Energy Usage Analysis at the University of Virginia

Yuxiang Xiang, Zhuoyang Zhou, Dongheon Lee
University of Virginia, yx5na, zz4nz, ddl3vd@virginia.edu

***Abstract* - The topic of this research is to construct models that capture the energy usage patterns of buildings at University of Virginia (UVa) and predict their future energy usage. We also applied classification and clustering methods to UVa buildings based on energy usage tendency to better our understanding of UVa's energy consumption mechanism. Currently, the energy consumption makes up a large amount of all UVa expenses and is projected to increase steadily due to the growth of UVa. Using the daily energy consumption data including total energy usage, electricity, heating, steam, chilled water and hot water usage from over 200 buildings at UVa over the past year, this research focuses on analyzing the energy usage trends, which leads to better prediction of future energy usage, improved construction intelligence, and better maintained facilities. The basic methodology of this research revolves around the architecture of regression models and time-series models of energy usage consumption and prediction based on these models. Autoregressive moving average (ARMA) models for different types of building based on past energy consumption were constructed at first, combined with the analysis of the ACF & PACF plots and residuals. Then, generalized autoregressive conditional heteroscedasticity (GARCH) model is fitted to account for the changing variance of the error term. Finally, decision tree models and logistic regression models are used to evaluate the performance of our models.**

Index Terms - Clustering, Energy Trends, Linear Regression, Time Series.

INTRODUCTION

Since the turn of the new century, the aggregate energy usage within our society, or, more specifically, within the University of Virginia has been on the rise. It is our responsibility to make data-driven decisions on energy usage.

Based on data analytics methods, we could accurately describe the energy usage patterns of over two hundred buildings at UVa. Our model construction and prediction allow for more effective control and smarter decision-making. Data analytics methods can improve the overall forecast accuracy, which helps better understanding of energy usage patterns and allocation of resources.

Previous research on this subject revealed that a combination of linear model and a seasonal autoregressive integrated moving average (SARIMA) model could be useful in predicting energy usage [1]. We took a multivariate approach, using autoregressive moving average (ARMA) model, generalized autoregressive conditional heteroskedascity (GARCH) model on non-seasonal types of energy and linear regressions on seasonal types of energy. The ultimate goal of this research is to have a better understanding of the energy usage pattern and make future predictions.

Our approach focused on the construction of suitable models for different energy and building types. Our research revealed that the linear regression models, ARMA models, and GARCH models to be most effective at accurately describing energy usage patterns at the university. And also, based on the parameters of time series models and correlations, we performed supervised classification to evaluate the performance of our models.

The original data set we received from the UVa facilities management consisted of over 20 millions of records of energy usage at UVa buildings from September 1st 2014 to August 31th, 2015. However, only a fraction of the buildings had consistent data throughout the whole year without any missing values. For each building, there were six different types of energy usage: total energy, electricity, heating, chilled water, steam and hot water. There were 262 buildings which had no missing data in all six kinds of energy records, so we constructed our model mainly based on the data of those 262 buildings. All energy usages were recorded by the hour; we selected the energy records at 4 PM to represent the daily energy usage.

For each building, the sum of heating water and steam usage corresponded with the heating value. One possible explanation is that each building uses one type or the other for heating purposes. Thus, for the remainder of the research, we narrowed our scope to energy, electricity, chilled water and heating consumption.

The following plots describe the energy usage of O-hill dining hall, a representative building at UVa:

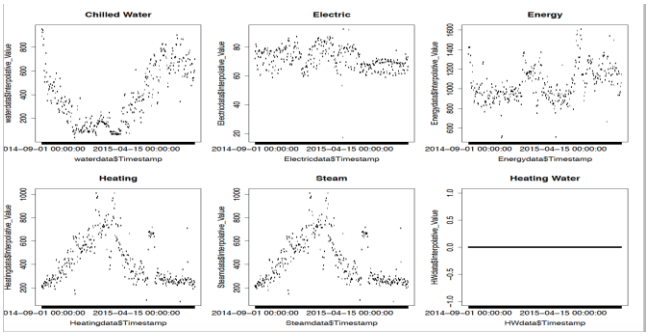

FIGURE I
OVERVIEW OF O-HILL DINNING HALL ENERGY USAGE

LITERATURE REVIEW

Recently, the topic of saving energy and decision making that leads to efficient use of energy has become popular [2]. Based on the growing interest in energy efficiency and the significant rise in energy demand, policy makers are making use of various energy forecasting models in their decision making [3].

Many theories have been crafted based on classical time series analysis. N. Fumo, and M. Biswas [4] applied both simple and multiple linear regression models on prediction of residential energy consumption and suggested the use of principal component analysis (PCA) method to reduce the collinearity of the models. P. Chujai et al. [5] used Akaike Information Criterion (AIC) and Root Mean Square Error (RMSE) and discovered that ARIMA models are the most suited at forecasting household electric power consumption at monthly and quarterly intervals, while ARMA models are the best at forecasting daily and weekly power consumption. Z. Shen and M. Ritter [6] applied different GARCH-type models, such as exponential GARCH (EGARCH) and threshold GARCH (TGARCH) in order to account for the volatility of wind speed and its nonlinear and asymmetric time-varying properties. They also fitted Markov regime switching GARCH (MRS-GARCH) model for forecasting the volatility of wind energy.

Various methods have been used to cluster buildings for energy usage analysis. R. Silipo and P. Winters [7] applied K-Means algorithm to group together meter IDs based on similarities in daily and weekly values of the electricity usages. They were able to detect temporal trends in some of the major clusters. Z. Yu et al. [8] created energy demand predictive model based on decision trees. Their features included outside temperature, room air conditioners (RAC) and house conditions.

METHODOLOGY

I. Data Cleaning

From Figure 1 in introduction, it should be noted that there are several inconsistent observations. For example, some points are negative; others suddenly drop to zero out of context. Also, there are a few points that are extremely high or low and do not line up with the general trend. It is reasonable to attribute these outliers to data entry errors. We replaced the negative values with the observation of one day before. For missing values, we filled them in with the average consumption from one day before and one day after. Finally, we set a criteria of average plus three times the standard deviation. If an observation exceeded this threshold, we replaced it with the criteria value.

II. Data Analytics of the Building Energy consumption

With the help of data visualization, we were able to see that the four types of energy could be divided into two groups. Chilled Water showed a strong seasonal trend with high values in the summer and low values in the winter [9]. On the contrary, Heating consumptions were reported with high values in the winter and virtually zero values in the summer. On the other hand, electricity and energy usage did not show such a strong seasonal trend.

II.I Analysis on data with strong seasonal trend, with Chilled Water as an example

Figure 2 shows the number of buildings with zero values for chilled water by day. These values could be attributed to final weeks and spring break. The seasonal trend becomes clear in Figure 3, which displays the aggregate sum of chilled water consumption over all buildings at UVa.

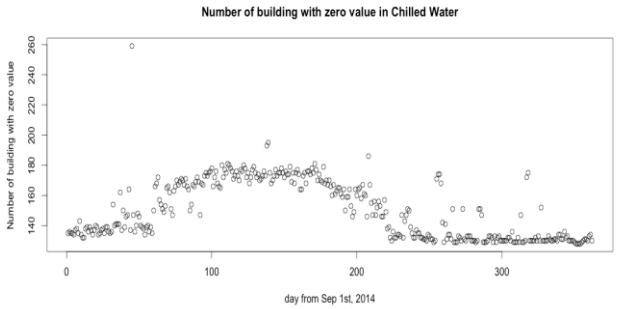

FIGURE II
NUMBER OF BUILDINGS WITH ZERO VALUE IN CHILLED WATER

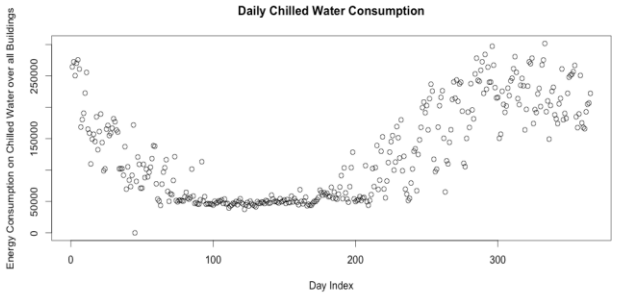

FIGURE III
DAILY CHILLED WATER CONSUMPTION

Using the temperature data, we constructed a multiple linear model of chilled water consumption with daily maximum temperature, daily minimum temp, CDD (the number of degrees that a day's average temperature is above 65° Fahrenheit), and HDD (the number of degrees that a day's average temperature is below 65° Fahrenheit). There was a strong correlation between HDD and CDD, so we decided to drop CDD from our model.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \qquad (1)$$

TABLE I
VARIABLE SIGNIFICANCE IN THE LINEAR MODEL

Variable	P - value
HDD	< 2e-16 ***
HIGH_TEMP	1.60e-15 ***

LOW_TEMP	< 2e-16 ***
poly(dayIndex, 2)1	2.49e-05 ***
poly(dayIndex, 2)2	0.00111 **
dayBefore	0.00272 **

The weather could explain the majority of the variation in the chilled water consumption. All explanatory variables are significant in our model. The R-squared value was over 0.85.

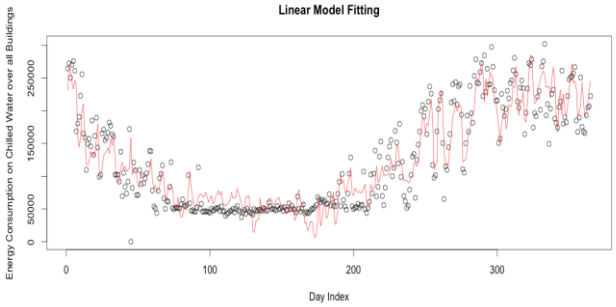

FIGURE IV
LINEAR MODEL FITTING OF CHILLED WATER

Our preliminary results with the multiple linear model captured the overall trend of chilled water consumption to a satisfactory degree. In the following section, we incorporated the seasonality of the chilled water consumption and constructed a model based on piecewise function.

We split the plot we have just created into three parts: from Sep 1st, 2014 to the last academic day of fall semester, Dec 5th, 2014, from Dec 6th, 2014 to Mar 6th, 2015, the beginning of spring break, and from Mar 7th, 2015 to August 31st, 2015.

The model we built based on variables temperature, HDD, dayIndex and dayBefore was satisfactory with a good overall performance. However, due to the large variance in the data, it might not capture all the variations in the data. We would re-discuss this problem towards the end of our research.

TABLE II
MODEL PERFORMANCE ON EACH PERIOD OF DATA

Adjusted R^2	Overall	Period One	Period Two	Period Three
Chilled Water	0.8609	0.7676	0.4711	0.8304
Heating	0.8778	0.7463	0.6605	0.821

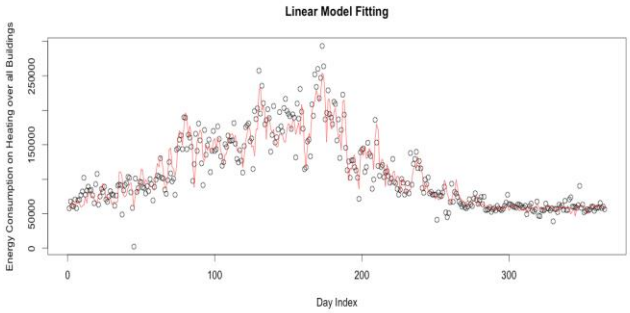

FIGURE V
LINEAR MODEL FITTING OF HEATING

The heating consumption at UVa displayed a similar seasonal trend with a somewhat higher variance at its peak period. We omitted further discussion on this matter due to its similarities with chilled water consumptions.

II.II Analysis on the data without strong seasonal difference

II.II.I Fitting time series models for total energy consumption

For the group of energies that did not show a strong seasonal trend, we constructed time series models. We also studied the time series coefficients for patterns among buildings, which could facilitate the building clustering process.

Due to the impracticality of finding the best fitted model for every single building, we resorted to using the daily sum. In this case, the time series models we constructed might not be perfect for each individual building, but could help explain the total energy usage.

Before constructing our time series model, we first applied a smoothing spline to the time series data. Then, we fit the ARMA model and GARCH model. After obtaining the best ARMA and GARCH model for daily sum of energy usage, we applied the models to each building and compared their coefficients.

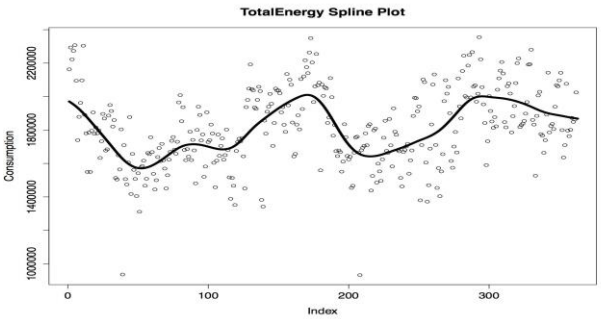

FIGURE VI
SPLINE PLOT OF TOTAL ENERGY DAILY SUM

After using the smoothing spline to de-trend the time series data, we generated ACF and PACF plots. Figure 7 shows the representative plots of total energy usage. For our

time series model, we did not use any differencing term given the following reasons:

An ARIMA model is classified as an ARIMA(p,d,q) model, where:

- p is the number of autoregressive terms,
- d is the number of no seasonal differences needed for stationary, and
- q is the number of lagged forecast errors in the prediction equation.

Since our time series data was not affected by any seasonality, we set **d** equal to zero. Thus, we constructed a ARMA(p,q) model, using the following equation:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \qquad (2)$$

Then, we compared the AIC values of different potential ARMA models and arrived at ARMA (2,1) as our final model for total energy sum. In order to see if this ARMA model fits well, we used the Ljung-Box test, which yielded a p-value of 0.7605. The null hypothesis for a Ljung-Box test is that the model is a good fit. Therefore, given our large p-value that is greater than any statistical significance level ($p > \alpha$: Fail to reject Ho), we failed to reject our null hypothesis. Hence, ARMA (2,1) was an acceptable model.

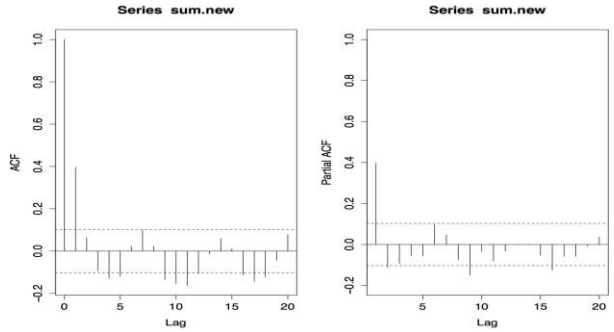

FIGURE VII
ACF AND PACF PLOTS OF TOTAL ENERGY DAILY SUM

Next, we fit ARMA (2,1) to each building and stored all the coefficients of each building.

Below are the coefficients ar1 and ma1 of each building.

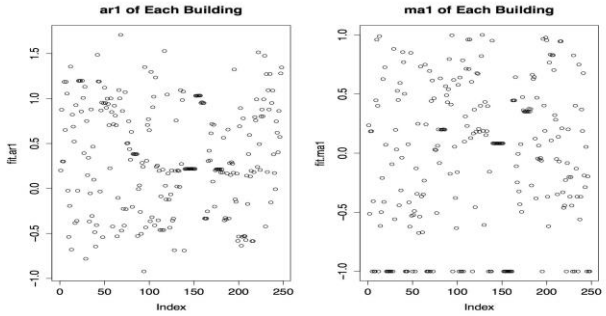

FIGURE VIII
COEFFICIENTS AR1 AND MA1 OF ARMA (2,1) MODEL

Then we applied the same method again, but with GARCH model. Compared to the ARMA model, GARCH model that assumes the error variance [11]. We repeated the steps as before and picked the model with the lowest AIC value. Our GARCH (1,1) model had a similar AIC value with our GARCH (2,1) and (2,2). Taking into consideration that GARCH (1,1) is most widely used, we decided to use GARCH (1, 1) as our model for daily total energy usage. Once again, we fit the same model to each of the buildings and stored the coefficients.

The coefficients we obtained from fitting the ARMA and GARCH to each building would later be used in both supervised and unsupervised classification.

II.II.II Fit time series models for specific buildings' energy consumption

In order to check whether the general time series models we fit above would also work well on specific, representative buildings, we selected several buildings to evaluate. We defined features based on each building's daily energy usage and its correlation with temperature. Here, we posited the idea that our time series models are suitable for those buildings with high self-correlations.

Since each building data consists of records from 365 days, we computed its correlation with lag-1, lag-2, lag-3 and temperature. The results are given as following:

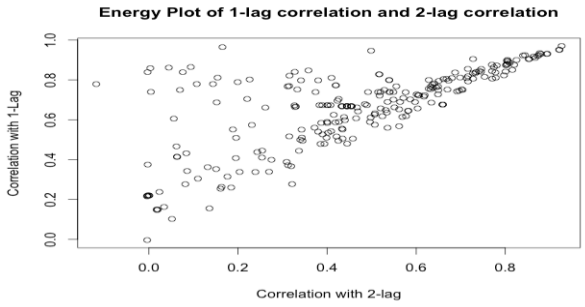

FIGURE IX
ENERGY PLOT OF 1-LAG CORRELATION AND 2-LAG CORRELATION

Then we looked into the absolute value of correlation between 1-lag and temperature.

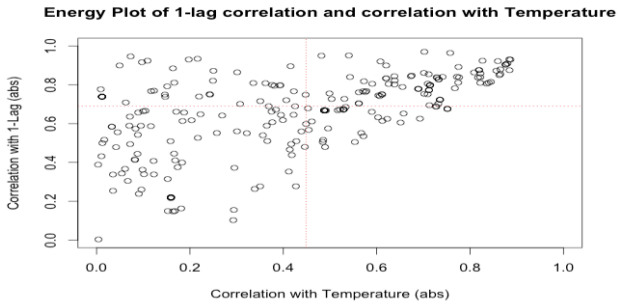

FIGURE X
ENERGY PLOT OF 1-LAG CORRELATION AND CORRELATION WITH TEMPERATURE

The two lines in Figure 10 correspond with the mean absolute value of correlation with lag-1 and correlation with temperature, which are 0.6891465 and 0.4487195, respectively. In other words, the energy usage of most buildings showed a moderate dependency on the energy usage of the day before or the temperature, making it significantly difficult to make accurate predictions. As for the buildings with a high correlation, it was possible to fit a fairly accurate model.

Then we picked specific buildings to apply the time series models based on what we have found above. For example, the energy usage of Old Cabell Hall, a representative building with a large concert hall at UVa, had a correlation with lag-1 of 0.8411568. Since it had a very high correlation with lag-1, we could safely assume that our time series model was adequate.

Thus, we fit the ARMA (2,1) model on Old Cabell Hall's energy consumption. We applied the Box-Ljung test to the residuals from the ARMA (2,1) model fit to determine whether the residuals are random. In this example, the Box-Ljung test reported a p-value of 0.9155, implying that the residuals are indeed random and our model is a good fit.

Then we created the ACF and PACF plots. The graphical evaluation of the residuals from this model is shown below in Figure 11. The residual plots of ACF and PACF indicated that we cannot reject the hypothesis that the residuals are uncorrelated. Thus, we conclude that the ARMA (2,1) model provides an adequate to the energy usage at Old Cabell Hall.

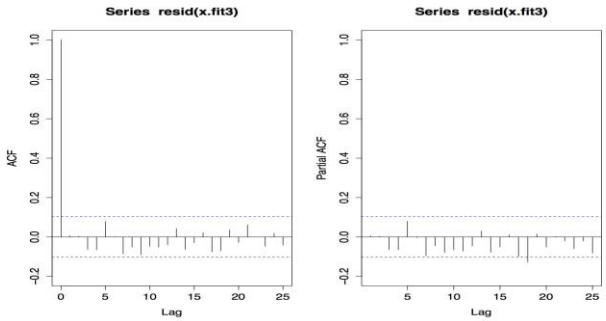

FIGURE XI
ACF AND PACF PLOTS OF RESIDUALS

III. Building Classification

In this section, we built a classification system based on supervised learning methods. First, each of the building was categorized into four groups, according to their primary usage information. Then, we used logistic regression and decision trees to predict the results of our classification system.

We used a logistic regression of binomial family with an indicator variable for whether a building is classified as "housing" or not.

TABLE III
VARIABLE SIGNIFICANCE IN LOGISTIC REGRESSION MODEL FOR HOUSING

Variable	P – value
Correlation_Lag1	0.0106*
Correlation_TEMP	0.0026**
sfwithfloor	0.0269*
ar2	0.0056**

Variable Correlation_Lag1, Correlation_TEMP, and ar2 are consistent with those that have been used earlier in this research. Variable sfwithfloor measures the net square footage of a given building that has been scaled according to the number of floors. All variables have p-values less than 0.05, and deviance and AIC values indicate that the model is adequate. Using cross-validation methods, we found out that our classification based on logistic regression on housing was 77.0%.

Then, a decision tree was built using the same variables. The variable sfwithfloor (net square footage of a building that has been scaled according to the number of floors) was selected as the root node in this case, since it had the most impact on the classification results.

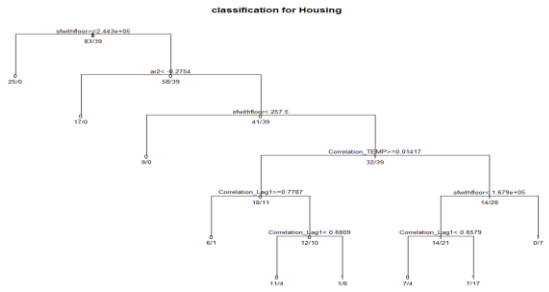

FIGURE XII
DECISION TREE FOR CLASSIFICATION OF HOUSING

Each method was applied to all four building groups. Presented below are the reported accuracies of each classification method.

TABLE IV
CLASSIFICATION ACCURACY

Classification Accuracy	Logistic Regression	Decision Tree
Housing	0.770592	0.811475
Student Life	0.868853	0.844262
Academics	0.836066	0.823942
Utility	0.680328	0.614754

CONCLUSION & FUTURE WORK

Linear model proved to be the most powerful tool at capturing the energy usage pattern for chilled water usage and heating usage. The strong seasonality that was apparent in both data types called for a model based on piece-wise function. A multiple linear model with temperature, day index and 1-lag provided the most adequate fit.

For energy usage and electricity usage, both of which displayed no particular seasonality, we adapted ARMA models and GARCH models. We built our models based on

the daily energy sum and applied them to representative buildings. Buildings that showed a high self-correlation were better fitted by our time-series model.

Finally, we created a decision tree model and a logistic regression model for building clustering purposes. We found out that our classification system built using the coefficients from our linear model and time-series models could effectively predict the primary usage of the building. Our classification accuracy could benefit from any supplementary information regarding the buildings, such as the building material, building location, average occupancy level, and so on.

ACKNOWLEDGMENT

We would like to thank our advisor Daniel Keenan for his guidance and advice throughout this project. We would also like to thank E. Scott Martin from UVA Energy Services, who provided us valuable data and clear instructions; his expertise helped us complete this project. We'd also like to show our appreciation to Professor Donald Brown as well as all faculty members from Data Science Institute in UVa for their generous help and assistance in the whole procedure. All authors contributed equally to the paper.

REFERENCES

[1] C. Heylman, Y. Kim and J. Wang, "Forecasting Energy Trends and Peak Usage at the University of Virginia", *2015 Systems and Information Engineering Design Symposium*, pp. 362-268, 2015

[2] L. Suganthi, A. Samuel, "Energy Models for Demand Forecasting—A Review", *Renewable and Sustainable Energy Reviews*, Vol.16, Issue 2, 2012

[3] E. Worrell, S. Ramesohl and G. Boyd, "Advances in Energy Forecasting Models Based on Engineering Economics", *Annual Review of Environment and Resources*, vol. 29, no. 1, pp. 345-381, 2004.

[4] N. Fumo and M. Biswas, "Regression Analysis for Prediction of Residential Energy Consumption", *Renewable and Sustainable Energy Reviews*, vol. 47, pp. 332-343, 2015.

[5] P. Chujai, N. Kerdprasop and D. Kerdprasop, "Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models", i*nternational Multi Conference of Engineers and Computer Scientists*, Hong Kong: Newswood Ltd., 2013.

[6] Z. Shen and M. Ritter, "Forecasting Volatility of Wind Power Production", *SFB 649 Discussion Paper 2015-026*, 2015

[7] R. Silipo and P. Winters, "Big Data, Smart Energy, and Predictive Analytics", *KNIME*, 2013.

[8] Z. Yua, F. Haghighat, B. Fung, H. Yoshinoc, "A Decision Tree Method for Building Energy Demand Modeling" *Energy and Buildings*, Vol. 42, Issue 10, 2010

[9] O. Gerin, B. Bleys, K. Cuyper, "Seasonal Variation of Hot and Cold Water Consumption in Apartment Buildings", *Proceedings of CIB W062 40th International Symposium on Water Supply and Drainage*, 2014

[10] Z. Yua, F. Haghighat, B. Fung, H. Yoshinoc, "Multi Model Prediction and Simulation of Residential Building Energy in Urban Areas of Chongqing, South West China", *Energy and Buildings*, Vol. 81, 2014

[11] Y. Wang, C. Wu, "Forecasting Energy Market Volatility Using GARCH Models: Can Multivariate Models Beat Univariate Models?", *Energy Economics*, Vol.34, Issue 6, 2012

AUTHOR INFORMATION

Yuxiang Xiang, Zhuoyang Zhou, Dongheon Lee
Masters Students, Data Science Institute, University of Virginia.