A framework for urban building energy use modelling

Narjes Abbasabadi[1], Rahman Azari[1]

[1]Illinois Institute of Technology, Chicago, IL

ABSTRACT: Reliable quantification of energy consumption by buildings plays a key role in development of sustainable cities. However, there are methodological uncertainties embedded in the most common urban scale energy use modeling methods and tools which affect the reliability of these tools and their applicability for decision-making purposes. This article presents a novel bottom-up data-driven framework for urban energy use modeling (UEUM) to help predict energy use more precisely through utilizing disaggregated data at building level, incorporating the actual urban spatial patterns, and testing different algorithms to propose an enhanced prediction model. This framework integrates the influential factors in the model including building characteristics; i.e., height, as an urban intensity metric, urban attributes; i.e., sprawl indices, that are captured in a multidimensional way representing compactness and connectivity of neighborhoods, and occupant characteristics. A case study on 800,000 buildings in seventy-seven neighborhoods in Chicago was used to test the framework. This framework has the potential to help better understand the existing urban energy use profiles and provides a more holistic image of urban energy use at multi-scales of building, block, neighborhood, and urban levels.

KEYWORDS: Urban energy modeling; data-driven model; building operational energy use

INTRODUCTION

Buildings are the most significant contributors to urban energy use and associated emissions. In the United States, buildings operation accounts for 41% of primary energy use, and 38% of GHG emissions (EIA 2012). Achieving energy and emission reduction goals requires understanding the existing profiles of building energy use at urban scale. While urban energy use modeling and prediction are essential in urban energy management and understanding of energy performance of cities, there are a limited number of methods and tools to accurately model urban energy use in cities (Sola et al. 2018). Also, there is a lack of an integrated approach to incorporate actual urban spatial patterns with urban energy use models. The extant literature on energy performance of urban spatial patterns tends to examine energy use either at the scales of individual buildings or collection of buildings of limited typologies (Ruby and 2014, n.d.; Resch et al. 2016) that do not incorporate actual urban context effects; and at city-scale level studies (Howard et al. 2012) rely on aggregated data that do not allow energy characterization at individual building level.

The main approaches for urban building energy use modeling are classified into two main groups: *top-down* and *bottom-up*, according to their specific input, output and applied method (Swan and Ugursal 2009; Reinhart and Cerezo Davila 2016). The Top-down approach relies on aggregate energy data and does not consider disaggregated and individual characteristics of building system. Hence this model is less reliable for building level energy analysis. The bottom-up approach, identified as the dominant model, is founded upon two methos of engineering or simulation and data-driven or statistical techniques. Either simulation or data-driven techniques have their own limitations about urban energy quantification. The simulation methods often suffer from oversimplification of building and system data for the city-scale energy use estimations. They rely on a limited number of typologies or archetypes that represent buildings in a city to achieve time and computational efficiency. The interaction between the individual buildings and the city has been shown to impact the accuracy of operational energy use estimations at both building and urban scales (Zhou, Huang, and Cadenasso 2011; Reid Ewing 2010); however, it is often times overlooked. The estimation

methods are also founded upon arguable assumptions, particularly in denser and taller urban areas where urban microclimate has a noticeable impact on the building operational energy consumption (Martin et al. 2017). So the methodological uncertainties embedded in simulation methods affect the reliability of results and their applicability for decision-making purposes.

The data-driven models could present more accurate urban energy modeling if data be available and If sufficient variables are captured in the model (C. Kontokosta, Bonczak, and Duer-balkind 2016). Hence, their accuracy and reliability of results in these studies, however, depend on availability and quality of large sets of data and representative variables (C. Kontokosta, Bonczak, and Duer-balkind 2016; Hsu 2015). However, the previous data-driven studies rely on the generalized empirical data provided by building energy surveys; yet, limited number of surveys provide local energy data at building level. Previous studies (e.g. Howard et al. 2012), when conducted at city scale or a zip code level, use aggregated data and do not allow for energy characterizations at an individual building-level. Recently, as a part of disclosure law which adopted by many cities in the US, energy benchmarking was released. Energy benchmarking provides more transparency and provides disaggregated building-level energy data. However, it has limitations regarding availability for all buildings in the city. In case of Chicago, it covers less than 1% of Chicago's buildings, which account for approximately 20% of total energy used by all buildings ("Chicago Energy Benchmarking, 2016, City of Chicago, Data Portal," n.d.). In addition, previous data-driven urban energy models apply mostly traditional statistical techniques such as Multi Linear Regression (MLR) for urban energy prediction and explaining the association between influential factors such as urban spatial patterns and building characteristics and energy use because of its simple design and interpretability [12,78–80]. However, MLR method dese not allow capturing non-linear and complex patterns.

The article presents an urban energy use modeling (UEUM) framework which employs a bottom-up data-driven approach through using disaggregated data, incorporating the localized variables in the model and applying Machine Learning (ML) based algorithms. Machine learning based algorithms allow capturing non-linear and complex features and provide higher precision level [70–72]. This model helps predict urban energy use more precisely and comprehensively through utilizing disaggregated data at building level, incorporating the localized variables in the model, and testing different machine learning techniques and algorithms. This model proposes an enhanced prediction model and provides a multi-scale analysis and visualization at neighborhood, census tract, census block, and building levels. This research also has the potential to provide insights on urban energy use dynamics across morphological patterns and helps planners and policy-makers develop more energy efficient cities. Chicago has been selected as a pilot case study to test the applicability of this framework for urban energy use modeling.

1.0. METHODOLOGY

This research develops a data-driven framework for urban energy use modeling. The conceptual framework of this research is presented in Figure 1. The framework is built upon a three-step model concept. First, the *Pattern Extraction* phase which studies urban spatial patterns to extract new features and incorporates localized variables in the model, and second, *Prediction* phase is applied to estimate urban energy use through learning the mathematical relationship between variables and tests different machine learning techniques and algorithms to propose an enhanced prediction model, and finally, the third step provides a multi-scale analysis at neighborhood, census tract, census block, and building levels.

In this research, the urban energy use is outlined as building operational energy use intensity (EUI) at a city scale. The Site EUI (kBtu/sq ft) per year was used as the dependent variable. The model was run based on Log Site EUI (kBtu/sq ft) per year to properly fit the nonlinear relationships between variables. The influential factors which affect the urban building energy consumption were identified as three main groups including *Building Characteristics (*variables such as building type, building height, building size, and construction year), *Urban Attributes (*functioning as density, accessibility, connectivity and land-use mixed which are captured via

urban sprawl index), and *Occupancy Characteristics (*including total population, household size in residential buildings, worker density in commercial buildings, and percentage of occupied units).

Table 1. Key variables incorporated in the model.

	Category	Variable	Unit
Independent	Building Characteristics	Building Height	Number of floors
		Building Size	Square meter
		Building Type	-
		Built Year	-
	Occupant Characteristics	Total number of occupants	-
		Household size	-
		Worker density	-
		Weekly working hours	hour
		Percentage of occupied units	-
	Urban Attributes	Sprawl Index (density, accessibility, connectivity and land-use mixed)	Unitless
Dependent	Building Operational energy use	Site EUI	kBtu/sq. ft./year (kWh/m2/yr)

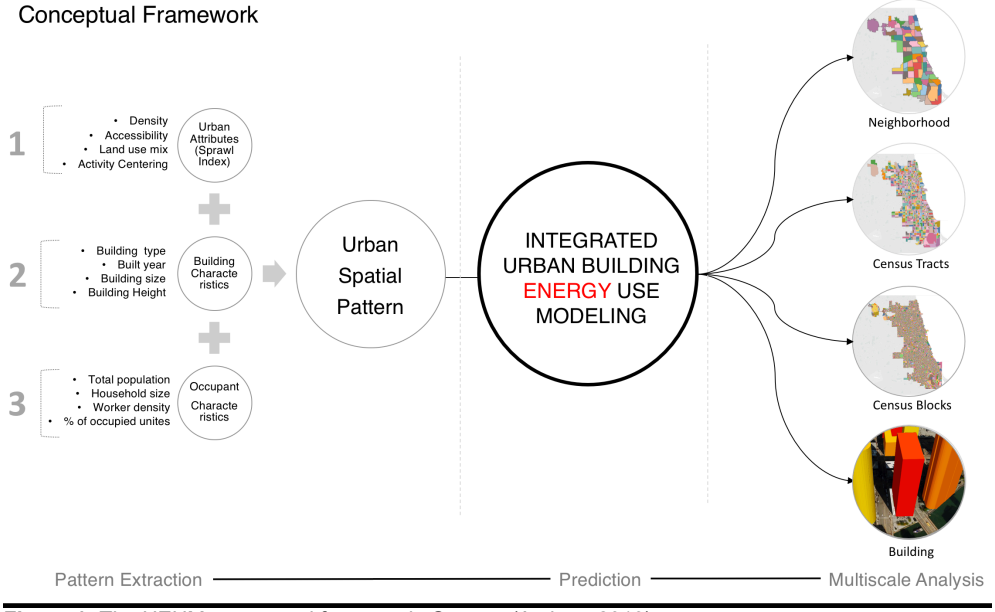

Figure 1: The UEUM conceptual framework. Source: (Authors 2019)

To test the framework, a case study on 800,000 buildings in seventy-seven neighborhoods in Chicago was selected. The merged urban spatial and energy dataset was built upon utilizing several datasets including GIS data representing the explicit geographical location and building characteristics such as Chicago building footprints (CBF) dataset (City of Chicago n.d.); the sprawl index representing the connectivity, compactness, land use and accessibility features of neighborhood as an indicator of urban attributes in this research was built upon the U.S. Urban Sprawl Data (National Cancer Institute n.d.) developed by Ewing (Reid Ewing and Hamidi 2014; R Ewing et al., n.d.); the building operational energy dataset was built through coupling of two unique datasets including Chicago Energy Benchmarking (2016) (City of Chicago n.d.) and Chicago Energy Usage (2010) (City of Chicago n.d.) datasets. The Chicago

Energy Benchmarking dataset provides disaggregated data at building level for buildings greater than 50,000 sq.ft. While Chicago Energy Usage (2010) dataset provides energy data for buildings of all sizes with block-level geographical identification.

The UEUM workflow, as illustrated in Figure 2, proceeds as follows:

- *Data Preparation* including three steps: locate data, treat missing data, and process and clean data. After locating data as discussed, to maximize use of available information in the CBF dataset, several statistical methods including multiple imputations (Rubin 1996) was done to handle missing data; e.i. building height information by applying the valid frequency inference. The outliers as extreme observations in the building energy datasets were identified through statistical tests (Kutner, Nachtsheim, and Neter 2004) and their influential impacts on individual regression parameters were assessed through the Cook's Distance test (Cook 1977). Finally, extreme outliers with significant influence were dropped out of the datasets. To test the model regarding normal distribution, Quantile–normal (q-norm) plot which is considered as a common normality test (Miller 1997 Of and Statistics, n.d.), p-norm and Kornel Density plots were applied.
- *Pattern Extraction is* applied through using the most promising Machine Learning clustering algorithm, k-means, (Ahmad et al. 2018; Jovanović, Sretenović, and Živković 2015; Amasyali and El-Gohary 2018a) to extract the actual archetypes/typologies of buildings with certain similarities together and learn underlying patterns. The K-means algorithm generates K clusters by dividing M points into N dimensions to minimize the sum of squares of errors within clusters (Hartigan and Wong 1979). Then localized variables such as building height typologies were added to the model. Incorporating the actual urban spatial patterns, building characteristics and urban context improve the accuracy of the city scale energy use prediction significantly.
- *Prediction* compasses the train model, validate, compare and predict energy consumption for all buildings in the city where the energy use data is not available based on the enhanced model. The energy use prediction as a regression problem approximates a mapping function from input variables; e.i. building characteristics, urban attributes, and occupant characteristics and building operational energy consumption as the output variable. Six machine learning models were trained including Multiple linear regression (MLR), Nonlinear Regression (NLR), Random Decision Forest (RDF), Classification and Regression Trees (C&RT), K-Nearest Neighbors (K-NN), and Artificial Neural Networks (ANNs), which are among promising data-driven techniques (Ahmad et al. 2018; Jovanović, Sretenović, and Živković 2015; Amasyali and El-Gohary 2018a) on the merged dataset.
- *Validation* process was done to achieve solid results. The cross-validation method as a most effective validation technique (Torabi Moghadam et al. 2018; Amasyali and El-Gohary 2018b) based on Random Sub-sampling was applied to avoid biased results. Data was split to train and test of 80% / 20%. Then the models were compared regarding their prediction performance based on the most widely used evaluation metrics including the Mean Absolute Deviation (MAD), Mean Square error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). These performance metrics are computed by measuring the errors between the predicted and actual values that means the lower the values of MAD, MSE, RMSE, and MAPE show the better performance of the model. As the final step, the results are compared and an enhanced prediction model was proposed and validated against aggregated city-level data.
- *Visualization* was done through developing of a GIS web-based platform which allows communication and visualization of the urban energy use predictions at multi-scales including building, block, neighborhood, and city levels.

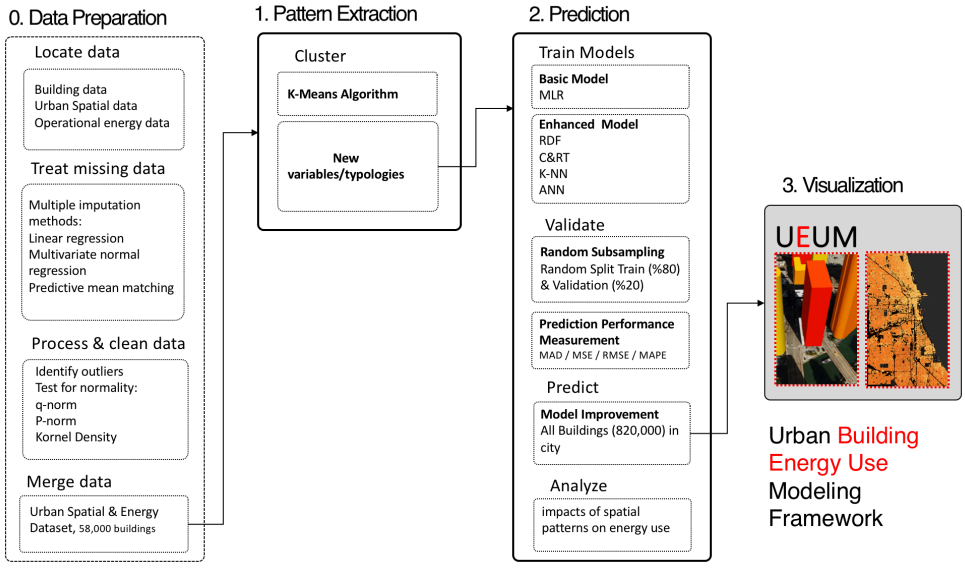

Figure 2: The UEUM workflow. Source: (Authors 2019)

2.0. RESULTS

The Figure 3. presents the performance evaluation of the prediction models applied in this research including the Multiple linear regression (MLR), Nonlinear Regression (NLR), Random Decision Forest (RDF), Classification and Regression Trees (C&RT), K-Nearest Neighbors (K-NN), and Artificial Neural Networks (ANNs) algorithms. The most common predictive model evaluation metrics: Mean Absolute Error (MAE), Mean Square error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) were employed to assess and compare the performance of the models. These metrics are calculated based on measuring the errors between the predicted and actual values. Therefore, the lower the values of MAD, MSE, RMSE, and MAPE show the better performance of the model. The results suggest that among the six algorithms modeled in this research, K-NN provides the best predictive performance with MAE of 0.08; while MLR provides the weakest predictive model with MAE of 0.22. The results show that the K-NN model performs best. The other algorithms provide results better than MLR but no significant differences observed between RDF, ANNs, C&RT models and MLR model. The result suggests that MLR model enables the energy use prediction fairly well with no significant difference, compared to RDF, ANNs, C&RT which are computationally expensive and time-consuming models for energy use prediction at city level. R2, the coefficient of determination, is a measurement metric of how well the regression model describes the observations and shows the percentage of variations that are explained by independent variables (Ohtani 2000). In the MLR model, the R2 value of 0.28, indicates that the model explains 28% of the variance in building operational EUI for buildings in Chicago. MLR as a common method for energy use prediction which has been employed widely in the previous studies (C. E. Kontokosta 2015), shows a R2 value of 0.28, indicating that the model explains 28% of the variation in energy use of the buildings in the model. R2 for other models (NLR, RDF, C&RT, K-NN, and ANNs) was calculated based on actual vs. predicted energy use values, shown in Figure 3. While among these models, K-NN significantly provides an improved perdition model with R2 of 0.75 that shows K-NN is able to explain 75% of the variation of energy use of the buildings in the model. It should be noted that the R2 is used to explain the linear association between variables and it fails to capture such association in non-linear models. Here R2 is used only for comparison purposes between the linear and non-linear models which we used the developed equation from the nonlinear models through predicting energy use values and plot them against the actual values on y-axises and then estimated the R2 values for each model. Then the improved model was applied to predict

energy use for around 820,000 buildings in the city. The model evaluates the energy performance of city in a multi-scale resolution analysis which maximizes the use cases and allows for a more comprehensive energy decision-making and policy (Figure 4).

3.0. DISCUSSION

The results of this research suggest that the urban energy prediction accuracy can be increased significantly by using disaggregated data at building level and incorporating the actual urban spatial patterns. Also, the more advanced machine learning methods enable an improved perdiction model. The findings of this study also provide empirical evidence on how spatial characteristics of neighborhoods impact the urban energy performance. The results of this study on the combined urban energy and spatial pattern dataset across neighborhoods in Chicago show the impact of the three major energy use determinants incorporated in the model including building characteristics (building height, type, size, and age); urban attributes representing sprawl dimensions; and occupant characteristics (total occupants, household size, worker density, weekly working hours, and percentage of occupied units) were found to be statistically significant at 95% confidence level, as measured by p-value of below 0.05. However, there are other energy use determinants such as socioeconomic and behavioral factors and other important building factors such as renovation year and building systems have been excluded from this research because of lack of data availability. This highlights a limitation in this research. Considering other influencing factors could contribute to more comprehensive low carbon urban policies and could be explored by future research.

CONCLUSION

The UEUM framework developed by this research is able to predict energy use at multiple scales of building, block, neighborhood, and city scales. This model captures the urban building operational energy use. The results of this research suggest that urban energy prediction accuracy can be increased by using disaggregated data at building level and incorporating the actual urban spatial patterns. Also, the more advanced machine learning methods enable an improved perdiction model. Among the six promising machine learning algorithms tested in this research, K-NN showed the best predictive performance. The finding of this study also provides empirical evidence on how spatial characteristics of neighborhoods impact the urban energy performance. As this research continues, in another article we intend to examine the association between variables in the model in detail. This framework has the potential to provide a more accurate and holistic image of urban energy use and the impact of different design decisions on energy consumption and to help designers, planners, and policymakers better understand the existing urban energy use profiles and project the environmental impacts associated with alternative scenarios of urban development.

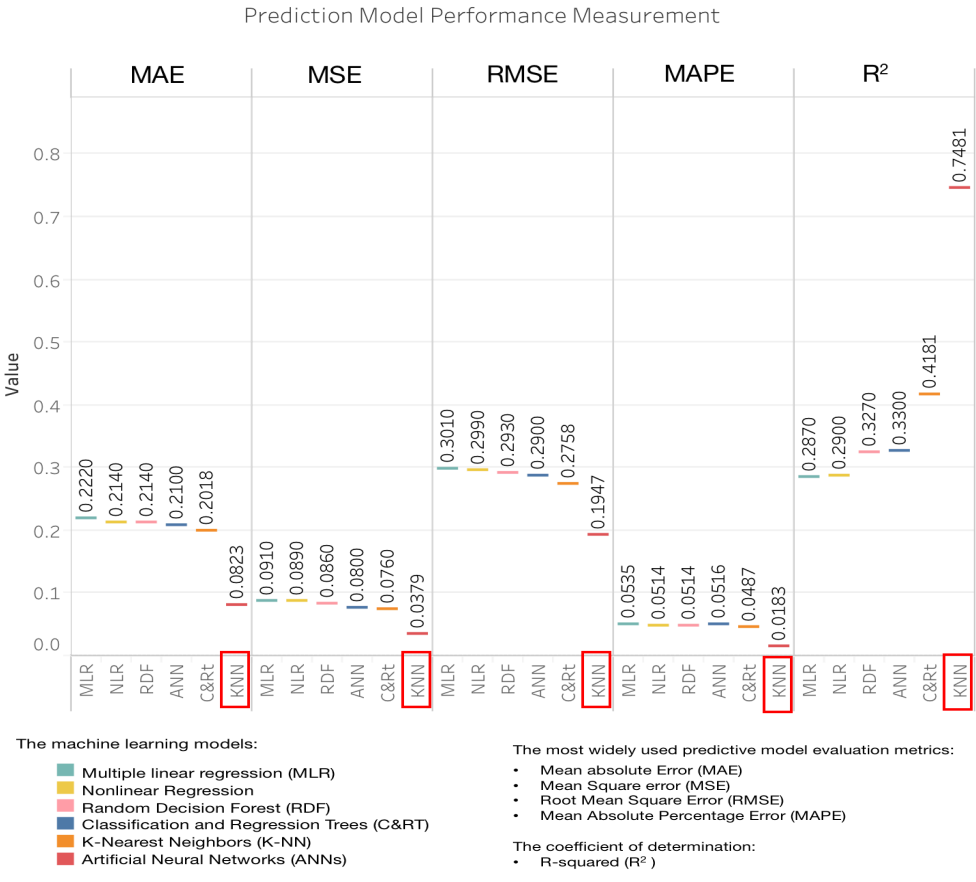

Figure 3: The performance evaluation of the prediction models. Source: (Authors 2019)

Figure 4: Multi-Scale Building Energy Use Modeling and Analysis. Source: (Authors 2019)

REFERENCES

Ahmad, Tanveer, Huanxin Chen, Yabin Guo, and Jiangyu Wang. 2018. “A Comprehensive Overview on the Data Driven and Large Scale Based Approaches for Forecasting of Building Energy Demand: A Review.” *Energy and Buildings* 165: 301–20. https://doi.org/10.1016/j.enbuild.2018.01.017.

Amasyali, Kadir, and Nora M. El-Gohary. 2018a. “A Review of Data-Driven Building Energy Consumption Prediction Studies.” *Renewable and Sustainable Energy Reviews*. https://doi.org/10.1016/j.rser.2017.04.095.

———. 2018b. “A Review of Data-Driven Building Energy Consumption Prediction Studies.” *Renewable and Sustainable Energy Reviews* 81 (March 2017): 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095.

“Chicago Energy Benchmarking, 2016, City of Chicago, Data Portal.” n.d.

City of Chicago. n.d. “Building Footprints (Current), City of Chicago, Data Portal.” Accessed June 27, 2018a. https://data.cityofchicago.org/Buildings/Building-Footprints-current-/hz9b-7nh8?category=Buildings&view_name=Building-Footprints-current-.

———. n.d. “Chicago Energy Benchmarking - 2016 Data Reported in 2017, City of Chicago, Data Portal.” Accessed June 27, 2018b. https://data.cityofchicago.org/Environment-Sustainable-Development/Chicago-Energy-Benchmarking-2016-Data-Reported-in-/fpwt-snya.

———. n.d. “Energy Usage 2010, City of Chicago, Data Portal.” Accessed June 27, 2018c. https://data.cityofchicago.org/Environment-Sustainable-Development/Energy-Usage-2010/8yq3-m6wp.

Cook, Dennis. 1977. “Detection of Influatial Observation in Linear Regression.” *The SAGE Handbook of Regression Analysis and Causal Inference* 19 (1): 15–18. http://dx.doi.org/10.4135/9781446288146.

EIA. 2012. “Annual Energy Review, US Energy Information Administration (EIA).” 2012. https://www.eia.gov/totalenergy/data/annual/showtext.php?t=ptb0201a.

Ewing, R, T Schmid, R Killingsworth, A Zlot - Urban Ecology, and undefined 2008. n.d. “Relationship between Urban Sprawl and Physical Activity, Obesity, and Morbidity.” *Springer*.

Ewing, Reid. 2010. “The Impact of Urban Form on U . S . Residential Energy Use.” *Housing Policy Debate* 19 (April 2013): 37–41. https://doi.org/10.1080/10511482.2008.9521624.

Ewing, Reid, and Shima Hamidi. 2014. “Measuring Urban Sprawl and Validating Sprawl Measures.” *Metropolitan Research Center, University of Utah.* https://doi.org/10.1017/CBO9781107415324.004.

Hartigan, J. A., and M. A. Wong. 1979. “Algorithm AS 136: A K-Means Clustering Algorithm.” *Applied Statistics* 28 (1): 100. https://doi.org/10.2307/2346830.

Howard, B., L. Parshall, J. Thompson, S. Hammer, J. Dickinson, and V. Modi. 2012. “Spatial Distribution of Urban Building Energy Consumption by End Use.” *Energy and Buildings* 45: 141–51. https://doi.org/10.1016/j.enbuild.2011.10.061.

Hsu, David. 2015. “Identifying Key Variables and Interactions in Statistical Models of Building Energy Consumption Using Regularization.” *Energy* 83 (April): 144–55. https://doi.org/10.1016/J.ENERGY.2015.02.008.

Jovanović, Radiša Ž., Aleksandra A. Sretenović, and Branislav D. Živković. 2015. “Ensemble of Various Neural Networks for Prediction of Heating Energy Consumption.” *Energy and Buildings* 94 (May): 189–99. https://doi.org/10.1016/J.ENBUILD.2015.02.052.

Kontokosta, Constantine, Bartosz Bonczak, and Marshall Duer-balkind. 2016. “DataIQ – A Machine Learning Approach to Anomaly Detection for Energy Performance Data Quality and Reliability.”

Kontokosta, Constantine E. 2015. “A Market-Specific Methodology for a Commercial Building Energy Performance Index,” 288–316. https://doi.org/10.1007/s11146-014-9481-0.

Kutner, Michael H., Chris. Nachtsheim, and John. Neter. 2004. *Applied Linear Regression Models.* McGraw-Hill/Irwin.

Martin, Miguel, Nyuk Hien Wong, Daniel Jun Chung Hii, and Marcel Ignatius. 2017. “Comparison between Simplified and Detailed EnergyPlus Models Coupled with an

Urban Canopy Model." *Energy and Buildings* 157 (May): 116–25. https://doi.org/10.1016/j.enbuild.2017.01.078.

Miller 1997 Of, Basics, and Applied Statistics. n.d. *BEYOND*.

National Cancer Institute. n.d. "Urban Sprawl Data for the United States- Geographic Information Systems & Science." Accessed June 27, 2018. https://gis.cancer.gov/tools/urban-sprawl/.

Ohtani, Kazuhiro. 2000. "Bootstrapping R2 and Adjusted R2 in Regression Analysis." *Economic Modelling* 17 (4): 473–83. https://doi.org/10.1016/S0264-9993(99)00034-6.

Reinhart, Christoph F., and Carlos Cerezo Davila. 2016. "Urban Building Energy Modeling – A Review of a Nascent Field." *Building and Environment* 97 (February): 196–202. https://doi.org/10.1016/J.BUILDENV.2015.12.001.

Resch, Eirik, Rolf André Bohne, Trond Kvamsdal, and Jardar Lohne. 2016. "Impact of Urban Density and Building Height on Energy Use in Cities." *Energy Procedia* 96 (1876): 800–814. https://doi.org/10.1016/j.egypro.2016.09.142.

Rubin, Donald B. 1996. "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association* 91 (434): 473–89. https://doi.org/10.1080/01621459.1996.10476908.

Ruby, P Rode - The economy of sustainable construction., and undefined 2014. n.d. "The Politics and Planning of Urban Compaction: The Case of the London Metropolitan Region." *Src.Holcimfoundation.Org*.

Sola, Alaia, Cristina Corchero, Jaume Salom, Manel Sanmarti, Alaia Sola, Cristina Corchero, Jaume Salom, and Manel Sanmarti. 2018. "Simulation Tools to Build Urban-Scale Energy Models: A Review." *Energies* 11 (12): 3269. https://doi.org/10.3390/en11123269.

Swan, Lukas G., and V. Ismet Ugursal. 2009. "Modeling of End-Use Energy Consumption in the Residential Sector: A Review of Modeling Techniques." *Renewable and Sustainable Energy Reviews* 13 (8): 1819–35. https://doi.org/10.1016/j.rser.2008.09.033.

Torabi Moghadam, Sara, Jacopo Toniolo, Guglielmina Mutani, and Patrizia Lombardi. 2018. "A GIS-Statistical Approach for Assessing Built Environment Energy Use at Urban Scale." *Sustainable Cities and Society* 37 (February): 70–84. https://doi.org/10.1016/J.SCS.2017.10.002.

Zhou, Weiqi, Ganlin Huang, and Mary L. Cadenasso. 2011. "Does Spatial Configuration Matter? Understanding the Effects of Land Cover Pattern on Land Surface Temperature in Urban Landscapes." *Landscape and Urban Planning* 102 (1): 54–63. https://doi.org/10.1016/J.LANDURBPLAN.2011.03.009.