Applied Energy 197 (2017) 303–317

Contents lists available at ScienceDirect

Applied Energy

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

A data-driven predictive model of city-scale energy use in buildings

Constantine E. Kontokosta [a,*], Christopher Tull [b]

[a] Center for Urban Science and Progress & Tandon School of Engineering, New York University, 1 Metrotech Center, 19th Floor, Brooklyn, NY 11201, United States
[b] Center for Urban Science and Progress, New York University, 1 Metrotech Center, 19th Floor, Brooklyn, NY 11201, United States

HIGHLIGHTS

- This paper provides insight into urban energy dynamics.
- Machine learning is used to predict building energy use at the city scale.
- Actual energy use data for more than 20,000 buildings is used.
- Energy use intensity is predicted for all 1.1 million buildings in New York City.
- Predictions are validated using actual building and zip code level energy data.

ARTICLE INFO

Article history:
Received 21 November 2016
Received in revised form 25 March 2017
Accepted 1 April 2017

Keywords:
Machine learning
Building energy
Energy efficiency
Urban dynamics
Energy prediction

ABSTRACT

Many cities across the United States have turned to building energy disclosure (or benchmarking) laws to encourage transparency in energy efficiency markets and to support sustainability and carbon reduction plans. In addition to direct peer-to-peer comparisons, the benchmarking data published under these laws have been used as a tool by researchers and policy-makers to study the distribution and determinants of energy use in large buildings. However, these policies only cover a small subset of the building stock in a given city, and thus capture only a fraction of energy use at the urban scale. To overcome this limitation, we develop a predictive model of energy use at the building, district, and city scales using training data from energy disclosure policies and predictors from widely-available property and zoning information. We use statistical models to predict the energy use of 1.1 million buildings in New York City using the physical, spatial, and energy use attributes of a subset derived from 23,000 buildings required to report energy use data each year. Linear regression (OLS), random forest, and support vector regression (SVM) algorithms are fit to the city's energy benchmarking data and then used to predict electricity and natural gas use for every property in the city. Model accuracy is assessed and validated at the building level and zip code level using actual consumption data from calendar year 2014. We find the OLS model performs best when generalizing to the City as a whole, and SVM results in the lowest mean absolute error for predicting energy use within the LL84 sample. Our median predicted electric energy use intensity for office buildings is 71.2 kbtu/sf and for residential buildings is 31.2 kbtu/sf with mean absolute log accuracy ratio of 0.17. Building age is found to be a significant predictor of energy use, with newer buildings (particularly those built since 1991) found to have higher consumption levels than those constructed before 1930. We also find higher electric consumption in office and retail buildings, although the sign is reversed for natural gas. In general, larger buildings use less energy per square foot, while taller buildings with more stories, controlling for floor area, use more energy per square foot. Attached buildings – those with adjacent buildings and a shared party wall – are found to have lower natural gas use intensity. The results demonstrate that electricity consumption can be reliably predicted using actual data from a relatively small subset of buildings, while natural gas use presents a more complicated problem given the bimodal distribution of consumption and infrastructure availability.

1. Introduction

Cities are increasingly adopting long-term sustainability plans designed to increase the efficiency of energy infrastructure, reduce operating costs, and mitigate the negative effects of climate change

* Corresponding author.
E-mail address: ckontokosta@nyu.edu (C.E. Kontokosta).

http://dx.doi.org/10.1016/j.apenergy.2017.04.005
0306-2619/© 2017 Elsevier Ltd. All rights reserved.

[1–4]. As buildings account for a majority of primary energy use and carbon emissions in dense urban areas, these plans tend to focus on the "greening" of existing buildings as a path to greater resource efficiency. One of the most notable urban policy innovations for buildings has been energy disclosure. These laws require annual energy use reporting for a subset of buildings in a city's inventory, typically those larger than a particular size threshold. In New York City (NYC), Local Law 84 covers approximately 20,000 buildings larger than 50,000 square feet each year. While this is a significant sample that accounts for approximately 45 percent of the City's total energy consumption, it represents only 2 percent of the NYC building stock ([5–8]). Beginning in 2018, this mandate will expand to cover all buildings over 25,000 square feet, similar to disclosure policies in other U.S. cities. There are legitimate political, financial, and privacy concerns that constrain the expansion of these laws to smaller buildings, particularly driven by the potential cost to building owners. Given this reality, it is imperative for policy-makers tasked with reducing city-wide energy use and carbon emissions to have alternative, but reliable, methods to understand energy consumption across multiple spatial scales.

This study develops a predictive model of energy use at the building, district, and city scales using training data from energy disclosure policies and predictors from widely-available property and zoning information. For city planners and energy policymakers, understanding energy use dynamics is critical to (1) knowing where and how energy is being consumed across the morphologic and socioeconomic contours of the city, (2) providing situational awareness of energy use to better allocate resources and target policy interventions to reduce consumption, and (3) identifying cost-efficient savings opportunities across the city. In addition to developing a time-invariant snapshot of consumption patterns, reliable energy use predictions can inform forecasting and policy scenario modeling over time. From an energy management perspective, a building-specific, city-wide energy profile provides opportunities for efficiencies through tracking, benchmarking, and impact evaluation of new programs. For planners, the ability to understand future energy demand – given expectations for land use changes, urban development, and other technological, architectural, and behavioral factors that could alter future energy use patterns – provides an important yardstick by which to evaluate policy alternatives, set goals, and measure progress.

In this paper, we evaluate several prediction algorithms, including ordinary least squares regression, support vector regression, and random forest, and two feature selection approaches to predict building-specific annual energy use and energy use intensity (EUI) from existing property and land use administrative records. We combine actual energy use and building attribute data from NYC's Local Law 84 (LL84) for 20,652 buildings with the NYC Department of City Planning's Primary Land Use Tax Output (PLUTO) database, covering all 1,082,437 properties in NYC, to predict in-sample and out-of-sample building energy use, and validate our prediction model output against actual zip code-level energy use data provided by New York City's energy utilities, ConEdison and National Grid. We begin by presenting the relevant literature and describing our data and data cleaning process. We then discuss our methodology and model results, and conclude with an exploration of the implications and significance of our findings for urban energy analytics and energy policy.

2. Literature review

Despite the importance of cities and buildings to the reduction of global energy use and carbon emissions, relatively little attention has been paid to the data-driven study of urban energy dynamics. The traditional focus of demand-side energy studies has been on building simulation and systems-level physical models of building technologies and components, rather than city-scale empirical models informed by data on actual consumption patterns. For instance, Dhakal [9] examines urban energy consumption in Chinese cities using a coarse estimate of energy use in urban areas derived from the energy intensity of economic activities (measured by Gross Regional Product). The study found that cities account for 84% of China's total commercial energy use, and that the 35 largest cities account for approximately 40% of national consumption and carbon emissions. Bennett and Newborough [10] present a framework for auditing citywide energy use. They highlight the importance of the city scale to identifying and promulgating carbon emission reduction strategies. The authors point out the need for various types of data, including surveys and direct monitoring, but do not provide any empirical examples of the citywide audit in practice. The paper highlights the need for such city-scale modeling to predict energy efficiency potential across urban sectors.

In a paper by Lin et al. [11], the authors utilize a LEAP energy simulation model to estimate future energy demand under various policy scenarios. The model applies demographic, socioeconomic, and macroeconomic variables for the city of Xiamen to forecast energy demand across the city over time. The authors rely on coarse consumption data that do not differentiate among building-specific characteristics beyond type of use. Although distinguished from the work of Bennett and Newborough [10] by the recommended data collection and by the fact that Lin et al. attempt to estimate energy use in Xiamen, both approaches rely on relatively low spatial resolution estimates of consumption predicated on numerous assumptions about the patterns and drivers of energy use. In another attempt to develop an urban energy demand model, Brownsword et al. [12] utilize a linear programming approach to estimate current and future energy use. The authors apply a broad typology to urban energy "consumers", separating the city into domestic, commercial, and industrial uses across size bins of small, medium, and large. While the authors stated goal is the replicability of the model, this approach significantly dismisses the variation in urban land use and building typologies and the impact of such characteristics on energy use (see [8,13]). Such assumptions limit the usefulness of an energy model to identify and predict future energy savings measures at scale.

Heiple and Sailor [14] estimate daily property-level energy consumption levels using annual building simulation results for prototypical buildings in the city of Houston, Texas. The authors apply energy simulation outputs from building prototypes and match these prototypes to existing buildings in the city using GIS-based tax lot data. The study attempts to scale up building-level estimations to the entire city. However, the use of prototypical buildings introduces significant error in the prediction, coupled with the uncertainty in matching prototype buildings to existing buildings based on sparse tax lot data. The limited data on actual energy use at the building level and building characteristics constrain the predictive power of the citywide model. Facing similar data constraints, Touchie et al. [15] attempt to identify building characteristics that influence energy consumption in residential buildings. Their findings are significantly limited by a sample size of only approximately 60 buildings, and they rely on multiple data providers to collect building-level characteristics. The limitations of the Touchie et al. study highlight the value of building energy disclosure data given the relatively large sample of buildings covered (particularly in New York City) and the richness of the building features included in the datasets.

Kavgic et al. [16] model the energy use in residential buildings in the context of Belgrade, Serbia. The authors highlight the limitations of city-scale energy models derived from the extrapolation of archetypical building profiles to an entire urban building stock.

The study, as with a majority of the attempts to develop reliable urban energy models, is constrained by the absence of actual energy use data to calibrate and validate the prediction model. Sensitivity analysis provides little value to policymakers without the quantification of the underlying error range of the prediction. As an example, Keirstead et al. [17] conduct a meta-analysis of urban energy systems modeling approaches. The authors indicate the need for activity-based modeling to improve the resolution of energy systems analysis. This requires new computational and data resources to be effective.

In work most directly relevant to this study, Howard et al. [18] attempt to downscale zip code-level energy data to the parcel level using standard end-use allocations by building type as derived from U.S. Department of Energy Commercial Building Energy Consumption Survey (CBECS) data. The significant limitation here is the lack of building-specific energy consumption data to validate their model, which results in an over-reliance on rules-of-thumb and simulation-based reference values. Ewing and Rong [19] focus on how the spatial characteristics of urban areas influence residential energy consumption. Using definitions of compactness that reference land use density at the county level, the study finds that residential units in compact counties use 20% less energy than similar units in sprawling counties. While the authors do not attempt to predict consumption or consumption patterns, the study provides a useful examination of how land use and the design of the built environment can impact energy use at regional geographies.

The prediction of building-level energy consumption has garnered greater attention in the academic literature, particularly as new energy data sources have emerged from city disclosure laws [3]. Zhao and Magoulès [20] provide an overview of several approaches to predicting building energy consumption, including engineering or physical models, statistical regression models, and artificial intelligence approaches. The benefits of using statistical or machine learning approaches to energy use prediction include the potential to achieve greater degrees of certainty in the output over simulation-based or rule-of-thumb approaches, and a more robust understanding of the effects of individual covariates, such as building characteristics, spatial variables, and use types. Coupled with audit data, machine learning techniques can be used to drive an evidenced-based approach to energy use and carbon emission reduction strategies [21].

Using data previously unavailable on actual building energy use data for approximately 20,000 buildings in New York City, Kontokosta [8] analyzes the determinants of commercial building energy consumption across building, system, spatial and occupancy characteristics. The author utilizes a robust multivariate regression model and finds that building size, age, use, occupancy characteristics, construction type, and building adjacencies all factor into a building's annual total energy consumption and intensity. Among other significant results, the study finds that newer buildings are less energy efficient, controlling for all factors in the model, than older buildings, particularly those built before 1930. Baker and Rylatt [22] seek to understand the drivers of energy use in residential buildings in the United Kingdom. Supplemented by survey data, they attempt to analyze the effects of some behavioral aspects of householders on consumption. While their findings on number of bedrooms and homeworker rate deserve further exploration, the very small sample size (n = 142), selection bias in survey responses, and homogeneity in the sample limit the generalizability and significance of the results. In another empirical study, Hsu [23] explores energy disclosure data from the perspective of identifying the minimum set of variables needed to predict consumption. Focusing on 361 multifamily buildings in New York City, the study uses energy use and systems-level data provided by a private third-party energy management company and applies a Bayesian multilevel modeling approach. The analysis finds that historical consumption data are the best predictors of future consumption, and that adding information on building systems adds little explanatory power to the model. While providing an important contribution to the literature, this approach has two significant limitations. First, there is serial correlation in time-series consumption data, and one would expect, *a priori*, that there would be little year-to-year variation in quality-controlled energy use data for a given building. Second, the model does not help to define what factors contribute to observed consumption patterns, and thus what methods might be most effective in reducing consumption given a set of building characteristics.

3. Material and methods

3.1. Data collection and description

One significant contribution of this analysis is the use of a substantial dataset of actual building energy consumption, combined with detailed property- and building-level attributes, covering a diversity of building use types and characteristics. There are three primary data sources utilized for this work: LL84 provides building energy use information and building occupancy and physical characteristics geocoded to the individual building and parcel for 20,000 buildings; PLUTO provides parcel-level physical characteristics, zoning, and use type data for all 1.1 million buildings in NYC; and our zip code energy dataset provides spatially-aggregate annual energy consumption for each of NYC's 176 zip code areas. This study uses these data to predict building-level energy use and energy use intensity for all building types, including office, residential, industrial, etc. More detailed descriptions of the data are provided below.

3.1.1. Local Law 84 energy disclosure data

LL84 is New York City's energy disclosure ordinance, which covers more than 15,000 properties and 20,000 buildings larger than 50,000 square feet. The resultant dataset includes detailed information on energy use by fuel type and source and building occupancy, use, and physical descriptors. A public version of the dataset is released every year on the NYC Open Data portal containing a subset of collected data limited to aggregate energy use and spatial identifiers. A confidential version of these data containing additional fields was provided to the authors by the NYC Mayor's Office of Sustainability, including time series data from 2010 to 2015. These data contain annual energy use, water use, and greenhouse gas (GHG) emissions for all reporting properties, as well as the primary property type (Office, Multifamily Housing, etc.), physical and occupancy characteristics, and the Borough, Block, and Lot (BBL) number used to match the reporting properties with city tax lot (PLUTO) data. Energy use is reported by total consumption and by energy use intensities for both site and source energy, and is normalized for weather. The confidential dataset contains a more specific breakdown of energy use by fuel type, including consumption of electricity, natural gas, steam, and various fuel oils, in total and per square foot. Data for calendar year 2014 were used in this study.

Several data cleaning and filtering steps were conducted prior to analysis [8,24]. The first removed duplicate properties, including entries with more than one BBL number specified, missing BBL numbers, and duplicate entries. The second cleaning step filtered the sample to remove entries with no reported energy use by excluding entries with zero or null values for weather-normalized source energy use intensity[1] (EUI).

[1] Defined as energy consumed per square foot of floor area (kBTU/ft^2), and adjusted for annual weather trends.

Of the reported fields, EUI was used to evaluate data quality and identify outliers in the sample. This field is calculated as the sum of consumption for each energy source divided by the reported gross floor area. The self-reported building size was compared against City administrative records of building size in the PLUTO dataset as a validity check. As such, if any of the component fields in the EUI calculation are drastically misreported,[2] it should produce outlier EUI values. Consumption values for individual energy sources were also analyzed for outliers and data entry errors. The EUI values for the entire dataset were observed to follow a log-normal distribution, and a logarithmic transformation was conducted to produce a normal distribution of the data for outlier detection. Observations were excluded from the analysis if they were greater than two standard deviations from the mean for their respective property type (e.g. office, residential, etc.). This was done for all property types with more than 50 observations; property types with 50 or fewer observations were instead filtered using the global mean and standard deviation of the sample. Given the distribution of the sample, 2.2 percent ($n = 355$) of the observations were removed. Fig. 1 shows the raw and log-transformed EUI distributions for the sample.

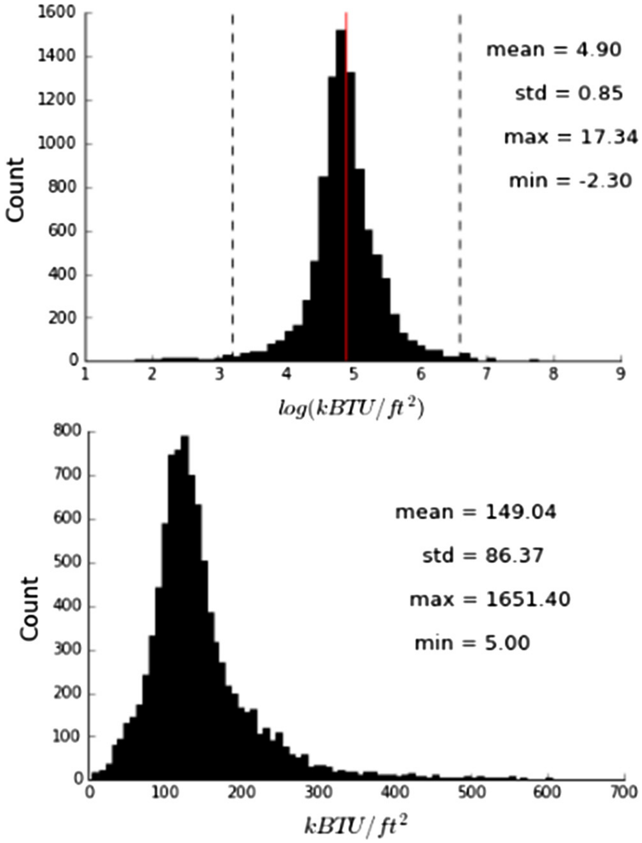

Fig. 1. Top: The histogram of log EUI with red line at distribution mean and dotted lines at ±2 standard deviations from the mean. Bottom: Histogram of EUI after removing outliers for each property type.

3.1.2. PLUTO tax lot and land use data

The NYC Primary Land Use Tax Lot Output (PLUTO) database is an extensive public land use and property dataset maintained by the NYC Department of City Planning. It contains location, land use, and physical characteristics for all 1.1 million properties in NYC, and can be used to identify and geolocate properties based on address, BBL, latitude/longitude geocoordinates, and zip code. In this analysis, PLUTO is used to provide a standardized feature space of building characteristics for our energy use predictions. These data are matched with LL84 to train the predictive model. The data are updated regularly, and version 14v2 was used in this analysis, which includes data current as of 2014.

The variables used in our statistical modeling methodology include bulk/architectural, zoning, and use characteristics for each property in NYC extracted from the PLUTO dataset. The variable definitions are shown in Table 1.

Population and demographic variables from the U.S. Census were considered, but these were not included as they are relevant only for certain property types (e.g. residential), and the census data would require an additional disaggregation step to align with our property-specific methodology, which would increase the margin-of-error in the estimates.

Table 1
Variables included in prediction models.

Log Building Area	Natural log of the gross floor area of the property in square feet
Surface Area to Volume Ratio (SVR)	Approximate ratio of surface area to volume of a property. Assumes a rectangular prism with width equal to lot width, depth equal to lot depth, and height proportional to the number of floors
Floor Area Ratio	The actual, as-built floor area ratio (FAR) of the building. The FAR is calculated as the building area divided by the lot area
Number of Floors	Total number of floors in the building
Inside Lot	A binary variable for whether the building is an inside lot or corner lot
Attached Lot	A binary variable for whether the building is attached or freestanding
Borough	Dummy variable for each of the five boroughs in NYC
Year Built	Year the property was built. For the OLS regression model this is encoded as five dummy variables for properties built 1930 or earlier, 1931–50, 1951–70, 1971–90, 1991 and later
Proportion Residential	Ratio of residential floor area to total floor area
Proportion Office	Ratio of office floor area to total floor area
Proportion Retail	Ratio of retail floor area to total floor area
Proportion Storage	Ratio of storage floor area to total floor area
Proportion Factory	Ratio of factory floor area to total floor area

3.1.3. Zip code energy consumption data

To validate our models, we use actual building energy consumption data for New York City aggregated to the 176 zip code areas, categorized by energy source for electricity, natural gas, steam, and fuel oils (#2, #4, and #5/6). The data were originally obtained from the local utilities Consolidated Edison (ConEd), National Grid and the Long Island Power Authority (LIPA), and were provided for this research by the NYC Mayor's Office of Sustainability. These data are used as a validation when predicting energy consumption in buildings not included in the LL84 sample (i.e. buildings less than 50,000 square feet and/or not required to report energy use data).

Due to data quality issues, only electricity and natural gas data are used in the current analysis. Fuel oil consumption is reported by fuel oil providers based on service deliveries, and the accuracy of these figures, as it pertains to actual consumption, is unreliable. Steam usage is excluded from the prediction models since the number of buildings using steam is constrained to those with the necessary installed infrastructure. For electricity and natural gas usage data, units were converted from kilowatt-hours (kW h) and therms, respectively, to kBtu for comparison with the LL84 energy use data.

[2] E.g. through accidental addition or omission of zeros or improper entry of units, for instance.

3.2. Methodology

Three different statistical models and two feature selection methodologies are compared in order to determine the accuracy attainable by city-scale energy prediction models trained on LL84 energy disclosure data. Accuracy is assessed both at the building level within the LL84 sample, and at the zip code level for all 1.1 million properties in NYC. The building-level output gives a direct measure of the accuracy of the model for predicting consumption in large properties, while the zip code model errors give a measure of model accuracy of the consumption prediction for buildings across the City (those out of the LL84 sample). Our approach is described in three components:

(1) Machine learning models for predicting energy use in buildings at the city scale.
(2) Feature selection using a stepwise selection process to identify feature sets f_g^e for each energy source e (electricity and natural gas) and error granularity g (building and zip code).
(3) Each of three different statistical models is fit using each of the selected feature sets, and used to calculate two error metrics. This results in a total of 3 × 4 × 2 = 24 error scores reflecting the performance of the building-level energy prediction models under a variety of different assumptions.

3.2.1. Predictive model comparison

Three machine learning algorithms are used to predict citywide building energy consumption: OLS Regression, Support Vector Machines (SVM), and Random Forest (RF). Each of these models has unique strengths and weaknesses, and can account for different degrees of nonlinearity [20,25]. Comparing model performance between these models gives insight into whether a simpler, and more easily interpreted, algorithm such as OLS can generalize more efficiently, or whether the nonlinear nature of RF and SVM grant an advantage in prediction accuracy. Each of the three models is fit using predictors determined through feature selection, and errors are evaluated both within the LL84 sample and across NYC as a whole at the zip code level.

3.2.2. Ordinary least squares

OLS is a conventional multivariate regression model that estimates a best-fit line designed to minimize the sum of squared errors (SSE). This linear modeling approach has been used extensively in the literature to predict energy use and to estimate the effects of various features on building energy consumption [8,20,26]. Computationally efficient and relatively straightforward to interpret, OLS models are constrained when there exists nonlinearity in the data and when the presence of outliers in the data can skew the coefficient estimates. The OLS model is given by:

$$y = \alpha + \beta_1 * X_1 + \ldots + \beta_n * X_n + \varepsilon$$

where y is the dependent variable, α is the constant, β are coefficients to be estimated (given by feature selection methods described above), X are a series of independent or explanatory variables, and ε is the error term.

3.2.3. Support vector regression

SVM is a supervised learning algorithm that can be used for classification and regression. We expect it would be appropriate here given the high dimensionality of the sample. Following [27], and applications as in Li et al. [28], Dong et al. [29], and Paudel et al. [30], SVM allows for the estimation of non-linear problems according to the primal classifier:

$$f(x) = g(w^T \varphi(x) + b)$$

where w and b are parameters for weight and bias, respectively, estimated based on the minimization of the regularized risk function for $w \in \mathrm{R}^D$ given by:

$$\min_{w \in \mathrm{R}^D} ||w||^2 + C\sum_i^N \max(0, 1 - y_i f(x_i))$$

where parameter C is a regularization term to control overfitting.

3.2.4. Random forest

Random forest provides what is considered to be one of the most accurate statistical learning algorithms available [31]. Breiman [32] introduced random forests by adding a component of randomness to the bagging of classification trees [33]. Each tree is constructed using a different bootstrap sample, and each node is divided based on a randomly selected set of predictors specific to that node [34]. The result is a classifier that is robust to overfitting, and random selection of predictors minimizes bias. However, the algorithm operates as a "black box" since individual trees cannot be observed, thus limiting the broad interpretability of results [35]. Although the importance of individual predictors can be evaluated, the ability to describe and present the results of a particular algorithm to a non-technical audience is an important consideration in the case of energy use prediction and its application to energy policy and decision-making. The results of the three approaches described here are presented in the next section.

3.2.5. Model fitting and error calculation

Our model fitting and error calculation approach is described here and illustrated in Fig. 2. First, the cleaned dataset D is randomly split into a training set D_{train} containing 70 percent of the sample and a test set D_{test} containing the remaining 30 percent. The test set, also referred to as a holdout set, is untouched by any of the model fitting procedures and is used to calculate the prediction error of the models on the buildings within LL84. This process produces a more reliable estimate of the out-of-sample prediction error.

In the case of ordinary least squares (OLS) regression, the model is fit using the dataset D_{train} and produces an estimate of prediction error. In the case of random forest and SVM, a process is used to first determine acceptable hyperparameters of the models. The method used here is similar to that described in Jain et al. [36,37]: a large grid G of plausible hyperparameter values is defined and a model is fit and scored for each set of hyperparameters θ_i in G. To avoid overfitting, error scores are calculated using fivefold cross validation. Cross validation proceeds by randomly splitting the dataset D_{train} into five nearly equal subsets (s). For each subset s_i, the model is fit to the data in the other four subsets and then used to predict energy use and calculate absolute errors for the buildings contained in subset s_i. After errors have been calculated for each subset, the mean absolute error ε of the predictions is calculated and used as the score for the model under the defined set of hyperparameters. From the score for each set of hyperparameters in G, the best performing hyperparameter set θ^* is defined as having the lowest error score ε^*. A smaller grid G' is then centered at θ^* and the process is repeated in a recursive manner until the absolute change in ε^* from one stage to the next is smaller than a defined threshold. The starting values and final selections for hyperparameters of the RF and SVM models are shown in Table 2 below. Any hyperparameters not displayed were left at the defaults used by the scikit-learn python package [38].

After the selection of relevant hyperparameters, EUIs are then predicted for each building and the accuracy of the prediction is assessed in two ways. The right path in Fig. 2 shows how accuracy is assessed at the building level for properties within D_{test} using the

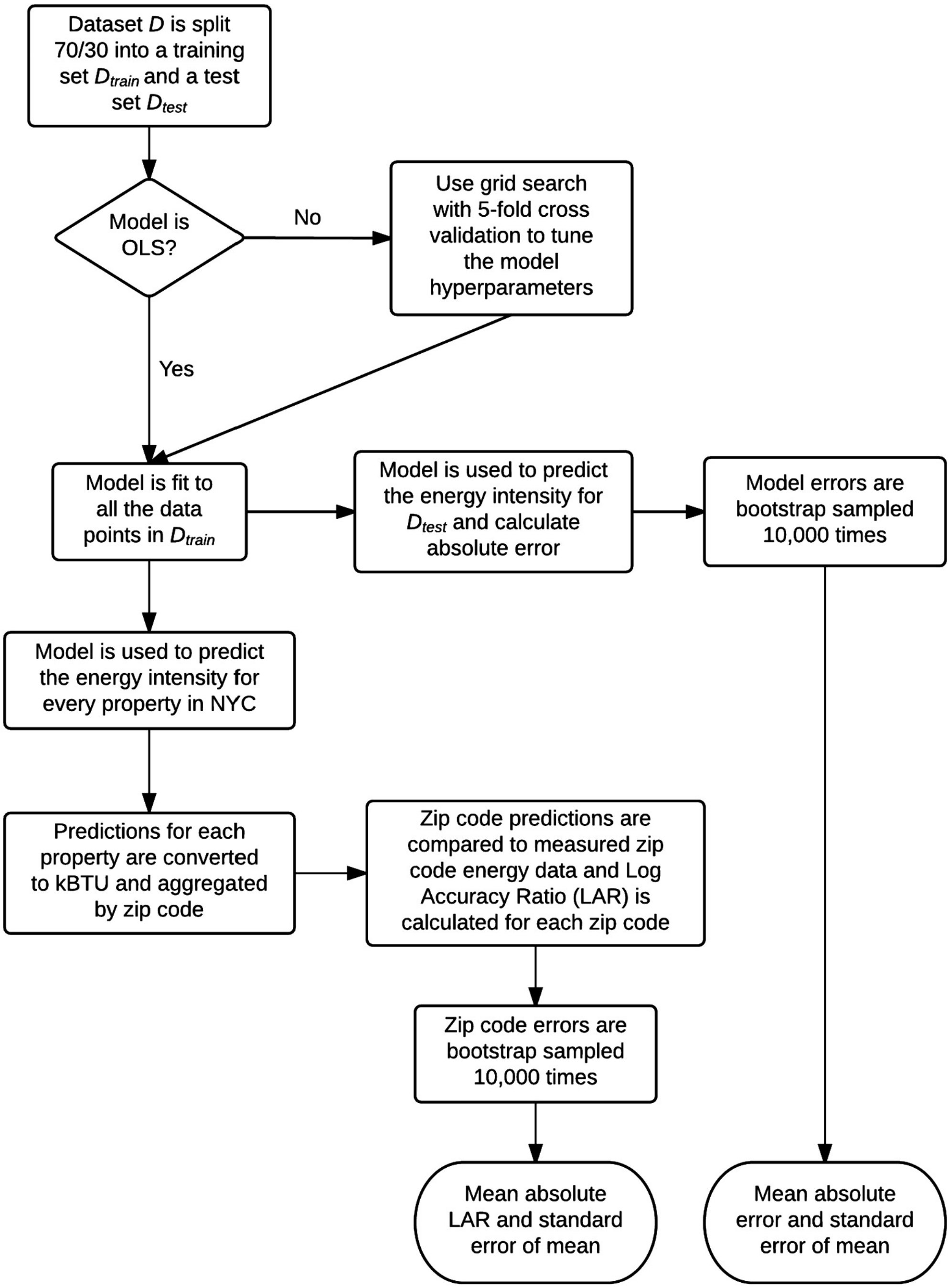
Dataset D is split 70/30 into a training set D_{train} and a test set D_{test}
Model is OLS?
No
Use grid search with 5-fold cross validation to tune the model hyperparameters
Yes
Model is fit to all the data points in D_{train}
Model is used to predict the energy intensity for D_{test} and calculate absolute error
Model errors are bootstrap sampled 10,000 times
Model is used to predict the energy intensity for every property in NYC
Predictions for each property are converted to kBTU and aggregated by zip code
Zip code predictions are compared to measured zip code energy data and Log Accuracy Ratio (LAR) is calculated for each zip code
Zip code errors are bootstrap sampled 10,000 times
Mean absolute LAR and standard error of mean
Mean absolute error and standard error of mean

Fig. 2. Schematic overview of the process used to train predictive models and calculate the mean prediction error. The same general process is used for each energy type (electricity, natural gas) and each statistical model (OLS, Random Forest, SVM). Note the diverging paths that produce both build-level error and zip code errors using two different error metrics.

Table 2
The initial grid locations and final selections for each of the chosen hyperparameters.

Model		Hyperparameters			
RF		Max depth	Max features	Min leaf samples	# Trees
	Initial grid	2, 4, 8, 12, 16	0.3, 0.6, 0.9	2, 4, 8, 12, 16	256
	Final selection	14	0.3	2	256
SVM		Kernel	C	Gamma	Epsilon
	Initial grid	RBF, Linear	0.01, 0.1, 1, 10, 100	0.01, 0.1, 1, 10, 100	0.1
	Final selection	RBF	1	0.1	0.1

actual EUI values available through LL84 and the mean absolute error (MAE). The MAE is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|p_i - m_i|$$

where p_i is the predicted log EUI of property i and m_i is the observed log EUI of the same property. MAE was chosen over competing metrics such as the mean squared error to avoid placing undue weight on outliers remaining in the data after cleaning. After obtaining a distribution of absolute errors for the properties in D_{test}, the bootstrap method with 10,000 iterations is used to estimate the sampling distribution of the mean. The mean and standard deviation of the sampling distribution become the reported MAE and the standard error of the MAE, respectively.

The left path in Fig. 2 details the steps taken to estimate prediction error for all properties in NYC aggregated to the zip code level for validation. After the model is trained, log EUI for electricity and natural gas use are predicted for each of the 1.1 million properties in NYC based on the property characteristics selected from the PLUTO dataset, discussed above. A number of properties (5.3 percent of the total) were missing information on year of construction, so a reduced model was specified without this feature. The predicted log EUI is converted to total annual energy use in *kBtu* through exponentiation of the predicted log value and multiplying by property area. The total predicted energy consumption for each individual property within a given zip code are then summed to obtain a predicted total annual energy use for each zip code. The predicted energy p_z of zip code z is given by:

$$p_z = \sum_{j=1}^{N}\exp(p_j) \times A_j$$

where N is the number of buildings within zip code z, p_j is the predicted log EUI of building j and A_j is the gross floor area of building j. The accuracy of the zip code predictions is then assessed using the mean log accuracy ratio (LAR), defined as

$$Mean\ LAR = \frac{1}{M}\sum_{z=1}^{M}\log\left(\frac{p_z}{m_z}\right)$$

where M is the number of zip codes, p_z is the total predicted annual energy use of zip code z and m_z is the measured (actual) annual energy use of zip code z as obtained from the local utilities. The LAR metric was chosen for several reasons [39]. First, a relative measure of predictive accuracy is appropriate given the significant variation in zip code total energy consumption, and other accuracy metrics, such as the MAE and MSE, would be influenced by the effects of the largest energy-consuming zip codes. Second, the mean LAR was chosen over other relative measures, such as mean absolute percentage error (MAPE), because it is symmetric around zero and has been shown by Tofallis [39] to provide for more accurate model selection.

3.2.6. Feature selection

The approach described above provides a robust prediction methodology and model validation, but it does not address feature selection. The selection of appropriate features to include in the model can have a non-trivial effect on the accuracy of predictive models [40]. This is especially true in the context of this study where models are being fit to a truncated sample of the population (large buildings) and statistical relationships that hold within the training sample may not generalize to the city as a whole.

We address the problem of feature selection using a stepwise selection procedure to choose a feature set f_g^e for each energy source type and prediction level (building or zip code). Here e is the energy type and corresponds to either electricity or natural gas, and g is the spatial resolution of the errors, either building-level or zip code-level. Thus f_{ll84}^{NG} is the feature set that results in the lowest building-level prediction error when predicting natural gas use intensity.

The stepwise selection process is illustrated in Fig. 3. It begins by fitting a series of OLS regression models and adding a single predictor at each stage. The first stage begins with only an intercept and one predictor. The predictor corresponding to the model with the lowest prediction error is then added to the feature set at each stage. The process then repeats and another round of regression models is estimated. This continues until an error score has been calculated for the full model with all features. The result is a greedy

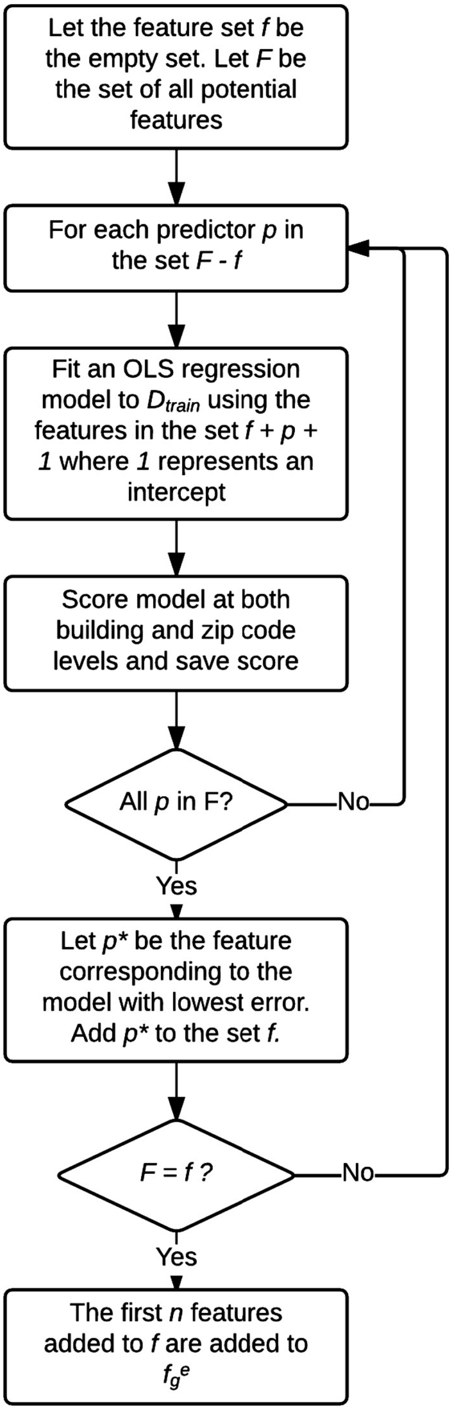
Let the feature set f be the empty set. Let F be the set of all potential features
For each predictor p in the set F - f
Fit an OLS regression model to D_train using the features in the set f + p + 1 where 1 represents an intercept
Score model at both building and zip code levels and save score
All p in F?
No
Yes
Let p* be the feature corresponding to the model with lowest error. Add p* to the set f.
F = f ?
No
Yes
The first n features added to f are added to f_g^e

Fig. 3. Stepwise model selection process.

algorithm that attempts to select a feature set with performance comparable to that obtained with an exhaustive search, but in a more computationally efficient manner. The algorithm is greedy in the sense that at each step the chosen feature p_i is optimal, conditional on the inclusion of the previous $i - 1$ features.

This process allows us to empirically analyze the tradeoff between model accuracy and model complexity [40]. Additionally, it provides a rough ordering of feature importance in terms of how each feature contributes to model accuracy. This ordering is imperfect, but useful. It is imperfect for the reason noted before that selection at each stage is conditional on the inclusion of the features selected at all previous stages. It is useful because it allows one to compare across energy source types and spatial resolutions. Specifically, one may compare the first few features selected as these may be interpreted as the most significant predictors of building energy use.

4. Results

We present our feature selection, EUI prediction, and model validation results below. The accuracy of our selected algorithms is discussed, together with the output of the OLS model and significance of individual predictors. We find the OLS model performs best when generalizing to the City as a whole, and SVM results in the lowest MAE for predicting energy use within the LL84 sample. Our median predicted electric EUI for office buildings is 71.2 kbtu/sf and for residential buildings is 31.2 kbtu/sf with MALAR of 0.17. Fig. 4 shows a map of our predicted site electricity use intensity for each building in NYC.

4.1. Feature selection

The feature selection results are shown in Fig. 5. We observe that the majority of the error reduction comes from the first few features. After approximately six features, accuracy ceases to improve, or, in the case of zip code errors, the models become less accurate as more predictors are added, most likely due to over-fitting. The contrast between the error ranges at the building- and zip code-levels is a function of data sampling. At the building-level, models are fit to data from within the LL84 sample and errors are evaluated on properties also from the same sample of large buildings. The randomization process to generate training and test groups ensures a relatively homogeneous mix of property types and attributes, leading to more predictable performance between these sets. The result is the monotonically decreasing errors visible in the left column of Fig. 4. The city-wide sample used to calculate zip code errors, on the other hand, involves no such expectation of homogeneity. In fact, the training set of large properties present in LL84 is known *a priori* to be fundamentally different from the broader NYC building stock, at least with respect to property size, since only buildings larger than 50,000 square feet (or multiple buildings on a lot totally more than 100,000 square feet) are required to report consumption data. This results in an apparent over-fitting effect whereby the first features selected capture statistical relationships that generalize well to the city at large, while later features capture relationships that are only true within the LL84 sample. This causes the model accuracy to begin to decrease after a certain point as less generalizable features are added.

To assess accuracy across different types of statistical models, four feature sets were then chosen by selecting the first six features

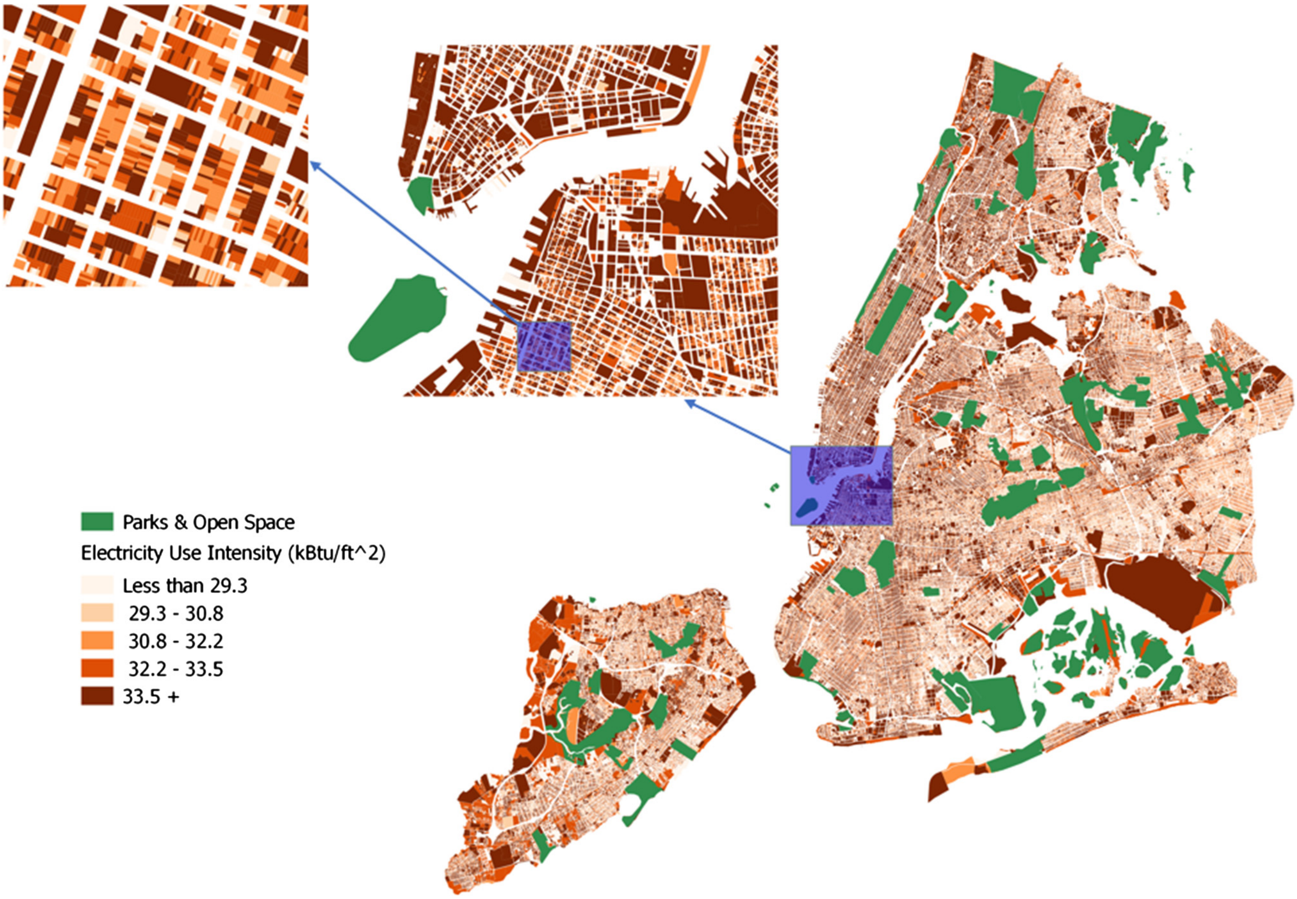

Fig. 4. Parcel-level map of NYC and the predicted electric energy use intensity for each property.

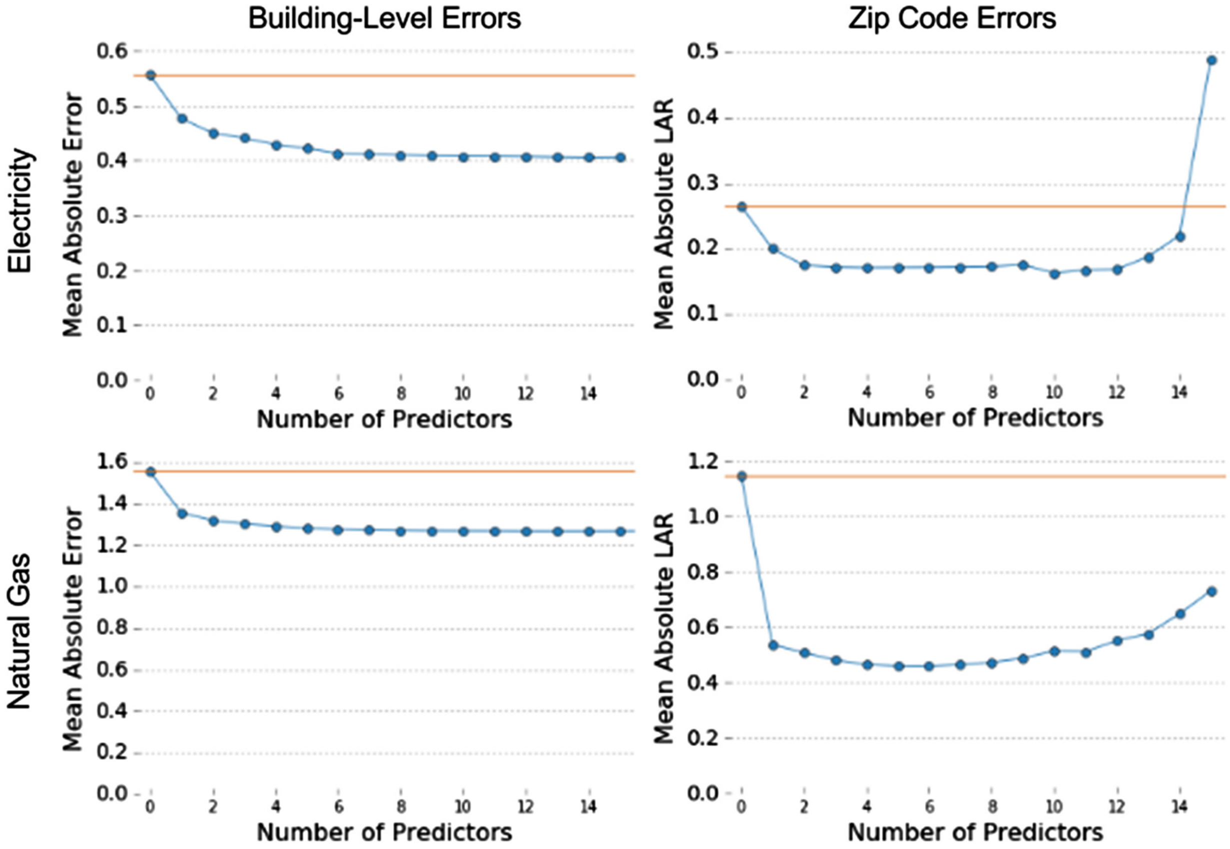

Fig. 5. The charts show the error at each step of the stepwise variable selection process as a new predictor is added to the linear regression model. The left column shows the mean absolute error at the building level when predicting energy for buildings within LL84. The error decreases monotonically as more predictors are added, although the curve flattens out after the first six. The right column shows the progression of errors at the zip code level. In all cases the horizontal orange line represents the error of a "naive" model with only an intercept, which is equivalent to prediction using the overall distribution mean.

Table 3
From left to right each column represents one of the feature sets f_{ll84}^{E}, f_{zip}^{E}, f_{ll84}^{NG}, and f_{zip}^{NG}, respectively. From top to bottom, the Feature # represents the order in which features were selected using the stepwise selection process. On can see that there is no general agreement between the building-level and zip code-level as to which features are the most predictive.

	Electricity		Natural Gas	
Feature #	LL84	Zip Code	LL84	Zip Code
1	Proportion Res.	Proportion Office	Borough	Log Building Area
2	Number of Floors	SVR	Year Built	SVR
3	Year Built	Built FAR	Log Building Area	Built FAR
4	Borough	Attached Lot	Proportion Storage	Number of Floors
5	Proportion Office	Inside Lot	Proportion Office	Proportion Garage
6	Proportion Retail	Proportion Storage	Attached Lot	Inside Lot

for each energy source type and spatial granularity (electricity use at building-level and zip code and natural gas use at building-level and zip code). Six features were chosen because, as is visible in Fig. 5, the accuracy in all four scenarios ceases to improve with additional predictors. The four feature sets chosen are presented in Table 3.

It is important to note the differences between the predictors that appear in the building-level models and those that appear in the zip code-level models. The *Proportion Residential* and the fixed effect for the *Borough* in which a building is located are the most significant predictors for electricity and natural gas use, respectively, among large LL84 properties. Yet neither of these appear to generalize to the city at large. The same is true of *Year Built*, which is an important feature in both building-level models, but does not appear in the zip code models. On the other hand, *SVR*, *Built FAR*, and *Inside Lot* all appear in the zip code models, but not in the building-level models, indicating that these effects generalize well. This point is examined further in the results.

4.2. Prediction model performance

Fig. 6 presents the overall error for each energy type, error spatial granularity, and predictive model specification. Error at the building-level among LL84 properties is assessed using MAE, while the error for aggregate zip code energy use is assessed using the mean absolute LAR. Error bars correspond to standard errors of the mean.

The colors of the points signify which feature set was used to fit the model. Orange denotes a feature set chosen as minimizing building-level error in the OLS model, while blue denotes a feature set chosen to minimize error across zip codes. The different feature sets are compared side-by-side to illustrate two modeling deci-

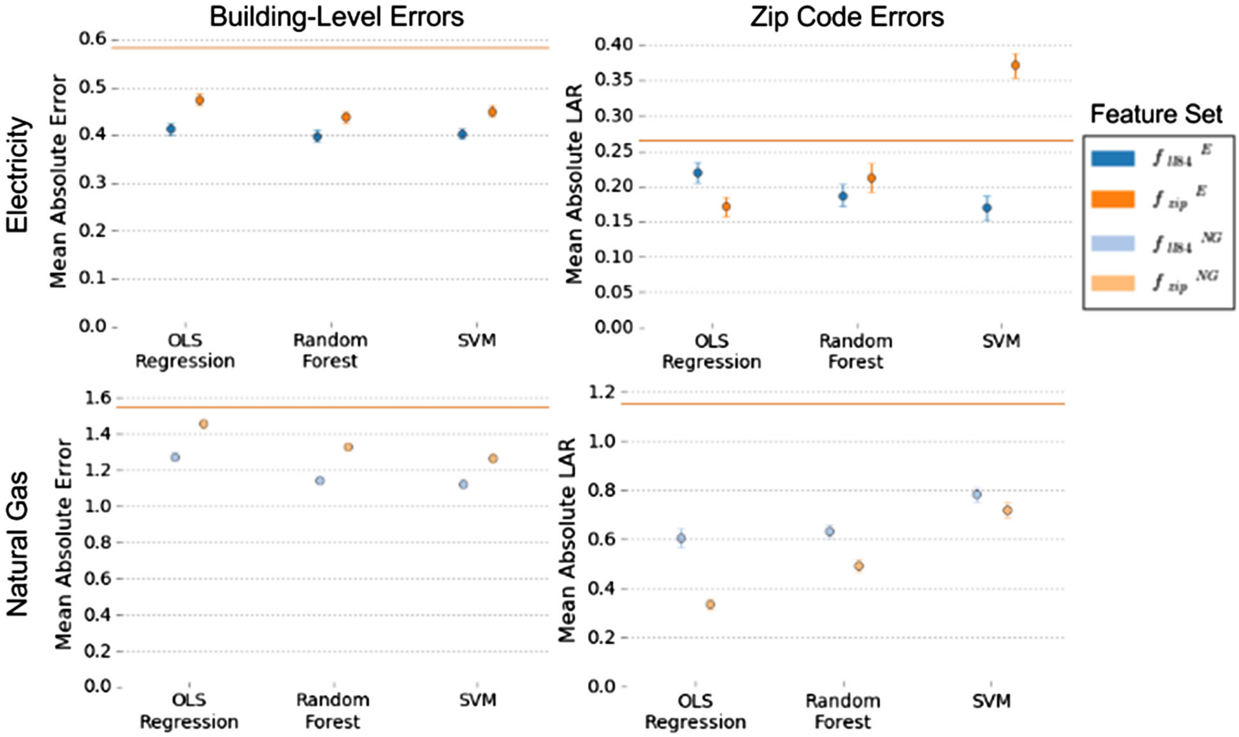

Fig. 6. Accuracy estimates derived using the methodology described in Section 3. The left column shows the errors among LL84 properties, while the right column shows accuracy at the aggregate zip code level. The top row shows electric energy intensity prediction errors, and the bottom row displays errors for natural gas intensity. Blue points are the errors achieved by models using the LL84 feature sets, while orange points correspond to zip code feature sets. Error bars display the bootstrapped standard error of each estimate. The horizontal orange line is the error of a naive model using the overall distribution mean to predict. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sions. The first is that the choice of features can have a significant effect on model performance. More importantly, it shows that a model that performs well among LL84 buildings will not necessarily do so among the broader population of buildings in NYC. This difference in model performance is important because it indicates that the statistical relationships that hold for large properties may not be true for smaller properties.

Specifically, the left column of Fig. 6 shows the MAE for the LL84 buildings in the test set D_{test}. Model performance is similar across model specifications, with a slight improvement in the nonlinear models. In all cases, the feature set optimized for LL84 performs better than the set optimized for city-wide (zip code) prediction. In contrast, the zip code errors show greater variation across models. At best, the trend is reversed (bottom right, predicting natural gas use) with the simpler OLS model being preferred over the nonlinear models. This can be partly explained by the truncated nature of the training data. The nonlinear models seem more likely to over-fit to the training data, capturing statistical relationships that may exist only among large properties. This over-fitting then leads to decreased performance when predicting outside of the LL84 sample, where a large majority of the buildings are less than 50,000 square feet. Models for predicting electricity use reveal mixed results based on MAE. Again, the OLS model generates the lowest absolute error, while the SVM and RF results do not demonstrate superior performance in this instance.

Some care must be taken when interpreting the zip code errors since these represent a summation of the building-level errors within that zip code. When the predicted energy use in a zip code deviates significantly from measured, this is likely due to many of the individual properties deviating in the same direction (either positive or negative). Conversely, if the aggregate error in a zip code is small, this is either due to the errors of the individual properties being small (and producing a small sum) or potentially due to properties with large individual errors that cancel out during the summation due to differing signs, producing a small error in the aggregate. In either of these cases, however, the model is still producing accurate predictions of zip code energy use using a building-level model. In addition, the errors from the in-sample (LL84 building-level) prediction give approximate upper and lower bounds to our expectations for out-of-sample estimation errors.

4.3. Predicting natural gas consumption patterns

The natural gas prediction models tend to be much less accurate than the electricity models. This is because patterns in natural gas use are less consistent than electricity use, and seem to follow at least two distinct consumption profiles. The first type corresponds to properties that use steam or fuel oil for the majority of their heating needs, while using a relatively small amount of natural gas for cooking or other ancillary purposes. The second type relies almost exclusively on natural gas for heating and domestic hot water.

These distinct patterns are visible in Fig. 7. Type 1 properties use little natural gas relative to other fuels and appear on the left of (a) while Type 2 properties use primarily natural gas and appear on the right of (a). This dynamic is also visible in the histogram of log natural gas use intensity. The histogram is bimodal, but it can be separated into two unimodal distributions by separating out Type 1 and Type 2 properties.

Given our attempt to predict consumption at the city scale, without *a priori* knowledge of energy sources for all individual buildings, we attempt to classify buildings as either Type 1 or Type 2 based on available property characteristics. This approach is pursued here using a binary logistic regression classifier with balanced class weights, assigning the value 1 to Type 1 buildings and 0 to Type 2 properties. The model is fit under both feature sets f_{ll84}^{NG}

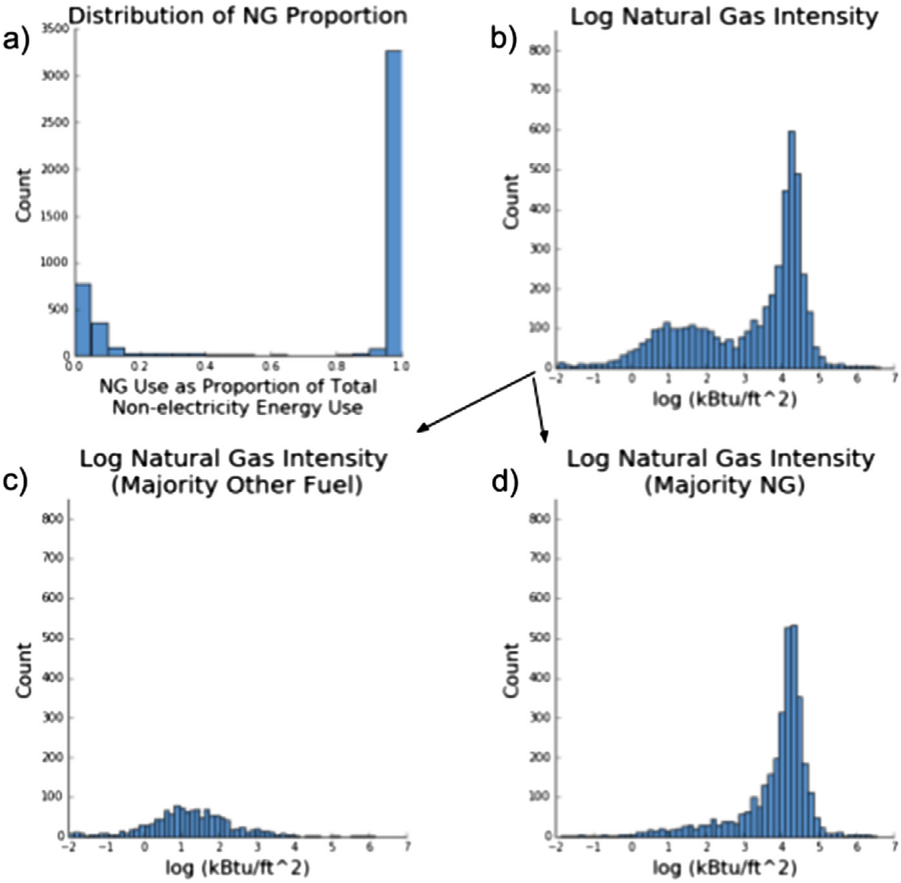

Fig. 7. The data show two different natural gas consumption patterns in the city. The distribution of natural gas use shows a split between properties which use natural gas as a majority of non-electricity energy and those with natural gas as a minority (a). The distribution of NGUI is bimodal (b) but separable by exploiting knowledge about these two different patterns (c and d).

Table 4
Regression model output for in-sample and out-of-sample models.

	Using LL84 Feature Selection		Using Zipcode Feature Selection	
Y=	Log Electric EUI	Log Natural Gas EUI	Log Electric EUI	Log Natural Gas EUI
N =	4149	4866	4149	4866
R2 =	0.325	0.256	0.166	0.109
F-statistic =	165.6	139.1	137.3	99.4
Year Built 1931–1950	0.026	0.171**	–	–
Year Built 1951–1970	0.060**	0.245***	–	–
Year Built 1971–1990	0.380***	0.832***	–	–
Year Built 1991-Present	0.404***	1.034***	–	–
Borough – Bronx	0.101***	−0.920***	–	–
Borough – Manhattan	0.258***	−1.524***	–	–
Borough – Queens	−0.100***	−0.307***	–	–
Borough – Staten Island	−0.336***	−0.861***	–	–
% Office Floor Area	0.550***	−1.252***	0.984***	
% Retail Floor Area	1.197***	–	–	–
% Residential Floor Area	−0.395***	–	–	–
No. of Floors	0.008***	–	–	0.023***
% Storage Floor Area	–	−1.380***	−0.096	–
Attached building/lot?	–	−0.449***	0.025	–
Total Gross Floor Area	–	−0.321***	–	−0.231***
SVR	–	–	−0.071***	−0.087***
Built FAR	–	–	0.016***	−0.134***
Inside Lot	–	–	−0.020	−0.068
% Garage Floor Area	–	–	–	0.795**
Intercept	2.921***	7.362***	2.836***	6.208***

* Significant at the 90% confidence level.
** Significant at the 95% confidence level.
*** Significant at the 99% confidence level.

and f_{zip}^{NG}. Accuracy is assessed using the area under the ROC curve (AUC), a common accuracy metric for binary classifiers that corresponds to the probability that a randomly drawn positive (value 1) property will be ranked higher than a randomly drawn negative (value 0) property. An AUC of 1 thus indicates a perfect classifier, while an AUC of 0.5 indicates a classifier that performs no better than random.

Initial results are promising, with an out-of-sample AUC of 0.78 on LL84 properties in the holdout set. Direct validation of the model is not possible for the city at large without specific data on the locations and connections of the natural gas distribution infrastructure, but initial analysis estimates that 13.6% of properties in NYC are Type 1 properties that use minimal natural gas relative to other fuels. This contrasts with 28.5% of LL84 properties being Type 1. This discrepancy is expected given the large number of buildings in the LL84 sample that use fuel oil or district steam for heating.

4.4. Regression model results

Finally, the OLS model output provides some explanatory insight into what drives consumption in urban buildings. The

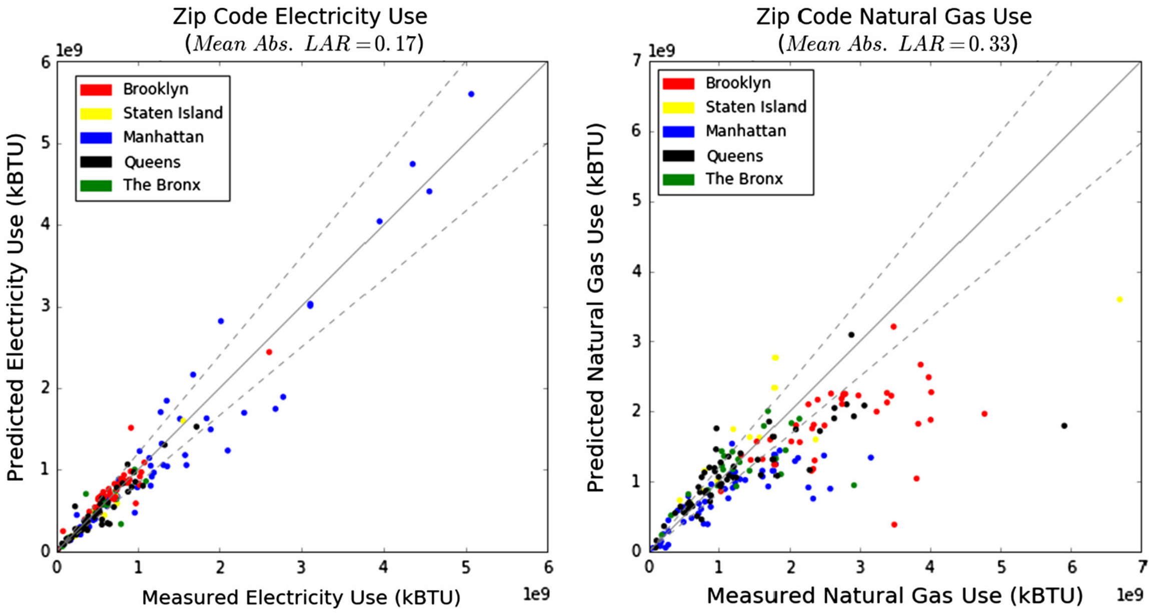

Fig. 8. Measured versus predicted energy use for each zip code derived from the OLS regression models. Each point represents a zip code colored by Borough. The solid grey line at 45 degrees represents zero error where predicted use is equal to measured use. The dashed grey lines represent ±20% accuracy.

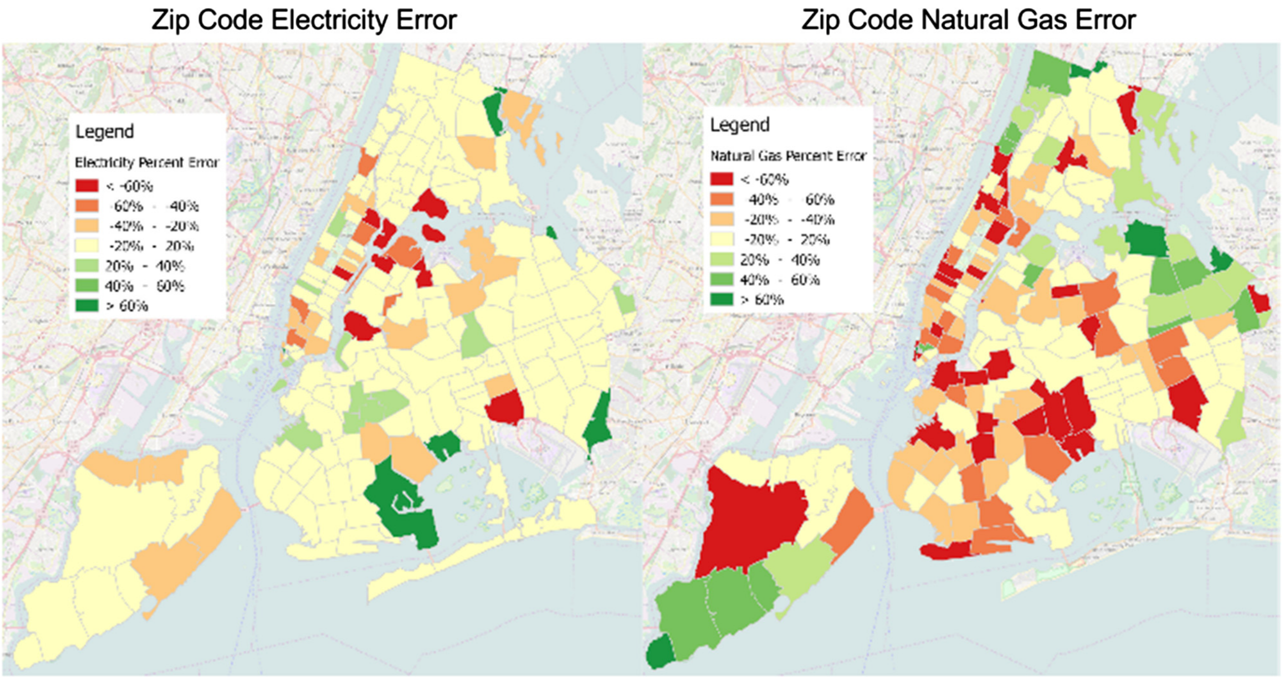

Fig. 9. Percent error of the zip code OLS prediction models displayed on a map of NYC. Red zip codes use more energy than predicted, while green use less. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

results of the OLS regression model, including features based on the selection results above, are shown in Table 4.

As expected, the OLS model predicts energy use in the LL84 sample buildings more accurately when using features selected from the in-sample (LL84) training set. In the LL84 models, building age is found to be a significant predictor of energy use, with newer buildings (particularly those built since 1991) found to have higher consumption levels than those constructed before 1930, given as the reference case. This is consistent with previous findings [8,13].

We also find higher electric energy consumption in office and retail buildings, although the sign is reversed for natural gas. In general, larger buildings use less energy per square foot, while taller buildings with more stories, controlling for floor area, use more energy per square foot. Attached buildings – those with adjacent buildings and a shared party wall – are found to have lower natural gas EUI. This finding is consistent with the thermal properties of buildings sharing a perimeter wall, and thus have adjacent conditioned space on at least one dimension of the structure.

4.5. Spatial dispersion of errors for top-performing models

Figs. 8 and 9 show a detailed view of the errors of the top-performing OLS regression model at the zip code-level for each electricity and natural gas use prediction, respectively. While the electric energy use model produces errors that are symmetric around zero, the natural gas model tends to under-predict in zip codes with high natural gas use. This could be due to a lack of representation in the LL84 data of buildings in these zip codes, which results in over-fitting to the large buildings used in the training set. The prediction for natural gas is also complicated by the bimodal distribution of natural gas use in buildings discussed previously.

4.6. Errors by building size

To explore the role of building size in zip code prediction error, Fig. 10 divides zip codes into four size categories using the median gross floor area of properties within the zip code (greater than 50,000 ft^2, 10,000–50,000 ft^2, 2,000–10,000 ft^2, and less than 2000 ft^2). The objective here is to determine the prediction error where buildings deviate significantly by size from the LL84 sample. Each of the plots corresponds to the OLS regression model fit with the stated feature set. Model results follow expectations: the building-level feature set minimizes errors on large properties most similar to those in LL84 and performs less well in predicting energy use in smaller properties (a), while the models chosen to minimize zip code error appear to do so by minimizing error among smaller, more numerous properties (b, d). The exception here is the model shown in (c) using the building-level feature set. Despite being optimized for LL84 properties, the model performs poorly among zip codes with large median property sizes. This again most likely relates to the challenge in determining the extent of natural gas use.

5. Conclusions and applications

Understanding building energy use at the city scale is a critical component of advancing urban sustainability, carbon reduction,

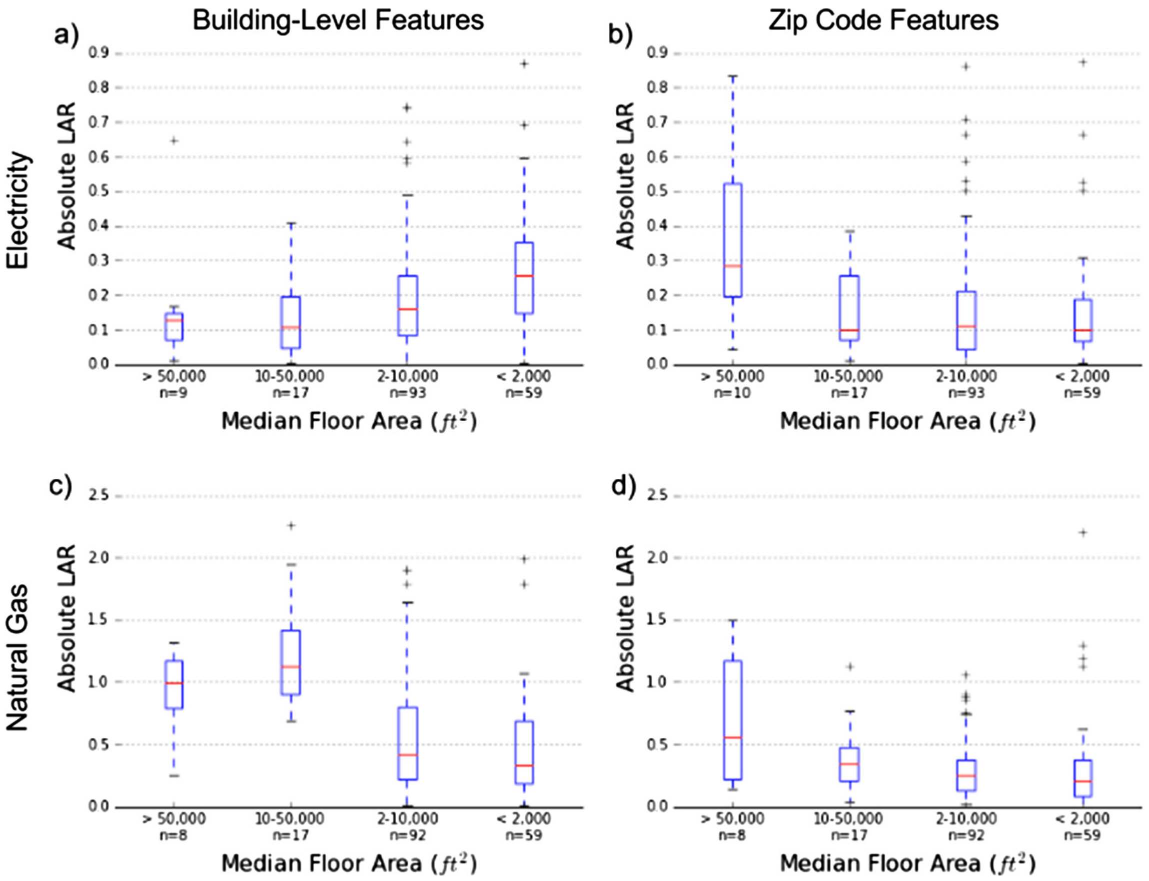

Fig. 10. Zip code errors binned by median floor area of the properties within the zip code. Each plot shows errors from the OLS model fit with each of the four feature sets. Red lines mark the median zip code in each category. Boxes mark the first (Q1) and third (Q3) quartiles. Whiskers extend to the furthest point within 1.5 times the interquartile range (1.5 * (Q3 – Q1)) outside of the boxes. The counts on the x-axis refer to the number of zip codes in each category. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and energy efficiency across the globe. Numerous cities have adopted energy use and carbon reduction plans focused on building efficiency, particularly for the existing building stock. To support these goals, the diffusion of energy disclosure polices, requiring buildings of a certain size to report energy use, has created new streams of data that can provide insight into urban energy dynamics and spur data-driven strategies for achieving greater energy efficiency.

The analysis presented here addresses the question of the generalizability of self-reported energy use data from a relatively small sample of buildings to the city as a whole. Using three different statistical learning algorithms and feature selection approaches, the results suggest that the data from the LL84 sample can produce reasonably accurate predictions of energy use across the city at the building scale, validated at the spatial aggregation of the zip code, particularly for electric energy use intensity.

Overall, we find little difference in machine learning methods used, based on the resultant LAR and MAE values. SVM provides the most accurate prediction for estimation of energy use within the sample of LL84 buildings, while OLS results in the lowest MAE when predicting total building energy consumption at the zip code-level for the entire city. The findings from the comparative analysis of algorithm results indicates that all three methods preform reasonably well and report a narrow range of MAE values. The straightforward and easily interpretable OLS models perform best overall, when considering both in-sample and out-of-sample prediction results.

We also find that building use, size, and morphology emerge as robust predictors of energy use at the building- and zip code-levels. The nature of the building occupancy, whether office, retail, or other use type, impacts electric and natural gas EUI. Larger buildings are found to be less energy intensive, while taller buildings, controlling for size, are more intensive. The shape of the building, operationalized as the surface-to-volume ratio, is also found to influence EUI, although additional study is warranted. Our approach is partially constrained by the temporal frequency (annual) of our data, as additional insight into seasonal and diurnal patterns through monthly and daily data could prove valuable. However, such data are not collected through disclosure policies, and there is often a trade-off between the temporal and spatial frequency of data for an entire city, due to limited availability of such data and privacy concerns.

The findings presented here create new opportunities for data-driven energy policy in cities and the ability to evaluate the impact of policy alternatives to advance energy use reductions. There are three specific applications of the model presented. First, the prediction of building-level consumption at the city scale, in this case accounting for the 1.1 million properties in New York City, can be used to estimate energy use in buildings not covered by current energy disclosure policies or where data are otherwise unavailable. Our building energy use prediction, validated at both the building and zip code scale, can provide energy policy-makers with a more complete understanding of the spatial distribution of energy consumption. Second, this city-wide understanding can enable targeted regulations and incentives by building type and geographic cluster of intensive energy users to more effectively allocate limited city resources, whether it be incentives for retrofits or building inspections and enforcement, to reduce energy use. As more cities begin to collect building energy data from a subset of buildings, it will be possible to develop a more accurate, high-resolution model of energy dynamics within and across cities and find common patterns of consumption and transferrable policy responses. Last, our prediction model provides a baseline to evaluate building and city-scale energy performance, and to assess the impact of changes in land use, energy policy, and other energy use drivers over time. Reliable building-level estimates of energy use can be used to generate higher resolution models of carbon emissions from the urban building stock, and directly contribute to the implementation and design of urban sustainability plans.

The results also suggest several opportunities to improve the scope of data collection of energy disclosure policies to facilitate city-wide energy analysis and tracking. First, rather than focusing on collecting data for buildings over a certain size threshold, it would be more useful – from the perspective of understanding city energy dynamics – to cover a diversity of building types, sizes, and locations. While there is certainly a cost to report energy data for building owners – a cost that may be more acute for small building owners and single-family homeowners – the value of these data would be significant to improving the accuracy of city energy models.

Second, this analysis identified the greater challenge of predicting natural gas usage given both the bimodal distribution of gas consumption (some buildings only use gas for cooking fuel, while others also use it for heat and hot water) and the uncertainty around the specific locations of natural gas distribution infrastructure. Collecting data on whether a property is adjacent to a natural gas branch or main line would help to determine if gas usage is feasible, thus improving the prediction of primary energy source and heating fuel type in individual buildings. Similar data collection for other critical infrastructure, such as district steam or distributed generation facilities, would be equally valuable.

Finally, this research highlights the need for greater data transparency and data access from utility providers. Utility companies hold valuable data resources that can help cities plan for, and evaluate, sustainability and carbon reduction strategies. Such data can be used to validate models of consumption and policy impacts, allowing for improved forecasting of energy use and a more granular understanding of urban energy dynamics. Several utilities have begun to make these data more widely available, if only to individual owners themselves due to privacy and confidentiality concerns, but greater access to these data could have significant implications for urban energy efficiency and national energy demand. This impact can be magnified through the merging of public and private data sources that span a range of physical, use, behavior, and environmental characteristics.

Building on the findings and methodology presented here, future studies could adapt similar machine learning models to predicting retrofitting and energy savings opportunities at scale, to segment buildings using various classifiers to target poorly-performing buildings, and estimate future energy demand based on anticipated changes to land use and zoning and the expected scale of urban development. Energy disclosure data, coupled with city administrative and land use records, represent a critical resource for urban energy planning and long-term energy use reduction applications.

Acknowledgements

The authors would like to thank the New York City Mayor's Office of Sustainability for providing the data necessary for this research. This material is based upon work supported by the National Science Foundation under Grant No. 1653772.

References

[1] Bassett E, Shandas V. Innovation and climate action planning: perspectives from municipal plans. J Am Plann Assoc 2010;76(4):435–50.

[2] Kontokosta CE. Greening the regulatory landscape: the spatial and temporal diffusion of green building policies in US cities. J Sustain Real Estate 2011;3(1):68–90.

[3] Kontokosta CE. Energy disclosure, market behavior, and the building data ecosystem. Ann N Y Acad Sci 2013;1295(1):34–43.

[4] Wheeler SM. State and municipal climate change plans: the first generation. J Am Plann Assoc 2008;74(4):481–96.

[5] City of New York. New York City Local Law 84 Benchmarking Report, August 2012. New York, NY: Mayor's Office of Long-Term Planning and Sustainability; 2012. <http://www.nyc.gov/html/gbee/html/plan/ll84_scores.shtml>.

[6] City of New York. New York City Local Law 84 Benchmarking Report, September 2013. New York, NY: Mayor's Office of Long-Term Planning and Sustainability; 2013. <http://www.nyc.gov/html/gbee/html/plan/ll84_scores.shtml>.

[7] City of New York. New York City Local Law 84 Benchmarking Report, 2014. New York, NY: Mayor's Office of Long-Term Planning and Sustainability; 2014. <http://www.nyc.gov/html/gbee/html/plan/ll84_scores.shtml>.

[8] Kontokosta CE. A market-specific methodology for a commercial building energy performance index. J Real Estate Financ Econ 2015;51(2):288–316.

[9] Dhakal S. Urban energy use and carbon emissions from cities in China and policy implications. Energy Policy 2009;37(11):4208–19.

[10] Bennett M, Newborough M. Auditing energy use in cities. Energy Policy 2001;29(2):125–34.

[11] Lin J, Cao B, Cui S, Wang W, Bai X. Evaluating the effectiveness of urban energy conservation and GHG mitigation measures: the case of Xiamen city, China. Energy Policy 2010;38(9):5123–32.

[12] Brownsword RA, Fleming PD, Powell JC, Pearsall N. Sustainable cities–modelling urban energy supply and demand. Appl Energy 2005;82(2):167–80.

[13] Kontokosta C. Local Law 84 Energy Benchmarking Data: Report to the New York City Mayor's Office of Long-Term Planning and Sustainability; 2012.

[14] Heiple S, Sailor DJ. Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles. Energy Build 2008;40(8):1426–36.

[15] Touchie MF, Binkley C, Pressnail KD. Correlating energy consumption with multi-unit residential building characteristics in the city of Toronto. Energy Build 2013;66:648–56.

[16] Kavgic M, Mumovic D, Summerfield A, Stevanovic Z, Ecim-Djuric O. Uncertainty and modeling energy consumption: sensitivity analysis for a city-scale domestic energy model. Energy Build 2013;60:1–11.

[17] Keirstead J, Jennings M, Sivakumar A. A review of urban energy system models: approaches, challenges and opportunities. Renew Sustain Energy Rev 2012;16(6):3847–66.

[18] Howard B, Parshall L, Thompson J, Hammer S, Dickinson J, Modi V. Spatial distribution of urban building energy consumption by end use. Energy Build 2012;45:141–51.

[19] Ewing R, Rong F. The impact of urban form on US residential energy use. Housing Policy Debate 2008;19(1):1–30.

[20] Zhao HX, Magoulès F. A review on the prediction of building energy consumption. Renew Sustain Energy Rev 2012;16(6):3586–92.

[21] Marasco DE, Kontokosta CE. Applications of machine learning methods to identifying and predicting building retrofit opportunities. Energy Build 2016;128:431–41.

[22] Baker KJ, Rylatt RM. Improving the prediction of UK domestic energy-demand using annual consumption-data. Appl Energy 2008;85(6):475–82.

[23] Hsu D. How much information disclosure of building energy performance is necessary? Energy Policy 2014;64:263–72.

[24] Kontokosta CE, Jain RK. Modeling the determinants of large-scale building water use: Implications for data-driven urban sustainability policy. Sustain Cities Soc 2015;18:44–55.

[25] Tso GK, Yau KK. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy 2007;32(9):1761–8.

[26] Pérez-Lombard L, Ortiz J, González R, Maestre IR. A review of benchmarking, rating and labelling concepts within the framework of building energy certification schemes. Energy Build 2009;41(3):272–8.

[27] Vapnik V. The nature of statistical learning theory. Springer science & business media; 2013.

[28] Li Q, Meng Q, Cai J, Yoshino H, Mochida A. Applying support vector machine to predict hourly cooling load in the building. Appl Energy 2009;86(10):2249–56.

[29] Dong B, Cao C, Lee SE. Applying support vector machines to predict building energy consumption in tropical region. Energy Build 2005;37(5):545–53.

[30] Paudel S, Nguyen PH, Kling WL, Elmitri M, Lacarriere B, Corre OL. Support vector machine in prediction of building energy demand using pseudo dynamic approach. arXiv preprint arXiv:1507.05019; 2015.

[31] Biau G. Analysis of a random forests model. J Mach Learn Res 2012;13(Apr):1063–95.

[32] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[33] Breiman L. Bagging predictors. Mach Learn 1996;24(2):123–40.

[34] Liaw A, Wiener M. Classification and regression by randomForest. R news 2002;2(3):18–22.

[35] Prasad AM, Iverson LR, Liaw A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 2006;9(2):181–99.

[36] Jain RK, Smith KM, Culligan PJ, Taylor JE. Forecasting energy consumption of multi-family residential buildings using support vector regression: investigating the impact of temporal and spatial monitoring granularity on performance accuracy. Appl Energy 2014;123:168–78.

[37] Jain RK, Moura JM, Kontokosta CE. Big data+ big cities: Graph signals of urban air pollution [exploratory sp]. IEEE Signal Process Mag 2014;31(5):130–6.

[38] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12(Oct):2825–30.

[39] Tofallis C. A better measure of relative prediction accuracy for model selection and model estimation. J Oper Res Soc 2014;66(8):1352–62.

[40] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction, vol. 2. Berlin: Springer; 2009. Springer series in statistics.