Web-Based Visualization and Prediction of Urban Energy Use from Building Benchmarking Data

Constantine Kontokosta
ckontokosta@nyu.edu

Christopher Tull
christopher.tull@nyu.edu

David Marulli
dm2783@nyu.edu

Renate Pingerra
rp2427@nyu.edu

Maha Yaqub
maha.yaqub@nyu.edu

Center for Urban Science and Progress (CUSP)
New York University
1 MetroTech Center, 19th Floor
Brooklyn, NY 11201

ABSTRACT

New York City has pledged to reduce its greenhouse gas emissions by 80 percent by the year 2050, and 60 percent of these reductions will need to come from the buildings sector. Unfortunately, the wide and rapid adoption of energy conservation measures is hindered by the lack of granular, comprehensive, and easily accessible energy usage data for buildings. To increase the volume of available building energy data, New York City's Local Law 84 requires large buildings to disclose their energy consumption.

This paper details two ongoing projects to increase both the availability and comprehensiveness of building energy data. The first is a web-based visualization tool which allows users to understand patterns of energy consumption in individual buildings and across the city. The second project attempts to generalize from disclosure data by creating a predictive model of annual energy consumption for each building in the city. Building-level predictions are then validated against aggregate zip code-level data from local utilities.

1. INTRODUCTION

The effects of anthropogenic climate change are well documented, leading to increased global average temperatures, decreased snow and ice, rising sea levels, and more extreme weather conditions. These events threaten the safety and stability of societies around the world, and place an even greater burden on already disadvantaged communities. In order to mitigate the worst predicted changes, the Intergovernmental Panel on Climate Change (IPCC) recommends that global average temperature increase be limited to 2 °C above pre-industrial levels. However, achieving this goal will require a concerted global effort to reduce greenhouse gas (GHG) emissions by 41-72 percent, with even higher reductions necessary in the most developed nations [9].

To achieve these goals a number of organizations, from the European Union [8] to individual cities [1, 5], have pledged to reduce their production of greenhouse gases by 80 percent by the year 2050: a process popularly referred to as *80 by 50*. As part of this process, buildings have been identified as a key area where cost-effective reductions can be made [20]. In the United States, residential and commercial buildings together accounted for 34 percent of all GHG emissions nationwide in 2013 [7]. However, the energy used by buildings is an even more significant source of greenhouse gas emissions in dense urban areas such as New York City, where 70 percent of all emissions are due to the heating fuel, natural gas, electricity, and biofuel used in buildings [4]. The city has calculated that in order for New York to achieve its 80 by 50 plan, more than 60 percent of its GHG emission reductions must come from improvements in building efficiency [3].

Despite the urgency of increasing the energy efficiency of buildings, progress has been slow, due largely to social factors at the intersection of governance, engineering, and market structures [19]. Many of these impediments to the wide adoption of energy efficiency improvements derive from persistent information asymmetries in real-estate markets, including asymmetries between building owners, tenants, service providers, lenders, and public agencies [14]. The presence of large quantities of accurate, available and granular data about building energy consumption could help restore the information balance between these various actors.

1.1 Energy Disclosure Laws

One way to even out the flow of information is through information disclosure laws, which have been shown to effectively increase transparency in areas as diverse as finance and nutritional labeling [21]. As of March 2015, ten U.S. cities and one county had adopted energy disclosure laws requiring some buildings to release their energy consumption data to the local government [18]. In 2009, New York City became the first of these cities with the passage of Local Law 84 (LL84) as part of the Greener Greater Buildings Plan. LL84 requires all buildings over 50,000 square feet to annually disclose their energy and water consumption along with a number of usage and occupancy characteristics [2]. This data set is the largest of its kind, with approximately 13,000 properties reporting their energy use during the year 2013.

By requiring and enforcing regular reporting of energy use, disclosure or benchmarking laws allow the energy consump-

Bloomberg Data for Good Exchange Conference.
28-Sep-2015, New York City, NY, USA.

tion of a building to be benchmarked against its own performance to track changes over time. These data also enable a building to be benchmarked against other peer buildings that share similar land use, construction, and location characteristics. For example, repeated observations of the energy consumption of a large number of buildings over time could allow for robust longitudinal studies to determine the effects of energy retrofit measures. A large sample size in these studies would be required to get accurate estimates and to determine how conservation measures interact with other variables such as property type and building physical characteristics. More accurate estimates of retrofit effects would lower uncertainty around retrofit financing, and allow more buildings to take advantage of the cost-reductions associated with increased energy efficiency.

New York's Local Law 84 has the capability to enable more informed decisions in the energy efficiency market, but LL84 and energy disclosure laws in general have several major disadvantages. First, information disclosure laws are most effective when the data they generate is embedded in the daily decision-making process of relevant actors [21]. However, the energy disclosure data from LL84 are currently released to the public in comma-separated value (CSV) format, either through the NYC Open Data portal, or through an obscure web page maintained by the Mayor's Office. CSV format may be ideal for analysts but it can provide a barrier to access for many building stakeholders. Second, while New York's LL84 data set is the largest of its kind, it contains data for only 1.5 percent of properties in NYC (as calculated by unique tax lots). Thus there are a huge number of properties for which no energy data are available. This leads to many outlying areas of the city having almost no representation in the data set. Lastly, because the data are self-reported by building managers, there is the possibility for errors due to incorrectly typed values or misunderstanding of reporting rules. As an example, errors have been identified when two buildings on separate parcels share the same meter, or when different buildings share a parcel [15].

This paper describes works in progress to address all three of these problems. The poor availability of building energy data is being addressed through the development of a public-facing web-based visualization tool. The tool will allow building stakeholders and the general public to explore energy use across the city, as well as to gain insight into the consumption of specific buildings. The lack of energy consumption data for the entire city is addressed through the construction of a predictive model of energy consumption. The model is fit to benchmarking data, and then validated against aggregate energy data from utilities. Finally, the issue of data quality is addressed through a general statistical data-cleaning methodology that applies broadly to building energy data.

2. DATA AND CLEANING

There were three primary data sources utilized for this work. LL84 and PLUTO are used in the visualization and energy modeling, while the zip code energy data is used only for validation of the predictive model.

2.1 Local Law 84

These are the energy benchmarking data for New York City. The public version of the data set available on the NYC Open Data portal was used for the visualization tool. These data contain annual energy use, water use, and GHG emissions for all reporting properties as well as the primary property type (Office, Multifamily Housing, etc.), and the Borough, Block, and Lot (BBL) number used to match the reporting properties with city tax lot data. The prediction task used a confidential version of the data provided by the NYC Mayor's Office of Sustainability. This data set contains additional fields for annual consumption of each fuel type such as electricity, natural gas, steam, and various fuel oils.

2.1.1 Data-Cleaning

The first cleaning steps were to remove redundant properties. This took the form of removing entries with more than one BBL number specified, as well as duplicate entries. The second cleaning step was to remove entries with no reported energy use. This was done by removing entries with zero or null values for weather-normalized source energy use intensity[1] (EUI).

Of the reported fields, it was decided that EUI was an optimal field to use for identifying entries with other incorrectly entered information. This field is calculated as a composite of all individual fuel types and the reported gross floor area. As such, if any of the sub-fields is drastically misreported[2], it should produce extreme EUI values. The EUI values for the whole data set were observed to roughly follow a logarithmic normal distribution. EUI values were then transformed with a natural logarithm producing a normal distribution of the data, and extreme values were removed if they were greater than ± 2 standard deviations from the mean. This corresponds to removing approximately the top and bottom 5 percent of entries[3]. Figure 1 details this process.

2.2 PLUTO

The NYC Primary Land Use Tax Lot Output (PLUTO) is an extensive public data set provided by the NYC Department of City Planning. It contains location, land-use, and physical characteristics for all the buildings in NYC, and can be used to identify properties based on address, BBL, and zip codes. In this work, the data set is used to provide a standardized set of building characteristics. These characteristics are matched with LL84 for training the predictive model, and also used to extrapolate energy use across the city. The corresponding MapPLUTO data contain location and shape information that is used for visualization. The data are updated regularly, and version 14v2 was used in this analysis.

2.2.1 Data-Cleaning

PLUTO contains fields representing how much floor area in a building is devoted to residential, office, retail, and other

[1]Defined as energy consumed per square foot (kBTU/ft^2), and adjusted for annual weather trends.
[2]E.g. through accidental addition or omission of zeros.
[3]Upon publication the data used for visualization still used an older methodology to remove the top and bottom 1% of entries.

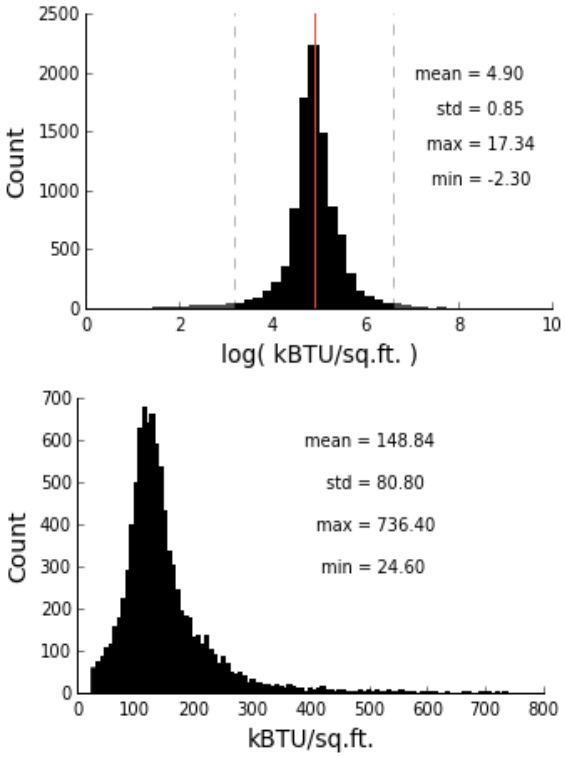

Figure 1: Top: Histogram of log EUI with dashed lines at ± 2 standard deviations. Bottom: Histogram of EUI after removing outliers.

uses. In approximately 8 percent of cases, the sum of these subdivisions did not equal the field denoting total building area. In order to avoid ambiguities, a new total floor area field was calculated as the sum of all floor area subtypes. This derived field was used in all subsequent calculations requiring a total built floor area.

2.3 Zip Code Energy Consumption

These data contain aggregate energy consumption for all of New York City at the zip code level, broken down by fuel type into electrical, natural gas, steam, and fuel oil. The data were obtained from the local utilities Consolidated Edison, National Grid and the Long Island Power Authority, and were provided for research by the NYC Mayor's Office of Sustainability. These data will be used as a ground truth when extrapolating energy consumption from large buildings to smaller buildings that do not report their energy usage under LL84.

2.3.1 Data-Cleaning

Due to unresolved definitions and data quality issues, only electricity and natural gas have been used in the current analysis. With these, the only cleaning performed was to convert all values to kBTU to allow comparison.

3. BUILDING ENERGY VISUALIZATION

In order to obtain the full value from energy disclosure laws, it is necessary for the data to be easily accessible and fully embedded into the decision-making processes of stakeholders. This paper describes work to develop an interactive website that enables users to easily navigate and query the data from LL84.

Much of the scholarly work on visualizing building energy data is focused on data for individual buildings. This includes tools designed for building managers and designers to visualize load profiles and better understand energy usage in real time [16, 17]. Other work discusses methods designed for end use consumers to promote energy-consciousness and reduce consumption by visualizing realtime energy use [11]. Software tools have also been proposed to monitor and manage energy use on a city-wide level by visualizing geolocated energy data [13], but once again this work was designed for real-time energy use.

The most relevant example of a comparable project is a web-based visualization created for the city of Philadelphia to display their 2014 energy benchmarking data [6]. This site informed the design of this groups work, particularly with regards to the visual encoding chosen for the interactive map. However, the Philadelphia site has a more passive nature overall that guides users through particular facts and figures. The design of the NYC site takes a more interactive approach with higher levels of user control.

3.1 Interaction Design

Figure 2 captures much of the core functionality of the visualization tool. At top left, users are able to switch between different metrics of interest, currently EUI, water use intensity (WUI), GHG emissions, and gross floor area. This is the highest level of selection and changes update all other portions of the visualization. Just below this is a search bar that allows users to query for particular buildings and addresses. This enables building stakeholders to seek out information about the energy consumption of their building, and to see how it compares to peer buildings.

Further down is a list of primary property types next to the number of properties of that type represented in the data. This is accompanied by a whisker plot that displays means and standard deviations for different property types, thereby highlighting both the difference in consumption and variance between different uses. This portion of the tool also allows users to select and focus on a single property type. This updates the map do show the spatial distribution of buildings of that type. Figure 2 shows the results of selecting only office buildings; the heavy concentration of properties in midtown Manhattan is clearly visible, as is the darker color (higher EUI) of midtown buildings.

Selecting a property type also displays a scatter plot with points representing buildings of that type displayed along two dimensions. This display allows for easy identification and selection of extreme outliers, thus allowing the user to investigate further. For now the dimensions are fixed with GHG emissions on the x-axis and the currently selected metric on the y-axis. Ultimately the user will be able to specify which relationship to display by selecting each of the dimensions.

The interactive map serves several purposes. Most obviously

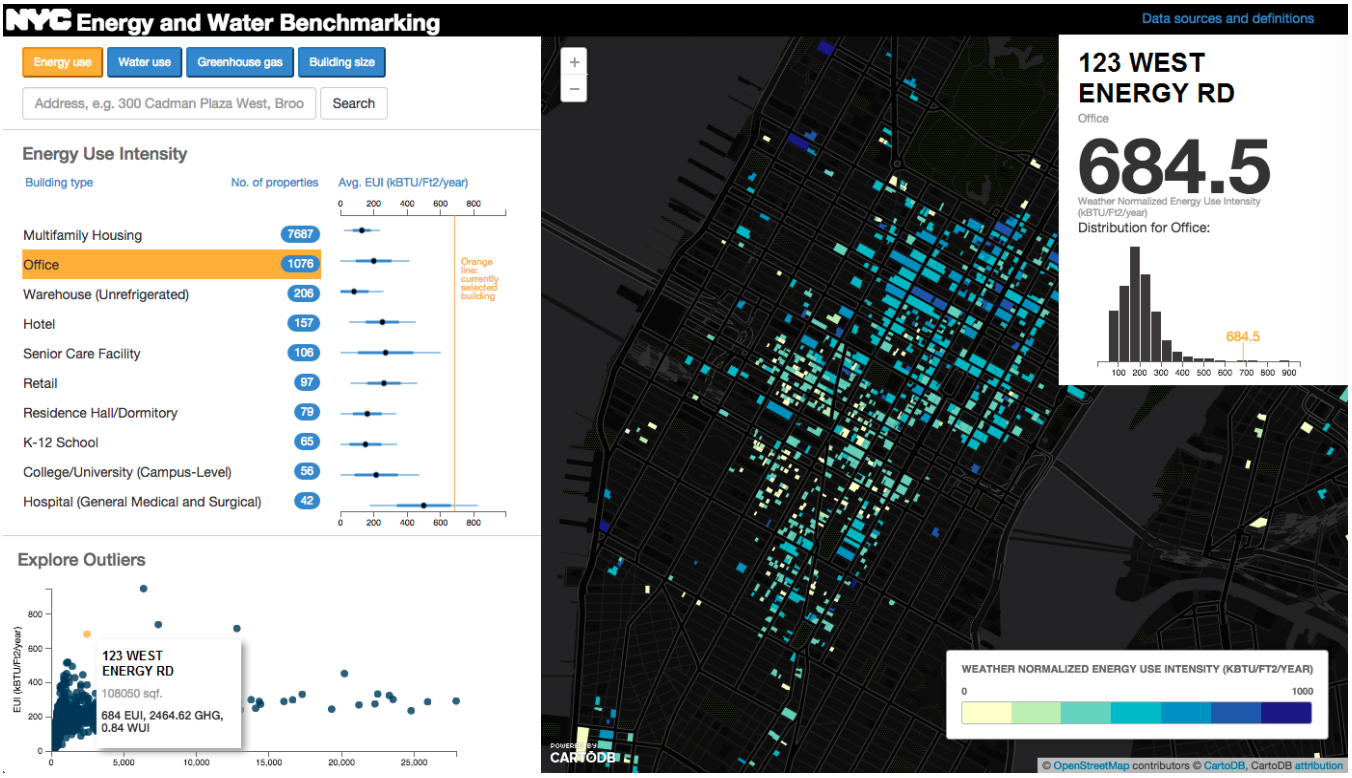

Figure 2: An overview of the building energy visualization site. The left hand side displays summary statistics and graphics for different property types. The right hand side displays an interactive map of the buildings in LL84 along with an info-box containing details about a specific property of interest.

it displays the spatial distribution of buildings which report under Local Law 84 and also displays the spatial distribution of energy use in the city. It also serves a s a navigation and query tool by allowing users to pan and zoom to locations of their choosing in order to examine the details of particular neighborhoods. Users can then select buildings which they recognize or which stand out as particularly high or low energy users.

The selection of an individual property on the map or scatter plot, or a successful search with the search bar will create the small information box visible in the upper right corner of Figure 2. This box displays a number of summary statistics about the selected property, and also displays a histogram of the metric of interest. This histogram is limited to properties of the same type as the selection, thus allowing more accurate comparisons to be made. This comparison is further aided by the appearance of an orange line on the histogram, marking the position of the selected property relative to the overall distribution within that property type. A similar line appears on the whisker plot, marking the location of that building relative to the means for different property types.

3.2 Implementation

The visualization tool is implemented using standard web technologies: HTML, CSS, and Javascript. The charts and graphs are created using the Data-Driven Documents (D3) library for Javascript, which allows easy binding of data points to SVG objects. The interactive map is implemented using the CartoDB.js library. Using CartoDB allows easy application of powerful mapping software, and also provides a spatially-enabled back-end for storing and easy querying of property polygons and energy data.

4. PREDICTING URBAN ENERGY USE

The second goal of this work is to identify how well energy benchmarking data can be extrapolated to buildings which aren't required to report their energy use. This is done by creating a predictive model of energy consumption based on the LL84 data and then using this model to predict energy use for every property in the city.

There are a variety of computational approaches to predicting building energy consumption. At the highest level these can be divided into engineering methods and statistical methods [23]. Engineering methods operate by forming elaborate physical models of building thermal properties and sub-component operation. They are useful in the building planning stage, but are often impractical for estimating the consumption of buildings already in operation because they require many detailed parameters that are often difficult or impossible to obtain. Statistical methods, on the other hand, operate by correlating energy consumption with influencing variables. This is done by collecting historical data for both the energy use and predictor variables and then using some sort of regression model to estimate the relationship between predictors and consumption. These

data-driven approaches are also known as inverse models and they can provide reliable estimation while being substantially simpler and less time-consuming than engineering methods [22].

There are several previous studies which directly relate to the approach used in this research project. Howard et al. [12] utilized zip code level energy consumption data from 2009 comparable to the zip code level energy data used in this project from 2013. Howard et al. utilize a robust linear regression model with building area devoted to various land uses as input and zip code level consumption data as output. This results in estimated EUI values for each land-use type across the city. These static EUI coefficients are useful, but they ignore a great deal of building-level variation that results from factors like building physical characteristics. A second related study utilized a robust linear regression model fit to LL84 benchmarking data to show that a number of building characteristics including occupancy and physical construction were significantly correlated with energy use [15].

4.1 Methodology

Work so far has been limited to predicting only electricity and natural gas consumption for New York City. Future work will extend this to consider total consumption

4.1.1 Electricity Prediction

The electricity use intensity of each property in LL84 was calculated as the annual total electricity consumption in kBTU divided by the total building floor area in square feet. Rows with zero building area or which lacked an electricity consumption were dropped from the data set, resulting in 8492 properties remaining. The electricity use intensity field was transformed using the natural logarithm to produce a more nearly gaussian distribution of values, and this was then used for prediction.

Although the data were previously filtered to remove properties in the extremes of total EUI, there was no specific filtering to remove outliers in electricity use. To compensate for these outliers, a robust linear regression model was used to fit the 8492 training examples. The specifics of the model are visible in Table 3 of the Appendix.

Robust regression has previously been shown in the literature to be a useful technique for estimating energy use as well as gaining insight into the factors relevant to increased consumption [12, 15]. The linearity assumptions inherent in linear regression tend to create a less "flexible" model with higher bias and lower variance [10]. In the case of this study, low variance is a benefit due to limitations of the training data. Specifically, the training data are drawn exclusively from large properties in NYC while a majority of properties in the city (54 percent) are classified as 1 or 2 family homes. Given the lack of representation of small properties in the training data, it is desirable to capture the most important and generalizable features of the data set without capturing those features which apply only to large properties. A rigid and inflexible model like linear regression is well-suited to this task. Table 1 in the Appendix shows the model used to predict electricity use intensity with a brief description of building characteristics used.

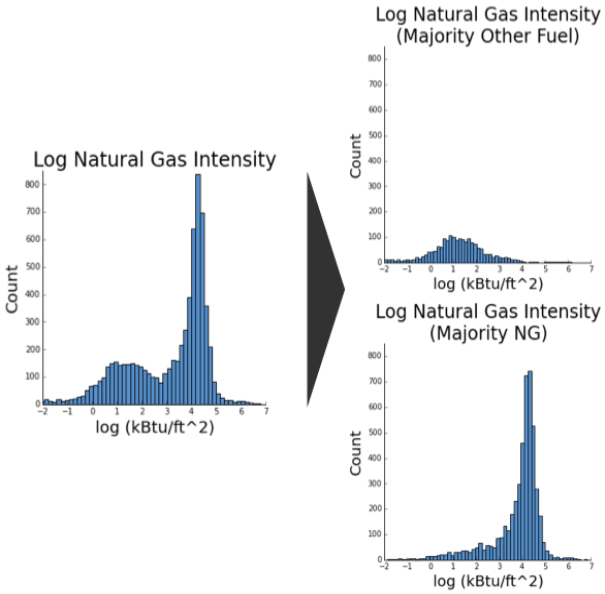

Figure 3: The bimodal distribution for NGUI was effectively split into two unimodal distributions by separating out buildings which use primarily natural gas as their non-electricity energy source.

After fitting the model to the LL84 training data, the model is then used to predict an electricity use intensity for every property present in the PLUTO data set. The use intensity is then multiplied by building floor area and aggregated at the level of the zip code for validation against the ground-truth utility data as discussed in the results.

4.1.2 Natural Gas Prediction

The prediction of natural gas use intensity (NGUI) proceeds in much the same way as that for electricity, with one major difference. The NGUI was first calculated by dividing the total annual natural gas use in kBTU by the floor area in square feet. Rows with zero building area or which lacked natural gas consumption were dropped from the data set, resulting in 6967 properties remaining.

The use intensity was again log-transformed, at which time a clearly bimodal distribution of NGUI became visible (Figure 3). It is hypothesized that this distribution is actually a compound of two different distributions arising from separate processes. The first process being buildings which primarily use an alternate fuel, which only a minority of natural gas use. This could correspond to buildings which use steam or fuel oil for heating, and natural gas only for secondary purposes like cooking. The second process would be buildings which use natural gas as their primary heating fuel as well as for secondary purposes. This hypothesis was supported after observing that separation of the buildings into two groups (majority natural gas and minority natural gas) produced two distinctly separate unimodal distributions for NGUI. Furthermore, this was the only distinction observed to nicely separate the distributions. Splitting the properties based on property type or on year built failed to produce unimodal distributions for the subgroups. This splitting process is visible in Figure 3.

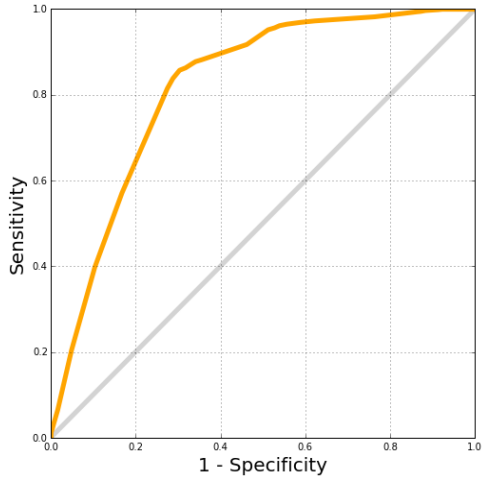

Figure 4: The ROC curve for the logit model to predict natural gas class. The area under the curve (AUC) is 0.818.

After determining that NGUI varies greatly based on which group a building falls into, this project set out to predict which buildings in NYC would use natural gas as their primary fuel. This was done by first creating a binary variable in Local Law 84 corresponding to natural gas class, equal to 0 if a building uses majority natural gas, and 1 for majority other fuel. The prediction was then done by training a logistic regression binary classifier on the LL84 data, and then using it to predict the gas class for every building in PLUTO. Figure 4 shows the Reciever Operating Characteristic (ROC) curve for the logit model evaluated on the LL84 data. This curve details how the model performs with regards to sensitivity (true positive rate) and specificity (true negative rate) as the cutoff threshold is varied. A curve near the diagonal performs roughly as good as random, while a curve tracing significantly above the diagonal performs much better than random. The curve was calculated using leave-one-out cross validation, whereby for each data point p, a model is fit to all of the data except for p, and then used to predict the class of point p.

The area under the ROC curve (AUC) is often used as a summary statistic for the performance of binary classifiers. The logit model used here has an AUC of 0.818 on the LL84 data. This is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Among the training data, a threshold of 0.3 appears to perform the best overall, but when generalizing to the entire city, a much higher threshold of 0.7 performed the best. This corresponds to only classifying buildings as majority other fuel if the classifier is very certain. Figure 4 shows the ROC curve with AUC.

The logit prediction of which class a building belongs to was then added as a feature to the PLUTO data set. This feature then became a predictor into a robust linear regression model as before. All of the previous arguments in favor of a linear regression model also apply here. It is worth noting that the optimal model in this case was simpler than in the electricity case, likely owing to the innate difficulty in predicting natural gas use, discussed later. The details for both the logistic regression classifier and the linear regression predictor are not included in this paper, but the regressors used are similar to those in Table 1 in the Appendix.

4.2 Prediction Results

After predicting the electricity and natural gas use intensity for every property in NYC, the results were multiplied by building total square footage and then aggregated at the zip code level. After aggregation, the predicted consumption was compared to the measured consumption obtained from the utilities. The results are visible in Figure 5, where measured consumption for each zip code is plotted on the x-axis, and predicted consumption is plotted on the y-axis. A perfect prediction would lie perfectly on the line y = x, while the dotted lines correspond ±20% of the correct value. The median absolute percent error (median APE) for electricity was 10.75 meaning that half of the predicted zip codes were within 10.75% of correct, and the R^2 was 0.93, meaning that 93% of the variance between zip codes was explained by the model. Likewise, the median APE of natural gas was 30 meaning that half of the zip codes were within 30% of correct, and the R^2 was 0.65, meaning that 65% of the variance between zip codes was explained by the model.

The large discrepancy between electricity and natural gas owes to the inherent difficulty in predicting natural gas use compared to electricity. Virtually every building can be expected to use electricity for lighting, computers, microwaves, and other purposes. On the other hand, not every building can be expected to use substantial quantities of natural gas. First, there is inherent variability in whether a building even uses natural gas. For example, a building may rely entirely on fuel oil for heating, and then use electricity for cooking. In this model, all buildings were assumed to use some level of natural gas, even if the amounts were very low. Second, there is uncertainty in the natural gas classification of buildings. Since only the most certain predictions were classified into the majority natural gas group, it is possible that a substantial percentage of the properties in NYC were incorrectly classified. Lastly, there is uncertainty as to what portion of a building's fuel use comes from natural gas. In the model used here, any building with less than 50% natural gas as their fuel would be classified as the zero group. This leaves a large variance that is likely unexplained by the relatively simple linear model used here.

5. CONCLUSIONS AND IMPACT

Energy disclosure laws have been proposed as an effective way to reduce information asymmetries among actors in the building sector. By doing so it is hoped that they will catalyze the transformation of the current building stock and enable rapid investments in energy efficiency. The work in this paper has discussed two attempts at overcoming the limitations of these disclosure laws. First, the design of an interactive visualization tool was discussed. This tool can unlock benchmarking data from the static form of CSV files to engage building stakeholders and increase data accessibility. Second, a method was discussed to generalize from benchmarking data to gain further insight into the spatial distribution of energy use in New York City. This method has the potential to inform the targeting of localized energy conservation measures. A complete energy consump-

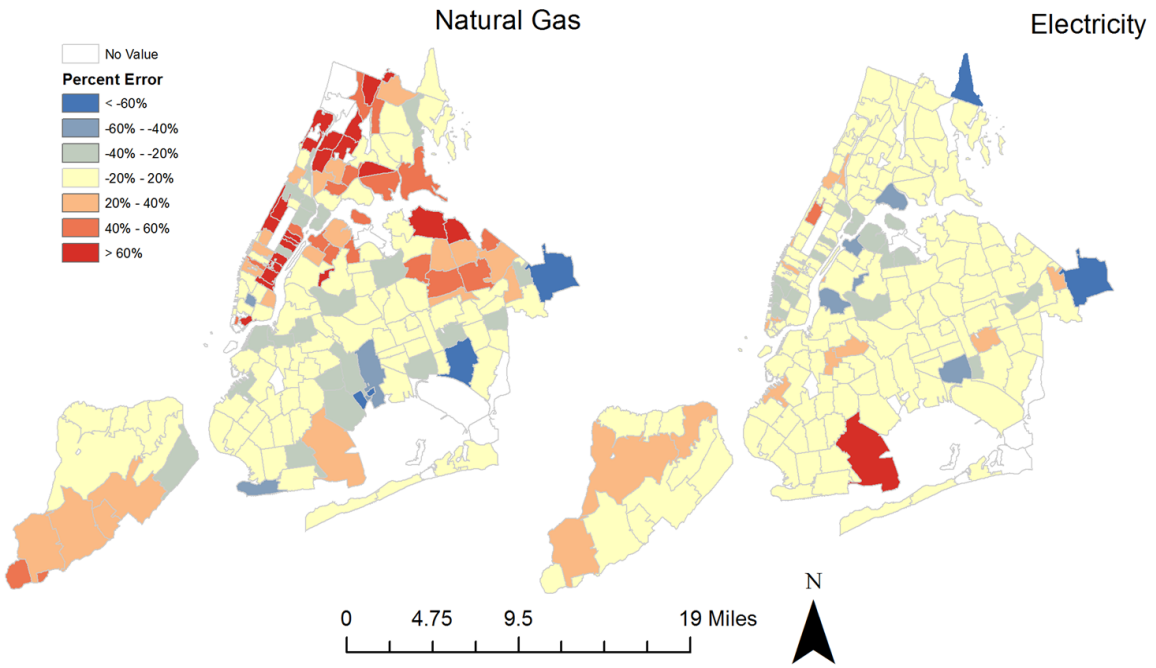

Figure 6: Percent error for each zip code in NYC. Red areas use more energy than predicted while blue areas use less.

tion map can also serve as the basis of a more general energy planning tool.

Future work will extend both of the projects and possibly combine them to provide interactive examination of actual and estimated energy consumption across NYC.

6. ACKNOWLEDGMENTS

The authors would like to thank the New York City Mayor's Office of Sustainability for supplying the data used in these projects.

7. REFERENCES

[1] City of Los Angeles. *The Sustainable City Plan*, 2015.

[2] City of New York. *Local Laws of The City of New York for the Year 2009 - No. 84*, 2009.

[3] City of New York. *New York City's Pathway to Deep Carbon Reductions*, 2013.

[4] City of New York. *Inventory of New York City Greenhouse Gas Emissions*, 2014.

[5] City of New York. *One City Built to Last*, 2014.

[6] City of Philadelphia. 2014 building energy benchmarking, 2014.

[7] EPA (US Environmental Protection Agency). *Inventory of U.S. Greenhouse Gas Emissions and Sinks 1990 to 2013*, 2014.

[8] European Commission. *A Roadmap for Moving to a Competitive Low Carbon Economy in 2050*, 2011.

[9] C. Field, V. Barros, D. Dokken, K. Mach, M. Mastrandrea, T. Bilir, M. Chatterjee, K. Ebi, Y. Estrada, R. Genova, B. Girma, E. Kissel, A. Levy, S. MacCracken, P. Mastrandrea, and L. White. *Summary for Policymakers. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 2014.

[10] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning, 2009.

[11] T. G. Holmes. Eco-visualization: combining art and technology to reduce energy consumption. In *Proceedings of the 6th ACM SIGCHI conference on Creativity & cognition*, pages 153–162. ACM, 2007.

[12] B. Howard, L. Parshall, J. Thompson, S. Hammer, J. Dickinson, and V. Modi. Spatial distribution of urban building energy consumption by end use. *Energy and Buildings*, 45:141–151, 2012.

[13] S. A. Kim, D. Shin, Y. Choe, T. Seibert, and S. P. Walz. Integrated energy monitoring and visualization system for smart green city development: Designing a spatial information integrated energy monitoring model in the context of massive data management on a web based platform. *Automation in Construction*, 22:51–59, 2012.

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

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

[16] D. Lehrer and J. Vasudev. Visualizing information to improve building performance: A study of expert users. *Center for the Built Environment*, 2010.

[17] N. Motegi and M. A. Piette. Web-based energy information systems for large commercial buildings.

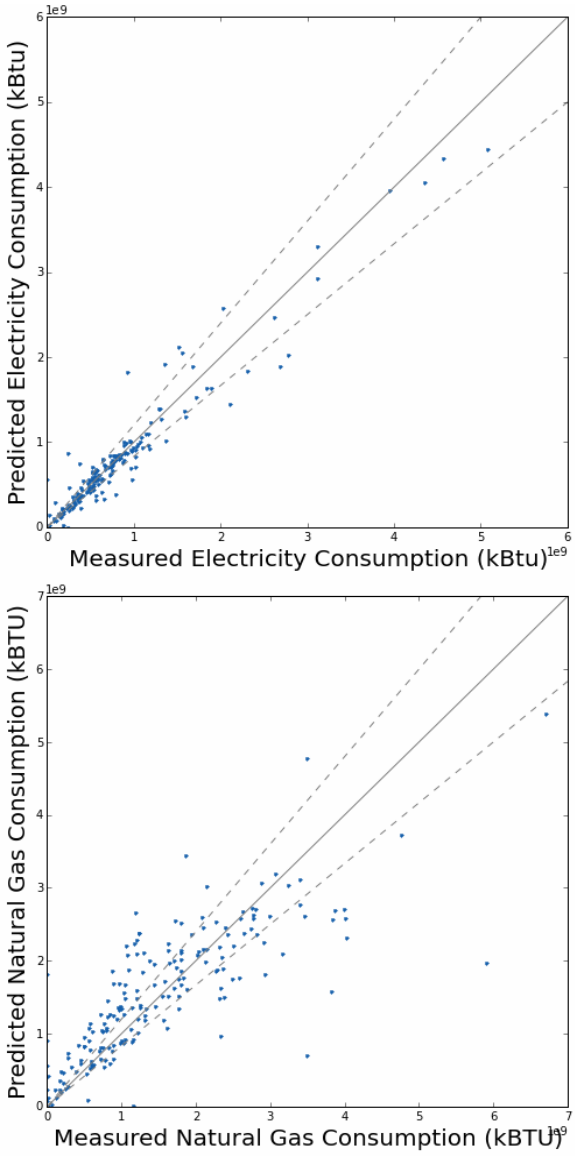

Figure 5: Measured versus predicted values for electricity and natural gas respectively for each zip code in NYC.

Lawrence Berkeley National Laboratory, 2003.

[18] K. L. Palmer and M. Walls. Can benchmarking and disclosure laws provide incentives for energy efficiency improvements in buildings? *Resources for the Future Discussion Paper*, 2015.

[19] M. Ryghaug and K. H. Sørensen. How energy efficiency fails in the building industry. *Energy Policy*, 37(3):984–991, 2009.

[20] D. Ürge-Vorsatz and A. Novikova. Potentials and costs of carbon dioxide mitigation in the worldś buildings. *Energy policy*, 36(2):642–661, 2008.

[21] D. Weil, A. Fung, M. Graham, and E. Fagotto. The effectiveness of regulatory disclosure policies. *Journal of Policy Analysis and Management*, 25(1):155–181, 2006.

[22] Y. Zhang, Z. O'Neill, B. Dong, and G. Augenbroe. Comparisons of inverse modeling approaches for predicting building energy performance. *Building and Environment*, 86:177–190, 2015.

[23] H.-x. Zhao and F. Magoulès. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16(6):3586–3592, 2012.

APPENDIX

A. REGRESSION RESULTS

Table 1 shows the robust linear regression model used to predict electricity use intensity. Both of the models related to natural gas use similar regressors, although the models are less complex.

The first variables in the model after the intercept are binary variables indicating the time period in which the building was built. The variables PercentRes, PercentOffice, etc. are proportions of the floor space in the building devoted to the specified use. These are then interacted with specific land use variables indicating how the building is zoned by the Department of City Planning. This allows floor space in each land use to have a unique coefficient. The number of floors and floor-area ratio come directly from PLUTO fields, while Attached Lot and Inside Lot are derived binary variables. Manhattan is a binary variable indicating whether a building is located in Manhattan. Finally, SVR is the surface area-to-volume ratio of the building derived from fields indicating lot width and depth, and assuming a rectangular building.

Table 1: Model Specification for Electricity Use Intensity

Regressor	Coefficient	Std. error	p-Value
Intercept	3.4865	0.079	0.000
Built 1931 to 1950	-0.0624	0.014	0.000
Built 1951 to 1970	0.0708	0.013	0.000
Built 1971 to 1990	0.2797	0.018	0.000
Built after 1991	0.3637	0.016	0.000
PercentRes:MultiFamWalkUp	-0.8945	0.081	0.000
PercentRes:MultiFamElev	-0.8438	0.079	0.000
PercentRes:MixedResOffice	-0.8039	0.080	0.000
PercentRes:MFOther	-1.3325	0.126	0.000
PercentOffice:MultiFamElev	0.4534	0.307	0.140
PercentOffice:MixedResOff	0.6402	0.166	0.000
PercentOffice:Office	0.1399	0.080	0.080
PercentOffice:Industrial	0.0131	0.133	0.922
PercentOffice:Public	0.4182	0.102	0.000
PercentOffice:OfficeOther	3.1747	1.398	0.023
PercentRetail:MultiFamElev	1.6883	0.208	0.000
PercentRetail:MixedResOff	1.2806	0.115	0.000
PercentRetail:Office	0.5756	0.086	0.000
PercentRetail:Industrial	0.9422	0.269	0.000
PercentRetail:RetailOther	0.6761	0.200	0.001
PercentGarage	-0.9947	0.104	0.000
PercentStrge	-0.7166	0.086	0.000
PercentFactry	-0.7973	0.090	0.000
PercentOther	-0.1324	0.081	0.100
Number of Floors	0.0055	0.001	0.000
Floor-Area Ratio	-5.122e-05	0.000	0.799
standardize(SVR)	-0.0227	0.005	0.000
Attached Lot	-0.0177	0.015	0.229
Inside Lot	-0.0220	0.010	0.021
Manhattan	0.1875	0.013	0.000