Forecasting Energy Demand in Large Commercial Buildings Using Support Vector Machine Regression

David Solomon*, Rebecca Winter†, Albert Boulanger‡, Roger Anderson‡ and Leon Wu‡§
*Department of Earth and Environmental Sciences, Columbia College, New York, NY 10027 USA
†Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027 USA
‡Center for Computational Learning Systems, Columbia University, New York, NY 10027 USA
§Department of Computer Science, Columbia University, New York, NY 10027 USA

***Abstract*—As our society gains a better understanding of how humans have negatively impacted the environment, research related to reducing carbon emissions and overall energy consumption has become increasingly important. One of the simplest ways to reduce energy usage is by making current buildings less wasteful. By improving energy efficiency, this method of lowering our carbon footprint is particularly worthwhile because it reduces energy costs of operating the building, unlike many environmental initiatives that require large monetary investments. In order to improve the efficiency of the heating, ventilation, and air conditioning (HVAC) system of a Manhattan skyscraper, 345 Park Avenue, a predictive computer model was designed to forecast the amount of energy the building will consume. This model uses Support Vector Machine Regression (SVMR), a method that builds a regression based purely on historical data of the building, requiring no knowledge of its size, heating and cooling methods, or any other physical properties. SVMR employs time-delay coordinates as a representation of the past to create the feature vectors for SVM training. This pure dependence on historical data makes the model very easily applicable to different types of buildings with few model adjustments. The SVM regression model was built to predict a week of future energy usage based on past energy, temperature, and dew point temperature data.**

I. Introduction

New York State has the lowest per capita energy use in the country partly due to the New York Metropolitan Region's transportation system, which accounts for 23% of energy consumption, while buildings consume 77% of energy. Large buildings specifically, such as Manhattan skyscrapers, consume 45% of energy in the region [14]. Therefore, one of the greatest opportunities to reduce carbon emissions in New York City, as well as in most urban areas, is through reducing energy consumption in buildings. Heating, ventilation, and air-conditioning (HVAC) is one of the principal systems in buildings that consume energy. Therefore, optimizing the energy use of HVAC systems has great potential in reducing energy consumption.

Rudin Management Inc. is one of the largest private real estate companies in New York City. The company owns over 50 buildings in Manhattan and is spearheading an environmental approach to real estate development. 345 Park Avenue, a 634 ft tall skyscraper between 51st and 52nd Street on Park Ave in Manhattan, is their flagship building in this effort. The 44-story building has been outfitted with a state-of-the-art energy monitoring system provided by MCEnergy. This log of energy demand makes it possible to predict and optimize their energy use, and the data it provides is the foundation for the predictive model presented here.

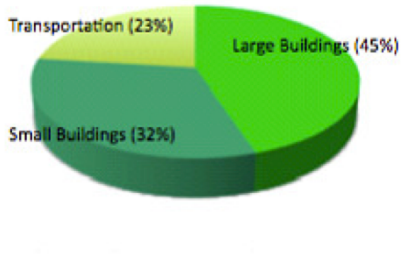

Fig. 1. Energy by sector for NYC [15].

345 Park Avenue is a large commercial building with tenants such as Bristol-Myers Squibb, the NFL, the KPMG accounting firm, the Blackstone Group, and others. Approximately 5,000 people work in the building, and there are about 1,000 visitors to the building daily. The building's regular hours are 7:00 AM to 7:00 PM Monday through Friday, and 8:00 AM to 1:00 PM on Saturdays. It costs approximately $2000 to $2500 in energy to run the HVAC system of 345 Park for an hour. 345 Park Avenue operates two large heating, ventilation, and air-conditioning plants. One is below ground level, serving the lower floors. The other is on the 34^{th} floor, serving upper floors. The building uses steam, electricity, and natural gas supplied by Con Edison to supply heat and cooling to the building.

A. Measuring Energy

There are two main ways of measuring energy use in a building. The first is the total consumption over a certain period of time, for example the number of kilowatt-hours consumed in a month. Most residential buildings are

charged solely based on their total energy consumption in a given billing cycle, and have no need for measuring their energy usage in any other way. However, the total amount of consumption over these large blocks of time does not give an accurate depiction of the energy usage trends for certain applications, particularly commercial buildings. This is because typical office buildings consume most of their energy during regular weekday office hours and use relatively little energy at night and on weekends, when the building is vacant. This requires the grid to supply these buildings with a large amount of energy at certain times, instead of an average amount constantly, making the total monthly consumption value a somewhat misleading indicator of the true energy requirements of the building. For this reason, the rate of energy consumption, called demand, is used in this study to show the amount of energy the grid must supply the building at a given time. Figure 2 shows hourly demand values of 345 Park in kilowatts; times of high demand closely follow the hours of operation of the building.

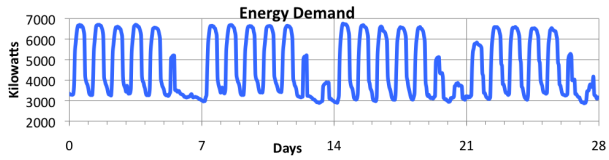

Fig. 2. Energy demand.

The energy demand of a building is determined by several factors. There are fixed building characteristics, such as size, materials, location, orientation, design that contribute to the energy consumption of a building. There are also dynamic characteristics that affect the energy demand of a building [4]. Weather has a significant effect on the amount of energy required to heat and cool the building. Figure 3 shows the correlation between temperature index (observed temperature) and energy usage for the entire New York City Region. Particularly hot days are marked by a significant increase in energy consumption. There is a less steep increase during very cold days because New York City heating comes largely from the burning of fossil fuels as opposed to electricity from the grid. A similar graph of 345 Park can be seen in figure 4. This graph shows a less steep slope with respect to weather changes because it reflects the relationship between average daily temperature and hourly energy demand during that day, not simply peak energy load as in Figure 3. Therefore, the trend is not as clearly expressed in the graph, because there is wider range of energy values.

Due to the cyclicity of 345 Park energy demand, one can see two major curves on this graph. The bottom curve shows when the building is vacated while the top curve shows when the building is in use. Points between these two curves primarily correlate to opening and closing times of the building, as the HVAC system is turning on or off. The graph becomes far more useful with these two primary curves in mind. One can see that on very cold days, the vacated building usage

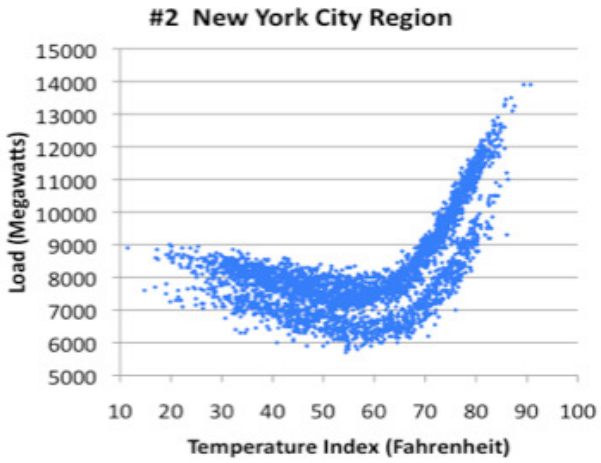

Fig. 3. Load versus temperature.

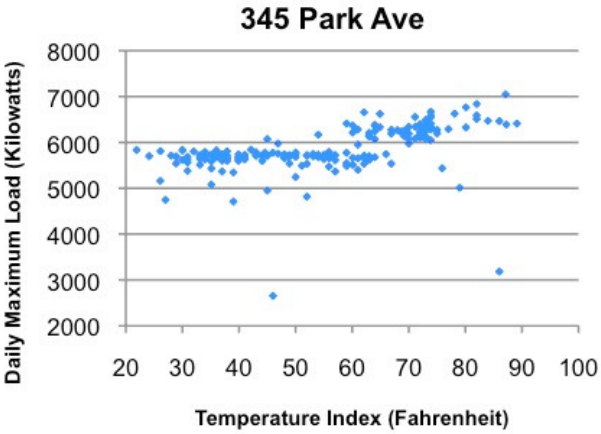

Fig. 4. Load versus temperature.

increases more steeply than the in-use building energy usage. This may be because cold winter nights can make the building far too cold, perhaps even damaging equipment overnight if the below-freezing temperatures are left unregulated. On very hot days, however, the most extreme temperatures generally occur during in-use hours, so more air conditioning is required to maintain a comfortable temperature.

Rudin explains that in order to help minimize the increased energy usage on days of particularly extreme temperatures, they try to recycle as much air in the building as possible. This reduces costs because when air is recycled, it does not need to be heated or cooled since it is generally already at room temperature. Unfortunately, recycled air can become quite high in carbon dioxide and low in oxygen because the people in the building continue to breathe in the same oxygen-depleted air without allowing it to be exposed to the outdoor atmosphere and greenery that replenish oxygen and moderate excess CO2. Therefore, 33% of the building's air is taken directly from outside and is cooled or heated to room temperature.

B. Modeling

Modeling the energy usage of 345 Park is an important step to improving the efficiency of the system. An accurate model of future energy usage can be compared to actual energy usage to look for anomalies in the actual data that may represent wasteful usage of energy. The short-term predicted energy usage, if accurate enough, could also be used to determine how much energy should be used now. For example, if the model predicts a large increase in energy usage in two hours, a moderate energy increase could be forced now to help combat the high future demand. Alternatively, if a low energy requirement is predicted for the day, pre-heating or pre-cooling times can be pushed later to decrease total consumption. Fundamentally, in order to control a pattern, one must first be able to model its behavior. Modeling the energy usage of 345 Park will allow the management to better understand their building's energy requirements, which inevitably leads to new and better ways of optimizing the system.

There are various kinds of models that can be used to predict and analyze the energy demand of a building. The DOE-2 model, created by the U.S. Department of Energy takes inputs that characterize the physical aspects of the building in order to predict its energy needs [18]. SimaPro is a tool that evaluates the embedded energy in the building's materials and construction history and also predicts operational energy. For this study, a purely operational approach to energy demand forecasting was taken. The Support Vector Machine algorithm developed by V. Vapnik in 1995 to perform Support Vector Machine Regression was used to predict energy demand for 345 Park Avenue. This method requires no inputs pertaining to the physical characteristics of the building. Rather, the method employs past hourly energy data and corresponding hourly weather data to create a model that predicts energy demand into the future. The strength of this model is that it can work for any building that has historical energy demand data.

The goal of this study is to create a highly accurate model to predict energy demand for 345 Park Avenue. In order to do so, the correct algorithm parameters were selected, weather variables to be used as data inputs were chosen, and issues of seasonal variability and timing were explored.

II. Methods

A. Data Collection and Processing

1) Energy Data: The energy demand data was collected by Con Edison and obtained from Rudin's online records provided by MCEnergy. Hourly data from January 1st, 2009 to July 13th, 2010 and September 3rd, 2010 to May 31st, 2011 was provided. In earlier data, the building's energy usage was less regular and generally higher, most likely due to inefficiency and fewer regulatory measures.

Due to the large gap in data during the summer of 2010, using an entire year long data set would have only been possible for 2010 predictions and would have required a data set beginning in 2009. Building a model based on this early data would have made it far less useful when applied to more recent data. In order to make more current predictions while still making use of the large data set available, the older data was organized as an additional variable set alongside weather variables. This variable gave the energy value exactly 52 weeks before a given data point. It was important to keep the day of the week consistent between the two years because of the strong weekly cycle, so the "one year ago" values were the same weekday as the "current" values, but not necessarily the same date in the previous year.

2) Weather Data: Weather data was collected from the Central Park weather station because of its accuracy compared to many smaller, amateur stations and its proximity to 345 Park Ave relative to stations at LaGuardia and John F Kennedy Airport. Since Central Park has more plants and greenery than paved areas of Manhattan, it is often a cooler temperature than what might be expected at 345 Park Avenue. This results from the Urban Heat Island Effect, in which paved urban areas absorb and trap more solar radiation compared to surrounding wooded areas or parks, causing these urban areas to heat up more. This effect is not particularly problematic in this application of weather data so long as the same weather station is used for all data because the computer model makes regressions based on relative temperature compared to the corresponding energy demand. Hourly weather data from the Central Park weather station was taken from the Weather Underground website (wunderground.com). On the website, hourly values of temperature, dew point temperature, pressure, wind direction, wind speed, humidity, precipitation, and conditions are provided for more than five years of history data. However, each day of hourly data is given in a separate comma separated value file online, so obtaining large sets of data is not feasible without the use of a computer script. Additionally, the data had many errors, additional points, missing points, and other difficulties, further confirming the need for programming to obtain data.

Using a Matlab script, each day of data was accessed in succession and compiled into a larger comma separated value file containing all the individual days of data. This file was then refined to exclude headings and other unnecessary information. Points with values of "-9999" represent an error and were eliminated, and points with a wind speed of "Calm", representing a wind speed too small for the equipment to detect, were changed to 0 mph. An additional Matlab script then took weighted averages to fill in missing hourly points and eliminated extra points. Only temperature, dew point temperature, pressure, wind speed, and humidity values were kept and added to energy data. Temperature and dew point temperature were ultimately used in the model.

3) Formatting Data: For Support Vector Machine Regression it is important to scale all data to between 0 and 1. This allows each variable to carry equal weight in the creation of the model. The energy, temperature, and dew point temperature data were scaled to be roughly between 0 and 1. Energy demand values in kilowatts were all divided by

10,000, and temperature and dew point temperature values in degrees Fahrenheit were all divided by 100. Some cold winter temperatures were below 0 degrees Fahrenheit, and were kept as small negative numbers in the regression for the sake of consistency.

Support Vector Machine Regression requires two sets of data to make its regression, the training set and the test set. A Matlab code was written to format the data into the two sets for use in SVM regression. The training set contains all of the data available. This set is used, along with the parameters chosen, to build a multi-dimensional model and apply the kernel function. The first column of data in the training set, called the y values, contains a list of energy values beginning with the most recent and going back in time. The task of the SVM regression is to output future y values so the y values must be the same type of data that will be predicted (in this case, energy demand). After the y value, the numbered time delay values are listed. Take, for example, a set that was to contain 48 hours of energy time delays plus 48 hours of temperature time delays to predict the next 24 hours of energy. Row 1 would first begin with the y value, the most recent energy usage value. Following this would be a "1:" denoting the first time delay and the temperature value 24 hours before the y value. There is a 24-hour gap between the y value and the first time delay because 24 hours of data will be predicted. For further explanation on the time gap necessary to predict 24 hours into the future, see the section on the invalid model. The second time delay would be denoted by "2:" and would contain the temperature value 25 hours before the y value. This pattern would be followed for 48 hours of energy data and would make up the first 48 time delays. The 49th time delay would be denoted by "49:" and would contain the first temperature value, the temperature 24 hours before the y value. The 50th would similarly contain the temperature value of 25 hours before the y coordinate. This would be repeated until the 96th time delay. Row 2 would begin with a y value of the second most recent energy value, and the first time delay would be the energy usage 24 hours before this new y value. This would be continued for the entire length of the available set.

Unlike the training set, the test set only contains the most recent data. It is used in conjunction with the model to predict future values. When Support Vector Machine regression is used, the y values predicted by the model are the values that would fit best as y values in the test set. Therefore, the length of the test set is equal to the number of hours that will be predicted by the model, so the example above would have a 24-row test set to predict 24 hours of data. Also, this means that the y values given in the test set do not affect the output. In formatting the test set, a y value of zero was chosen for all the rows. Continuing the example above, the first time delay in row 1 of the test set would be denoted by "1:" and would contain the most recent energy value. The second time delay would contain "2:" and the second most recent energy value. This would be continued for the first 48 hours of energy values and then the first 48 hours of temperature values. The test set would continue for 24 rows, so that the column of values in the 1st time delay are the 24 energy values skipped in the training set time delays.

4) The Invalid Model: This gap in values from the y coordinate to the first time delay is necessary because if the first time delay denoted the energy of just one hour before the y value, the test set would also need to follow this pattern. The last row of data, responsible for predicting just one hour ahead would work because the first time delay is energy usage of one hour before the y value of next hour, so this would just be the most recent energy value. It becomes a problem in the second-to-last row of the test set, in which the y is predicting 2 hours from now. The first time delay of this value is the energy demand of one hour in the future, which is not known. Similarly, three rows from the end would need an energy demand value for two hours in the future. Therefore, without a time gap equal to the hours one wants to predict, the test set can only be one row long, and would only predict the upcoming hour of data. If the test set is any longer, it will need to use data from the future to predict the future, which is not a valid model. This invalid method was tried, and yielded predictably excellent results.

5) The Triangle Cut-off: The way the training and test sets are formatted with time delays causes the sets to be shorter than the amount of data available. Take, for example, a fictitious data set with energy values 21, 22, 23, 24, 25, 26, 27, 28, 29 going back in time from newest to oldest. The test and training sets with three time delays for creating a two-value regression are shown in figure 5. This set has nine values, but can only make a four-row training set plus the two test set rows for prediction. The rows in red at the bottom are incomplete because there is not enough data to fill in all the time delay values, so the bottom of the set forms a triangle, in which each successive row has one less time delay value. The number of lines lost in formatting is equal to the sum of the number of time delays used and the size of the gap (which is the number of values that are being predicted).

B. Model Creation

In order to run SVM regressions on a personal computer, two different software packages can be used to implement the SVM algorithm. Both LIBSVM and SVM^{light} require the same data format and divide the regression into a training function that trains the model and a testing function that is used to validate the model and predict into the future. LIBSVM was chosen because the SVM^{light} training function took much longer to run when computationally expensive hyper parameters were specified. The average running time of LibSVM was approximately four minutes, while SVM^{light} often took over an hour. Since the goal was to predict hourly energy demand, in order for the model to be relevant for real-time predictions, it must run quickly. This might be possible with SVM^{light} on a larger server and so that software package should not be disregarded, but for use on personal computers, LIBSVM was the clear choice software package.

Test Set:

+0	1:21	2:22	3:23
+0	1:22	2:23	3:24
Training Set:			
+21	1:24	2:25	3:26
+22	1:25	2:26	3:27
+23	1:26	2:27	3:28
+24	1:27	2:28	3:29
+25	1:28	2:29	
+26	1:29		**3=time delays**
+27			
+28			
+29			**2=gap**

Fig. 5. Test set.

To apply the Support Vector Machine to the training data, a variety of parameters must be selected. The kernel function, C parameter and γgarameter specified in the training phase determine the performance of the model prediction. To create the most effective model, the parameters selected created the best "goodness-of-fit" and did not make the model too computationally expensive to run on a personal computer [4]. In order to evaluate the accuracy of the predictive capability of the model, Root Mean Square Error and R-squared statistical metrics were used.

Root Mean Square Error (RMSE) measures the error between predicted and actual values [7]. It is calculated by taking the sum of the squared differences between the actual and predicted values, dividing by the number of observations, and taking the square root. A lower value of RMSE (Root Mean Square Error) indicates smaller error. The RMSE equation is defined below:

$$RMSE = [\sum(Y^2 - Z^2)/n]^{1/2}$$

Where
Y = measured values
Z = predicted values
n = number of observations

R-squared, or the "coefficient of determination," is also used to measure the error between predicted and measured values. To calculate R-squared, one subtracts from 1 the sum of the squared distances between actual and predicted values divided by the sum of the squared distances between the actual values and their mean. The values range between 0 and 1, with those closer to 1 indicating a more accurate prediction. The R-squared equation is defined below:

$$R\text{-Squared} = 1 - \sum(Y\text{-}Z)^2/\sum(Y\text{-}Yavg^{)})^2$$

Where
Y = measured values
Z = predicted values
Yavg = average Y value

1) Kernel selection: The first step in the development of the appropriate Support Vector Machine model for predicting the energy demand of 345 Park Avenue, the appropriate kernel function was selected. Most literature and research that employed Support Vector Machine algorithms to predict energy demand and temperature employed the Gaussian function, which is the function included in the RBF kernel. As Dong *et al.* notes, "The RBF kernel nonlinearly maps samples into a higher dimensional space, and unlike the linear kernel, can handle the case when the relation between class labels and attributes is non-linear" [4]. The non-linear, dynamical nature of the influence of weather on energy demand in heating, ventilation, and air conditioning systems excludes the possibility of using a linear kernel. The polynomial kernel could be a possibility, however, it has many more hyper-parameters (which impacts the complexity of the model) than the RBF kernel, meaning the RBF kernel has less "numerical difficulties" than the polynomial kernel and has less of a tendency to produce values approaching zero and infinity. The polynomial kernel creates a less restrained curve.

2) Variable Selection: In order to determine which combination of weather and energy variables would create the most accurate model, regressions were performed for weeks in both February, to represent a winter prediction, and May to represent a late spring regression. Two different years of energy data and multiple years of weather data were available. The weather variables were collected from wunderground.com. The section on data retrieval, refinement, and formatting outlines the process for acquiring weather data from wunderground.com. Temperature, humidity, dew point, sea surface pressure, wind speed, precipitation, and solar insolation hourly data are provided.

Sea surface pressure was eliminated as a possible weather variable for the model due to the low impact of these variables on energy use in heating, ventilation, and air conditioning. Solar insolation was also quickly eliminated as a viable option for the model due to (as most literature indicates), variable cloud cover and the fact that in urban areas shadows are frequently cast across buildings, the actual amount of solar insolation that reaches the building would require a complex set of calculations to include. Furthermore, the low surface area to volume ratio makes insolation a less important variable for large commercial buildings such as 345 Park Avenue [16].

While precipitation may have a large impact on building temperature due to latent heat, the precipitation data available had many missing values, marked "null", and it was determined that replacing these values with averages or zero values

would drastically impact the accuracy of the model. Relative humidity and dew point temperature can be derived from each other and therefore including both would introduce redundancy in the model. Relative humidity can be defined as the ratio water pressure in an air/water mixture. Relative humidity, unlike absolute humidity (which is simply a measurement of water content in air) changes with pressure and temperature. Dew point temperature "is the temperature to which a given parcel of humid air must be cooled, at constant barometric pressure, for water vapor to condense into water" [8]. Both relative humidity and dew point take into account temperature, moisture content in air, and pressure. Dew point temperature was chosen for our model.

Therefore, the 5 variables available to create the model were 2 separate years of energy, temperature, dew point temperature, and wind speed. These variables were then ranked in terms of importance to the predictive model based on the variables that Rudin uses to determine their heating and cooling loads. Past energy demand and temperature index (a combination of temperature and humidity) are used by Rudin, so the variables were we prioritized in the order of 1^{st} year of energy, 2^{nd} year of energy, temperature, dew point temperature, and wind speed.

In order to evaluate which variable combinations produce the best model, the LIBSVM-3.1 software was used to train and test data. Training data was used starting September 4^{th} 2010 and ending February 19^{th} 2011 for the February regression and starting September 4^{th} 2010 and ending May 2^{nd} 2011. The following week was used as test data. Default parameters were used for each regression, since those values had not yet been specified. LIBSVM-3.1 produces an R-square and RMSE value after running the testing function. These values can be used to validate the accuracy of the model if the class labels (the prediction variable values for SVMR) in the test set are the true values. The output is not a prediction but rather a comparison of the model's ability to predict with the actual testing values. The methods section on the prediction cases outlines how replacing the class labels with random values creates the predictive capability of the model. However, in the case of validation of the model variables and parameters, the true values are used as class labels.

3) Parameter Selection: In order to select the C value, γ value, and number of time delays to construct the most accurate and efficient model, a step-wise search method was used. The step-wise method works by running regressions using values of different orders of magnitude for a specific parameter, calculating the MSE and R-square value to assess accuracy, then evaluating on finer scales until the appropriate value is established. The same method listed above for variable selection is used to selected C, γ and time delay values, where the test file uses real values as classifiers in order to compare the model's accuracy at predicting for those values. First the C value was evaluated at 1, 10, 100, and 500, and then it was evaluated at 200, 300, 400. γ was evaluated at 1, 0.1, 0.01, 0.001, and 0.0001. The number of time delays was evaluated at 24, 48, 96, 144, 192, 240, 288, and 336. These are all multiples of 24 because of the importance of ensuring that the daily cyclicity is not interrupted.

C. Regression

LIBSVM was invoked in a remote shell on CCLS's computation servers. The svm-train function was used to train the model with the given training set and parameters with the format:

./svm-train -s 4 -t 2 -g .1 -c 200 TrainFile.txt ModelFile.txt

for a gamma value of 0.1 and a c value of 200. The svm-predict function was then applied to the model with a given test set to output a prediction. This function was used in the format:

. /svm-predict TestFile.txt ModelFile.txt OutputFile.txt.

III. Results

A. Model Creation

1) Variable Selection: Following are the results for variable selection:

Figure 6 displays the R-square values for six trial models using six different variable combinations.

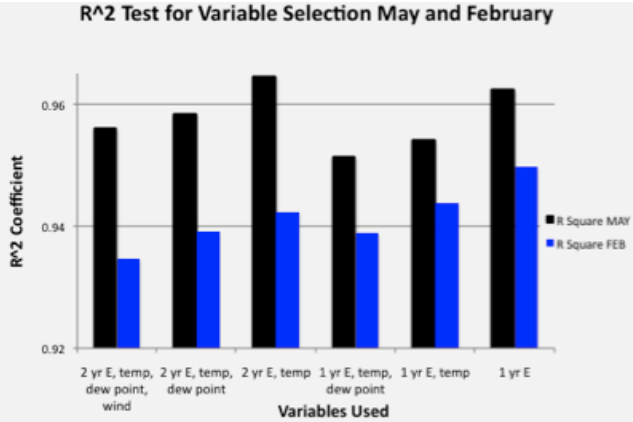

Fig. 6. R-Square coefficient versus variables.

Figure 7 displays the Root Mean Square Error values for six trial models using six different variable combinations.

2) Parameter Optimization: Following are the results for parameter optimization.

Figure 8 displays the R-square values for the optimization of C values.

Figure 9 displays the MSE values for the optimization of C values.

Figure 10 displays the R-square values for the optimization of [gamma] values

Figure 11 displays the MSE values for the optimization of [gamma] values.

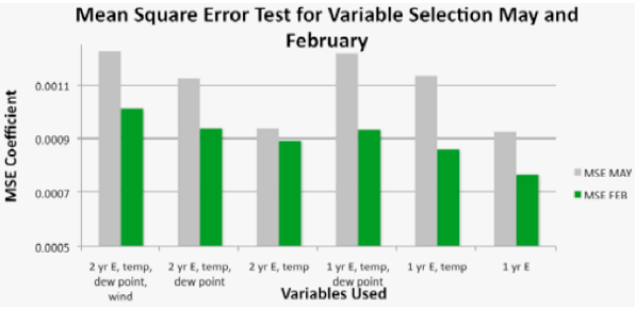

Fig. 7. MSE coefficient versus variables.

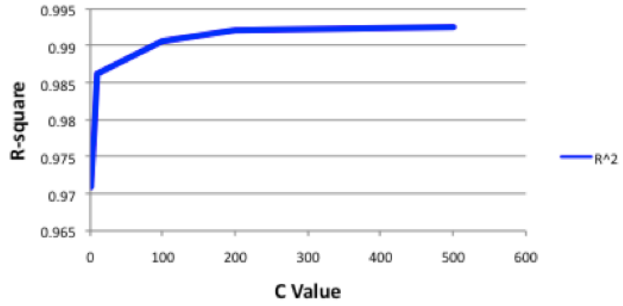

Fig. 8. R-Square versus C value

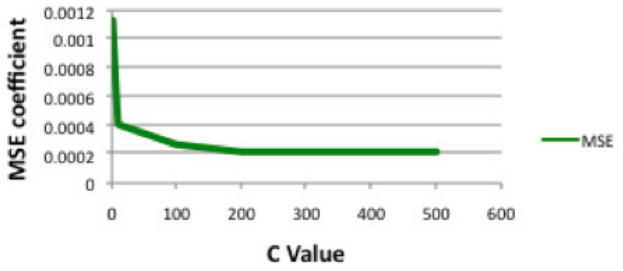

Fig. 9. MSE coefficient versus C value

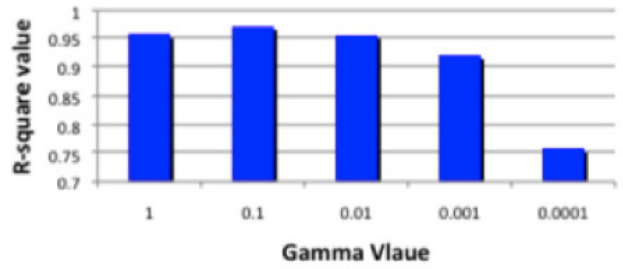

Fig. 10. R-Square value versus Gamma value

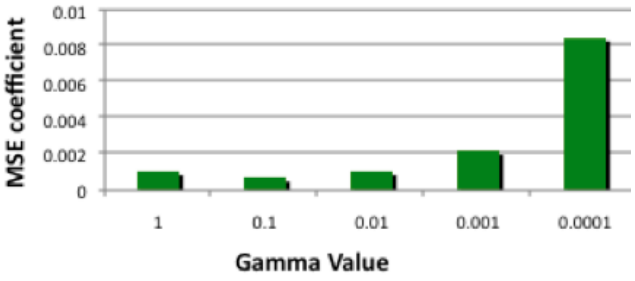

Fig. 11. MSE coefficient versus Gamma value

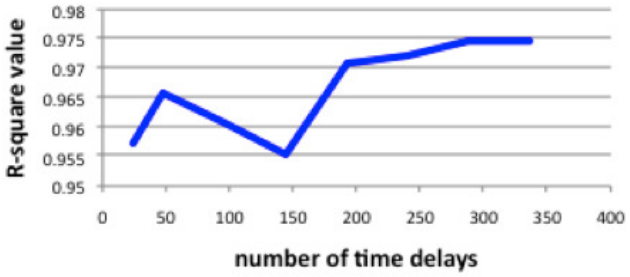

Fig. 12. R-Square value versus number of time delays

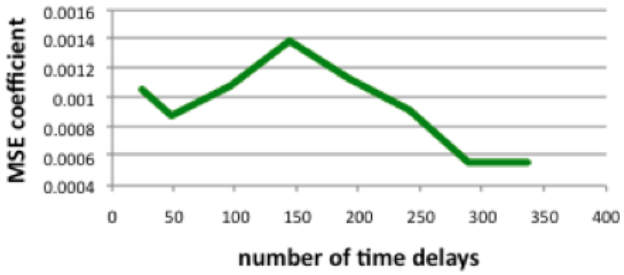

Fig. 13. MSE coefficient versus number of time delays

Figure 12 displays the R-square values for the optimization of number of time delays.

Figure 13 displays the MSE values for the optimization of time delays.

3) Regression: Following are the results of several regressions:

IV. Discussion

A. Model Creation

Based on the results of the R-squared and Mean Squared Error statistical tests, the best combination of variables for a February regression would be to use 1 year of energy. For May, the best combination of variables would be 2 years of energy and temperature. While these statistical tests prove the accuracy of these models, 2 years of energy, temperature, and humidity was instead used for all regressions. The reason for this is that response of heating, ventilation, and air conditioning to weather is very dynamic. In studying the physical HVAC plant at 345 Park Avenue, the operations management indicated that they employ the next day's heat

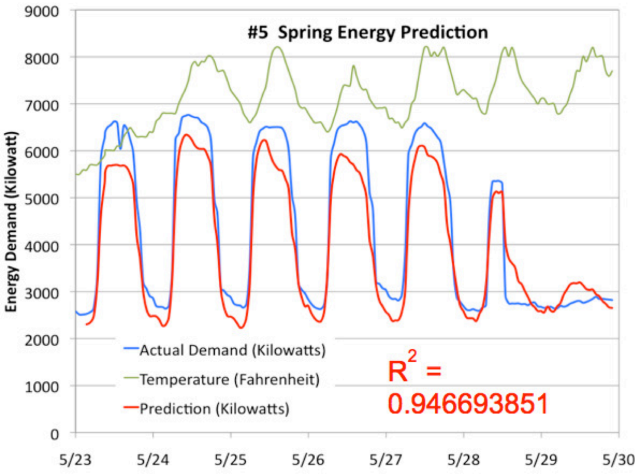

Fig. 14. Actual versus predicted energy demand versus time in spring

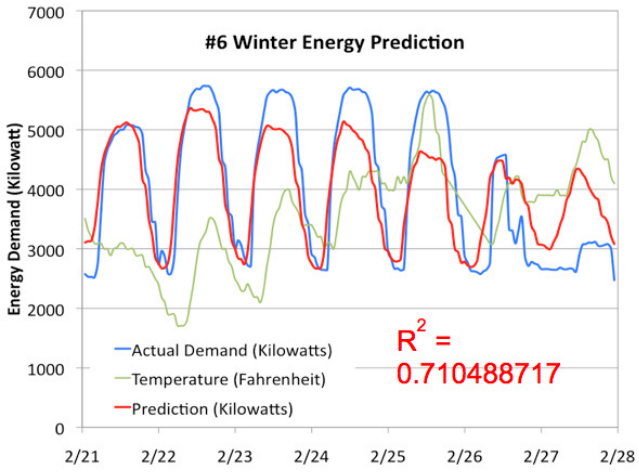

Fig. 15. Actual versus predicted energy demand versus time in winter

index in order to determine their heating and cooling load for the day. Temperature index is a combination of temperature and humidity that attempts to estimate what the temperature feels like to humans. Therefore, it was decided that including those variables in the creation of the model was important. A model using fewer variables produce smooth, highly cyclical curves, while the addition of more variables creates curves with more noise and statistically poorer fits. However, the inclusion of more variables allows the model to adapt more dynamically to changes in weather that occur within a single day or week, and it aides the model in predicting minima and maxima energy demand values.

In selecting the appropriate C and γ values and the number of time delays, it is important to create both a highly accurate model and one that is not too computationally expensive. In selecting a parameter, changing the value in one direction to make the model more accurate often simultaneously makes it more computationally expensive. Eventually, an asymptote of accuracy will be achieved, in which changing the value in the direction of better results does not significantly affect the accuracy of the model. Our goal is to choose a value near the beginning of the asymptote in order to optimize the trade-off between accuracy and computational expense. A C value of 200, gγgalue of 0.1, and 196 time delays were chosen. Following this section are the results of the regressions for different weeks throughout the year.

It was hypothesized that the reason why the results of R-square and MSE are much lower for the winter regression is that the model does not understand how to read low energy values as an indicator of higher energy demand.

B. Regression

Figures 14 and 15 show regressions for two different 5-month data sets at different times of the year. The spring graph is closer to the actual energy consumption of the building, with an R-squared value of about 0.95, while the winter graph is less accurate, with an R-squared of about 0.71. It was hypothesized, as above, that the reason for the less accurate winter regression is that the SVMR model has trouble handing low winter temperature values as needing increasing energy corresponding to higher temperatures.

V. Conclusion

A model was created that accurately predicts the energy demand of 345 Park Avenue. The R-squared values for the final regressions indicate that the ability of the model to predict energy demand is high. Even the lower R-squared value for winter regression provides a good estimate of how the building will behave in the future.

A variety of further research could be pursued to improve the accuracy of this model. First, in order to correct the problem of consistently low predictions for winter regressions, the data could be normalized such that low temperature values indicated higher energy demand, possibly by increasing the temperature values using some devised metric. Eliminating weekends from the training and test data could also be considered, so that the model only produces regressions for the business week. We would also like to consider making models for specific weekdays. This would reduce some of the complications created by the weekly cyclicity, which may account for low weekday energy demand predictions. Finally, access to more years of data without missing months would greatly improve the model and its predictive capacity.

This model provides a good example of the capabilities of Support Vector Machine Regression for modeling energy demand in commercial buildings. It could potentially be applied to many of Rudin's properties in order to inform operations management at each building of how to most efficiently reduce energy consumption for Rudin Management as a whole.

Acknowledgment

The authors would like to thank Rudin Management Company of New York, Columbia University Lamont-Doherty Earth Observatory, GE Ecomagination and FedEx for their support of this research.

REFERENCES

[1] Arciniegas, IE. And Marathe, A. Important Variables in Explaining Real-Time Peak Price in the Independent Power Market of Ontario', *Utilities Policy*, vol. 13: 27-39, 2005.

[2] Bauer, M. Scartezzini, J-L. A Simplified Correlation Method Accounting for Heating and Cooling Loads in Energy-Efficient Buildings.

[3] Burges, C. A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, vol. 2: 121-167, 1998.

[4] Dong, B. Cao, C. Lee, SE. Applying Support Vector Machines to Predict Building Energy Consumption in Tropical Region, *Energy and Buildings*, vol 37, 2005.

[5] Farmer, JD. and Sidorowich, JJ. Predicting Chaotic Time Series, *Physical Review Letters*, vol 59, 1987.

[6] Hour, Z., Lian, Z. "An Application of Support Vector Machines in Cooling Load Prediction", *School of Mechanical and Electrical Engineering, Shenzhen Polytechnic* and *Institute of Refrigeration and Cryogenics Shanhai Jiao Tong University*, 2009.

[7] Mean Sqaure Error, http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_mean_square_error.htmz.

[8] Dew Point, http://en.wikipedia.org/wiki/Dew_point

[9] Hsu, C-W. Chang, C-C. Lin, C-J. A practical Guide to Support Vector Classification, Department of Computer Science, National Taiwan University, Taipei 106, Taiwan, 2006.

[10] LibSVM "README" tutorial.

[11] Li, Q. Meng, Q. Cai, J. Yoshino, H. Mochida, A. "Predicting Hourly Cooling Load in the Building: A Comparison of Support Vector Machine and Different Artificial Neural Networks", *Energy Conversion and Management,* vol 5, 2009.

[12] Li, X., Lu, J.-H., Ding, L., Xu, G. and Li, J. Building Cooling Load Forecasting Model Based on LS-SVM, Asia Pacific Conference on Information Processing, 2009.

[13] Miller, J N. Outliers in Experimental Data and Their Threat', *Analyst*, vol. 118, May 1993.

[14] New York State Energy Planning Board. *New York State Energy Plan and Final Environmental Impact Statement*. (November 1998): 3-23, 1998.

[15] PLANYC, http://needigest.com/?p=627#more-627.

[16] Radhika, Y. and Shashi, M. "Atmospheric Temperature Prediction Using Support Vector Machines", *International Journal of Computer Theory and Engineering*, vol 1, no 1, April 2009.

[17] Shen, K-Q. Ong, G-J. Li, X-P. "Feature Selection Via Sensitivity Analysis of SVM Probabilistic Outputs", *Machine Learning*, vol. 70: 1-20, 2008.

[18] Simulation Research Group, Lawrence Berkeley National Laboratory, University of California, Overview of DOE-2.2, June 1998.

[19] Ward, J K. Wall, J. West, S. de Dear, R. "Beyond Comfort – Managing the Impact of HVAC Control on the Outside World", Proceedings of Conference*: Air Conditioning and the Low Carbon Cooling Challenge*, Cumberland Lodge, Windsor, UK, July 2008.

[20] Xi, X.-C., Poo, A.-N. and Chou, S.-K.. Support Vector Regression Model Predictive Control on a HVAC Plant, Control Engineering Practice, vol 15, 2007.