optimization Algorithm
Hourly Building-level Energy Data **(n=1million)**

Figure 1 - Overview of proposed urban building energy model.

For step 2 of our method, archetypal hourly building loads were obtained from reference building simulations models developed by the U.S. Department of Energy (DoE) (Office of Energy Efficiency & Renewable Energy (EERE) n.d.). We collected data for all 16 commercial building types and 2 residential building types (high and medium energy usage) under the TMY3 (typical meteorological year, version 3) conditions for Central Park. Additionally, we added one more hand constructed profile to this reference set which we named “Datacenter” that modeled energy usage as a flat energy curve. Each simulated building energy profile included a breakdown of loads by total facility electricity use, electricity for heating, electricity for cooling, electricity for interior lighting, electricity for interior equipment, total facility gas use, gas for heating, gas for interior equipment, and gas for water heating. The reference building simulations are based on typical meteorological years and do not normalize energy usage profiles to specific weather conditions in a given year. In order to overcome this limitation, we collected 2016 hourly weather data for NYC using the OpenWeatherMap website. Finally, hourly electricity demand for NYC was collected from the New York Independent System Operator (NYISO); this data is used to calibrate the urban building energy model. Several data points were missing for the NYISO hourly electricity load and were linearly imputed based upon the nearest two hours of load. Before modeling, we accounted for the 2016 leap year by appending 24 hours of data to the DOE reference buildings dataset for February 29th, which we assumed took on the same values as the previous day’s load.

Step 1: Predicting Annual Building Loads Using Machine Learning

Following a similar approach presented by Robinson, et al., we deployed four commonly utilized machine learning models (i.e., lasso regression, random forest, gradient boosting, and support vector machines) to predict the annual energy use of building in NYC (Robinson et al. 2017). For each model, we use 5-fold cross validation to prevent overfitting and ensure our models are generalizable. The mean squared error (MSE) is used to assess the fit of our model where $MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_i\right)^2$. Here, y is the log transformation of annual building energy use (in kBtu), $\hat{y}$ is the predicted output from the model, N is the total number of buildings and i refers to a specific building. We used the log transformation as this is common in the literature when modeling annual building consumption due to the wide energy consumption range and the heteroskedastic nature of building data (Kontokosta and Tull 2017; Yang et al. 2018). Each model examined has a different set of hyperparameters that can be tuned to increased performance. All tuning was performed in R using the 'caret', 'glmnet', 'svm', 'randomForest', and 'xgboost' packages. Lasso regression has one hyperparameter (lambda) which is a regularization term based on the L1 norm (Tibshirani 1996). To pick our final lasso model, we performed a linear search and selected the value with the lowest resulting cross-validation error. For random forests, gradient boosting, and support vector machines, we did a grid search over 12 different combinations of hyperparameters. The summary of the four examined models and the MSE from the best set of hyperparameters is shown in Table 1. All of the models had fairly similar performance, but the random forest model proved to be the best, therefore the results from this model was utilized in the second step of our method.

Table 1 -Summary of the hyperparameters examined for each of the models as tested using 5-fold cross-validation. The shown error rates are for the models with the lowest cross-validation MSE after performing the grid search and selecting the optimal final hyperparameters.

Models	Hyperparameters	Final Hyperparameters	Final MSE
Lasso Regression	Lambda penalization: $\lambda = 0:1$	Lambda penalization: $\lambda = 0.0104$	0.312
Random Forest	Max Features: *3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25*	Max Features: *5*	0.293
Gradient Boosting	# Boosting Iterations: *1000, 2000, 3000, 4000* Learning rate: $\eta = 0.01, 0.001, 0.0001$	# Boosting Iterations: *1000* Learning rate: $\eta = 0.0001$	0.343
Support Vector Machines	Kernel: *Linear* Penalty Factor: $C = 1, 3, 100$ Insensitivity parameter: $\varepsilon = 0.1, 0.4, 0.7, 1.0$	Kernel: *Linear* Penalty Factor: $C = 1$ Insensitivity parameter: $\varepsilon = 0.4$	0.316

Step 2: Constructing Building Hourly Energy Demand Profiles

In order to construct hourly demand profiles from each building's annual predicted energy use (output of step 1), we leveraged the simulation results from the 19 DoE reference buildings. First, we assign each building in NYC to 3 different reference profiles. The PLUTO dataset contains its own definitions of building type that are broken down into 25 categories, where we map each category to 3 different reference buildings. This one-to-three mapping was selected to reduce bias introduced by the authors. Let $D_{i,1}$, $D_{i,2}$ and $D_{i,3}$ denote the three DoE reference buildings mapped to PLUTO building category i, where $i = \{1,2,\ldots,25\}$. Because energy consumption varies among buildings even of the same type, different buildings will have different demand profiles. Let $Y_{i,j}$ denote the hourly energy demand for building j of PLUTO building category i, where $Y_{i,j} \in \mathbb{R}^{8784}$ (there are $H = 8784$ hours in the 2016 leap year). The primary assumption in this step is that each building will follow a load profile that is a linear combination of the 3 assigned DoE profiles. In addition, these profiles are adjusted for 2016 weather, with an hourly temperature vector T and a cooling-degree-hour vector C where $C^t = \max(0, T^t - 65)$ and t is the hours in the year. Finally, a vector W represents a final adjustment for business days, which are given a value of 1, while all weekends and holidays are given a value of 0. In total, 6 vectors define the energy consumption of each building.

Using the NYC electricity demand profile, denoted as $\in \mathbb{R}^{8784}$, we set up an optimization problem which minimizes the difference between this profile and the *aggregated* building profiles, given as A. The optimization selects parameter weightings, defined as $\beta_{i,j}^{k}$, for each building where $k = \{1,2,\ldots,6\}$ represents the 6 vectors that define a building's consumption profile. Given this problem formulation, and that $Y_{i,j}$ is a linear combination of the weighted vectors, each building profile is scaled such that $\sum_t Y_{i,j}^t = 1$. To achieve this, the parameters $\beta_{i,j}^{k}$ are found in relation to the scaled DoE, weather, and business day vectors. The 3 DoE vectors ($D_{i,1}$, $D_{i,2}$ and $D_{i,3}$) represent the building load, and therefore we set $\sum_t D_{i,1}^t = \sum_t D_{i,2}^t = \sum_t D_{i,3}^t = 1$; the final load is then set as a weighted sum of the three vectors such that $\sum_{k=1}^{3} \beta_{i,j}^{k} = 1$. The weather and business day vectors are considered adjustments to the final load and are therefore scaled such that $\sum_t T^t = \sum_t C^t = \sum_t W^t = 0$. We set up the optimization function to minimize the mean absolute percentage error (MAPE) between the NYC electricity profile and the aggregated building profiles as follows:

$$\min_{\beta_{i,j}^{k}} \text{imize} \frac{1}{H} \sum_{t=1}^{H} \frac{\left| A^t - N^t \right|}{N^t}$$

$$\textit{subject to } Y_{i,j} = \beta_{i,j}^{1} D_{i,1} + \beta_{i,j}^{2} D_{i,2} + \beta_{i,j}^{3} D_{i,3} + \beta_{i,j}^{4} T + \beta_{i,j}^{5} C + \beta_{i,j}^{6} W$$

$$A = \sum_{i,j} \left(Y_{i,j} * E_{i,j} \right)$$

$$\sum_{k=1}^{3} \beta_{i,j}^{k} = 1$$

$$\hat{\beta}_i^k = \frac{1}{n} \sum_{j}^{n} \beta_{i,j}^{k}$$

$$\left(\beta_{i,j}^{k} - \hat{\beta}_i^k \right) < \epsilon$$

The aggregated building load A is defined as the sum of the individual building profiles multiplied by their annual *energy* use, defined as $E_{i,j}$, which is given from the output of step 1.

Step 2 ensures that each building's total annual energy use is equal to the value predicted from the previous step. Although we want to include variation in building profiles *within* the same class i, these profiles should also be fairly similar since they serve similar functions. Therefore, a constraint is placed on each $\beta_{i,j}^{k}$ such that this value cannot deviate away from the mean of $\hat{\beta}_{i}^{k}$ (i.e., the mean over every building j in class i for parameter k) by more than $\epsilon = 0.1$.

Due to the high number of free parameters in the optimization function, validating the constructed model can be difficult. Our solution was to construct an iterative approach that allows us to understand the relationship between the error rate and the number of aggregated building profiles as buildings are sequentially added to the model. To fit the model, a random sample of 500 buildings is taken from the NYC dataset, where for each building, a random search is conducted over one-hundred $\beta_{i,j}$ vectors (given the constraints outlined above) until there is a decrease in the objective function (Solis and Wets 1981); otherwise a random vector is selected. To account for variations in building schedules, a random shift of $s = \{-2,-1,0,1,2\}$ hours to each building profile was also implemented. The fitting procedure was repeated 60 times to examine the stability of the fit parameters between independent samples. This assumes that a random subset of aggregated building profiles approximates that same profile of New York City's total electricity usage. The value of the objective function for each of the independent trials – as buildings are added to the aggregated profile – can be seen in Figure 2.

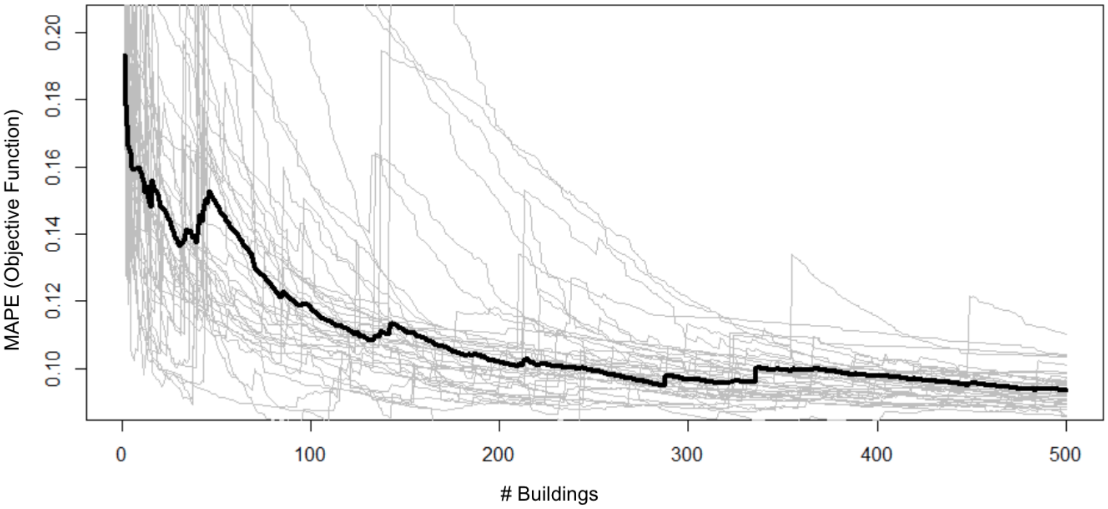

Figure 2 - The value of the objective function over 60 independent trials as buildings are iteratively added to the aggregated load profile. The dark black line is the average trend.

Figure 2 shows that the error rate between the NYC profile and the aggregated profile, which starts to plateau around 250 buildings; however we also observe that each trial contains a lot of noise, especially when few buildings have been added to the aggregated profile. The large amount of noise largely comes from the randomness of the buildings that have been added to the profile. For example, the first 50 randomly selected buildings may have small energy demands but the 51st building might be a skyscraper that dwarfs the cumulative demand of the previously added buildings, making the fitted profile of this building especially important. If no parameters are found to reduce the error between the aggregated profile and the NYC profile, this can lead to a large increase in the objective function value.

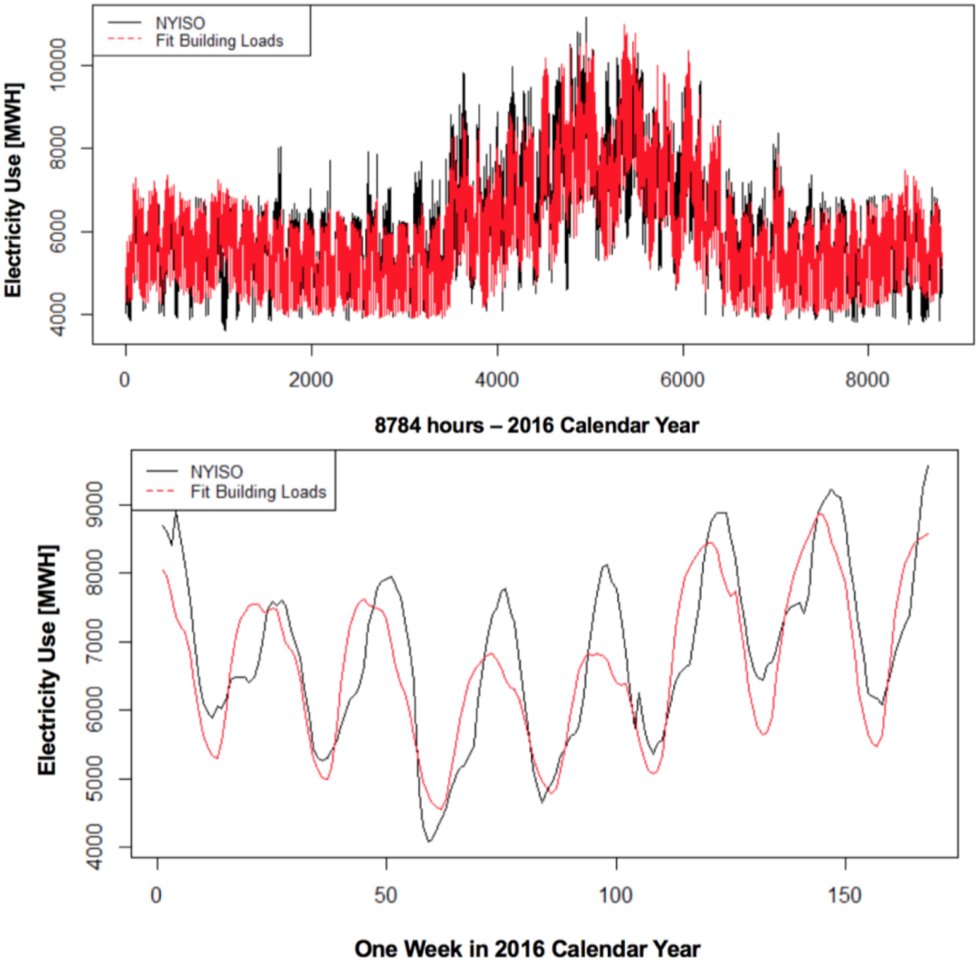

Figure 3 – (a) comparison between the NYC profile and the fit building loads for 2016; (b) zoomed in view of comparison between the NYC profile and fit building loads for a typical week in 2016

The fitted model comparing the aggregated building loads to the NYC electricity profile (from NYISO) is seen in Figure 3a and 3b. It can be seen that the model captures the main trend for NYC but struggles to capture certain anomalies, like the high spike that occurs near hour 1,600. It is important to note that the building profiles were scaled using their total energy consumption data while the NYISO data is measuring the electricity consumption of the city; however, such scaling allows us another opportunity to validate the model. First, the urban energy model is completed by constructing hourly loads for every building in the city by randomly sampling the $\beta_{i,j}$ vectors from the 60 independent trials for each buildings' PLUTO class i. For example, if building class $i=3$ had a total of 250 examined buildings across the 60 trials, then the $\beta_{i,j}$ vectors for each building in the city of that class is determined by taking a random sample from the 250 fitted $\beta_{i,j}$ vectors. Every building in NYC now has a weighting of 3 DoE reference buildings. Each DoE building has a different fraction of total energy demand that comes from electricity, which ranges from 34-94%. This large range is due to the fact that residential buildings use a lot of natural gas for heating, resulting in a lower electricity use fraction, while office buildings have a high electricity use fraction due to higher lighting loads.

Since the output of step 1 is the total *energy* demand for each building in NYC, we determine the total *electricity* demand by taking the weighted sum of electricity demand from each of the three assigned DoE buildings. After summing the newly calculated electricity demand profiles for each building in the city, we find that the demand totals 56.8 million MWH compared to the NYISO ground truth load of 53.6 million MWH. This translates to an error of 5.9% for city-wide electricity demand.

LIMITATIONS AND FUTURE WORK

The urban building energy model presented in this work makes several key assumptions. First, it assumes that a small subset of building profiles in aggregate approximates the city-wide profile for an entire city (i.e., New York City), and that building profiles are also scaled using their total energy demand rather than electricity demand (due to data constraints). Second, it assumes that each building profile is a linear combination of DoE reference building models. Though each building profile is adjusted for weather and business day operations, the model is not validated at the building-hourly level since this data is difficult to obtain. Despite such limitations, this represents a first attempt at producing building-level hourly loads for electricity and total energy demand for an entire city using only readily available public datasets. Future work aims to streamline the optimization algorithm to make it less computationally intensive. Moreover, we aim to construct new forms of validation by fitting the model to a 10 month period of the 2016 NYISO data and testing the error rate on the unseen data from the remaining 2 months. Finally, we hope to obtain building-level hourly profiles for a handful of buildings and use this to validate our proposed methodology.

CONCLUSION

This paper proposes an urban building energy model that determines hourly electricity and total energy demand profiles for every building in New York City by integrating physics-based simulation models and machine learning techniques. In doing so, our proposed model represents a first-step towards a validated and highly granular spatio-temporal characterization of urban building energy usage. A key feature of our proposed model is its reliance on only public and readily available data thus making it highly extensible to the numerous city's around the world that have building benchmarking initiatives. With the world continuing to urbanize, understanding the spatio-temporal dynamics of urban building energy use is integral to facilitating our transition towards more efficient, sustainable, and integrated energy systems.

REFERENCES

Chen, Y., Hong, T., and Piette, M. A. (2017). "Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis." *Applied Energy*, Elsevier, 205(July), 323–335.

Howard, B., Saba, A., Gerrard, M., and Modi, V. (2014). "Combined heat and power's potential to meet New York City's sustainability goals." *Energy Policy*, Elsevier, 65, 444–454.

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

Landerman, L. R., Land, K. C., and Pieper, C. F. (1997). "An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values." *Sociological Methods & Research*, SAGE PERIODICALS PRESS, 26(1), 3–33.

Office of Energy Efficiency & Renewable Energy (EERE). (n.d.). “Commercial and Residential Hourly Load Profiles for all TMY3 Locations in the United States.” *Department of Energy*, <https://openei.org/doe-opendata/dataset/commercial-and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states> (Nov. 18, 2018).

Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A., and Pendyala, R. M. (2017). “Machine learning approaches for estimating commercial building energy consumption.” *Applied Energy*, Elsevier, 208, 889–904.

Solis, F. J., and Wets, R. J.-B. (1981). “Minimization by Random Search Techniques.” *Mathematics of Operations Research*, INFORMS , 6(1), 19–30.

Tibshirani, R. (1996). “Regression Shrinkage and Selection via the Lasso.” *Source: Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.

Yang, Z., Roth, J., and Jain, R. K. (2018). “DUE-B: Data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis.” *Energy and Buildings*, Elsevier, 163, 58–69.