Solar Power Prediction for Smart Community Microgrid

Wellington Cabrera
Dept. of Computer Science
University of Houston
Houston, TX, USA

Driss Benhaddou
Engineering Technology
University of Houston
Houston, TX, USA

Carlos Ordonez
Dept. of Computer Science
University of Houston
Houston, TX, USA

***Abstract*— Urban areas host more than 50% of the world's populations, are responsible for 75% of energy consumption in the world, and they emit almost 80% of global carbon dioxide. There is an urgent need to develop "low carbon" cities that are smart and efficient and use renewable energy to foster the growth of the green economy. Smart grids are being developed to tackle these challenges through integration of renewable and green energy as well as energy efficiency. They are moving toward a concept of networked microgrids. Microgrids will enable the integration of distributed renewable energy such as roof top solar panels within smart city communities. For these microgrids to operate reliably and efficiently, prediction algorithms are important because of the fluctuation of solar energy and its dependence on weather. Prediction of energy is a component of microgrids energy management systems to optimize their operation. This paper presents a machine learning based algorithm, which learns a regression tree model with time of the day and humidity as main parameters. The regression tree model presents a promising accuracy. This work shows that solar panel prediction in Houston is heavily dependent on humidity of the region.**

Keywords- Prediction algorithms, regression Trees, Microgrid, Smart Grid, Smart Energy Management.

I. INTRODUCTION

Urban areas host more than 50% of the world populations, are responsible for 75% of energy consumption in the world, and they emit almost 80% of global carbon dioxide. There is an urgent need to develop "low carbon" cities that are smart and efficient and use renewable energy to foster the growth of the green economy. Smart grids are being developed to integrate renewable energy and improve the grid efficiency and reliability. One of the most important advancement in smart grid concept is microgrid. Microgrids have been developed as a mean to integrate green and renewable energy such as roof top solar panels, micro-turbines (MT) and Combined Heat and Power in campuses and Communities [1] and provide reliable power with economic, environmental and technical benefits.

Microgrids provide the ability to take energy from the grid as well as feed power directly to the grid through low voltage (LV) networks, thereby allowing the customer to become an active participant in the grid [2]. They should also be able to work in emergency state and be disconnected from the grid, called islanded mode. In disconnected state microgrids should be able to generate energy to power the loads; in addition, it should be able to adjust load demand in real time to avoid breakdown of the microgrid. To be able to operate microgrid in stable and efficient ways, it is very important to develop accurate solar prediction algorithms.

The concept of microgrid can be applied at the community level where consumers within a subdivision can implement solar energy in their roof tops and aggregate their energy in the community microgrid and form retail electricity providers (REP). For example, a university campus is a typical community where buildings implement solar energy in the roof top to produce local energy while interacting with other buildings to optimize energy generation and consumption in the campus. A building in the campus constitutes a building block of this microgrid and actively participates in the operation of the campus. One of the important task within the scope of managing a campus (a microgrid) is to implement prediction algorithms that can give accurate account of the energy produced. This information is coupled with consumption to make sure the microgrid is operational in efficient manner (see figure 1).

Prediction of renewable energy production plays a key role in microgrid management and control. Prediction algorithms are implemented part of the Energy Information System (EIS) of microgrids. In smart microgrid, EIS is coupled with power system to deliver a smart system that can provide energy in efficient manner. Energy information system plays, therefore, a key role in managing the resources within the microgrid and can be thought as a layer on the top of the power layer. EIS has the objective of making sure the microgrid is stable, reliable, and resilient (can work in normal or islanded mode). EIS has also the capability of interacting with the smart grid market as well as other nearby microgrids.

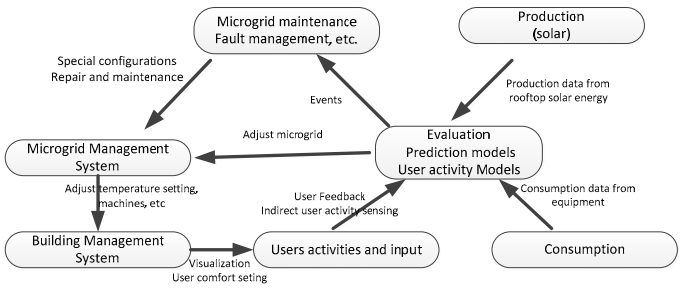


Figure 1: Architecture of the proposed EIS system.

Figure 1 depicts different components of our microgrid EIS system. Evaluation and prediction component collect data from local energy production (e.g. solar panels), consumption sensors/meters, and user activities. The data collected is used to

run solar energy prediction, user preference and activity models, besides to consumption profile models. The data is collected by the EIS middleware and is archived for visualization, aid in control and human decision making, as well as for future use for other developments. The data is accessed by a number of applications including user activity modeling, thermal and air flow setting, HVAC control, event generation for maintenance and failure detection. The evaluation and prediction component generate the information needed by the microgrid management. Microgrid management adjusts building operations through Building management system component.

Given the regulatory and geographical constraints, solar energy will play a significant role in energy production. This paper show how to apply regression tree algorithm for solar power prediction. It uses historical data collected from a solar energy plant at the University of Houston. The main contributions of the paper are as follow:

- Show that regression trees can be applied to solar power prediction, with promising accuracy.
- Use real data from solar roof top at the Central plant building of the University of Houston campus [12].
- Prediction algorithm in Houston is dependent on humidity and time of the year, which is unique characteristics of Houston area.

The paper is organized as follows: section 2 provides the literature review of different prediction algorithms, and section 3 describes the proposed algorithm. Analyses of the data collected and the results of the algorithm are presented in section 4. The paper is concluded in section 5.

II. LITERATURE REVIEW AND BACKGROUND

Solar energy forecasting can be categorized from time horizon point of view into hour-ahead, day–ahead or intra-hour types of forecasts. Considering the geographical scope, the solar forecasting can cover a large area (a big solar plant) or a collection of small locations. From the prediction methodology point of view, forecasting can be broadly characterized as *physical* or *statistical*. Physical approaches model physical atmospheric phenomenon as part of the solar irradiance prediction using Numerical Weather Prediction (NWP) methods or sky imagery. On the other hand, statistical approaches rely on an historical database of solar power generation, and the respective weather conditions. We review statistical approaches as the contribution of this paper is in the statistical approach methods.

Statistical approaches use a training dataset that contains PV power generation data, as well as various inputs or potential inputs, such as NWP outputs (Humidity, temperature or other), weather station or satellite data. This dataset is used to train models – such as autoregressive or machine learning models – that output a forecast of PV power at a given time based on past inputs available at the time when the model is run.

Persistence method is a simplest method that predicts the output based on the last data considering the variation of the sun angle [3]. Variation on cloud condition impacts strongly the accuracy of this method. It is useful only in very short term predictions. Autoregressive models use historical data to develop prediction models. P. Bacher *et al.* [4]developed Recursive Least Squares (RLS) Autoregressive models using 21 locations in Denmark. R. Huang *et al.* [5] based their PV forecast on the Autoregressive Moving Average model (ARMA), using data that corresponds to the UCLA zone, California.

Support Vector Machines (SVM) is another technique used for solar energy prediction. N. Sharma et al [6] used the SVM model and tested three types of kernels: Linear kernel, Polynomial kernel, and Radial Basis Function kernel (RBF). RBF shows better results compared to others. Likewise, J. Shi et al [7] studied Power generation prediction in a China location with SVM, using a RBF kernel.

Neural Networks technique was also explored by researchers for PV prediction. C. Cheng et al [8] present a predictive model for PV using Radial Basis Function networks, a kind of artificial neural networks. This 24 hours ahead prediction model is trained and tested with data from a China location in 2007. On the other hand, A. Mellit [9] applied a model based on Multi-Layer Perceptron to estimate the solar irradiance in Trento, Italy. Classification and Regression Trees is another statistical technique for energy production that was used mainly for wind energy prediction. Classification and Regression Trees (CART) is described in the seminal work of Breiman et al [10]. Even though Regression Trees has been applied in wind farm power [11] generation, it has not been applied and tested on solar energy prediction. This paper develops a classification and regression Trees for solar energy. The following table summarize different techniques and their characteristics.

TABLE 1: Forecasting Services/Tools

Provider/Tool	Time horizon	Method
Green Power Research	medium term	Geostationary satellite imaginary
AWS TruePower	Short – long	NWP
SolarAnyWhere	Short	Satellite imagery
	Long	NWP
SolarCasters	Hour/day ahead	NWP
SOLARFOR	0 – 48 hour	NWP

III. APPLYING REGRESSION TREES TO SOLAR POWER PREDICTION

Binary tree prediction algorithm is part of the family of "Classification and Regression Trees". Regression Trees are an alternative to linear regression in case the output presents a dissimilar behavior through the input space. As a splitting criteria, we can choose either the one that minimizes the sum of the squared differences between the response values for the data points in the current node and their sample mean, or the greatest reduction in the sum of the absolute differences between the response values for the data points in the current node and their sample median.

Regression Trees is a machine learning algorithm used to construct predictive models. As any Machine Learning algorithm, it has to be trained with a dataset (aka training set). In our problem, the training set consist on weather and power historical data. In contrast to linear regression, regression trees do not give us a unique global model; regression trees consist on multiple models, where each of them is a simple model, for instance a constant.

Although we use all the variables available both in the historical weather as in the historical power dataset, our results show that the most important variables to predict power are humidity, time of the day and sky condition (clear, overcast, etc.). In contrast to linear regression, regression trees do not provide a unique global model; regression trees consist on multiple models, where each of them is a simple model, for instance a constant. The fundamental step to build a regression tree is a binary splitting of the data in a recursive manner. The challenge is to get the best splitting of the learning samples that are present in a particular node. A recursive implementation of the algorithm is depicted in the following table. The input is a preprocessed data set X consisting of historical weather, historical PV generation, and current weather forecast, as described in section IV.

Algorithm 1: recPartition (recursive Partitioning)

Input: X
Output: A tree structure with partition values

1 **function** recPartition (X)
2 **for** each feature F in X do
3 $V_F = \underset{V_F}{\operatorname{argmin}}(\sum_{x<V_F}(x-\bar{x}) + \sum_{x\geq V_F}(x-\bar{x}))$
4 select the feature F and partition value V_F leading to the overall minimum
5 **if** stop criteria is met
6 **return** (F, V_F, Null, Null)
7 **else**
8 **return** (F, V_F, recPartition (Points on Left of V_F), recPartition (Points on Right of V_F)
9 **End**

IV. Data Collection and Analys

Historical Data for Model Learning:
Weather: The National Oceanic and Atmospheric Administration (NOAA) keeps a data repository of historical weather. Historical weather records are available for weather stations across the United States. To predict the solar power generated at University of Houston, we use the data for the closest weather station: Hobby Airport, Houston, TX. The available data includes several weather variables, for instance temperature, humidity, pressure, wind speed and wind direction. NOAA records weather observations in periods of 1 hour. The data is available as CSV files for monthly periods. We download data for a whole year.
Power: Our solar panels record the power level in a log every 5 minutes. The power data is retrieved and stored in a database, along with the weather data. Like the weather historical data, we consider the power data for one year. This dataset is comprised of about 100 thousand observations. The amount of energy is the dependent variable. The power production from photovoltaic panel is clearly seasonal, as shown in Figure 2. Therefore we analyze one year of data to consider the effects of every season.

Preprocessing.
In the preprocessing step we prepare the data to the analysis. String variables are converted to ordinal data, and redundant fields are removed. For instance, the original data presents a field called Sky Condition, which is a string of characters. The preprocessing step converts this field to an ordinal data (0=Clear, 2= Few, 4= Scattered, 6= Broken, 8=Overcast). The ordinal number corresponds to the fraction of the sky covered by clouds, in eights. Likewise, since temperature values are provided both in Fahrenheit and Celsius degrees, values in Celsius degrees are removed to avoid redundancy. A set of 12 redundant variables are removed, from the weather data set. The energy production data is logged every 5 minutes, but the weather data is recorded every hour. We assign the energy production measurement to the closest hour, for instance, the measurements for 4:35PM and 5:25PM are assigned to the 5:00PM record in the data set. Data collection and pre-processing was performed with custom programs

Exploratory Analysis.
In this section we analyze the relationship between energy and several weather variables, for the complete year 2014. Figure 2 shows that, in the Houston geographical area, the photo voltaic generation reaches peaks in the first and in the last third of the year. The middle of the year shows lower energy production than the colder part of the year. The colder seasons also present an energy generation with higher variance

Figure 3 shows a plot of the energy vs humidity. The data corresponds to the noon time for the whole year. The plot shows a linear, negative correlation between energy and humidity. We present several box plots, in order to understand important

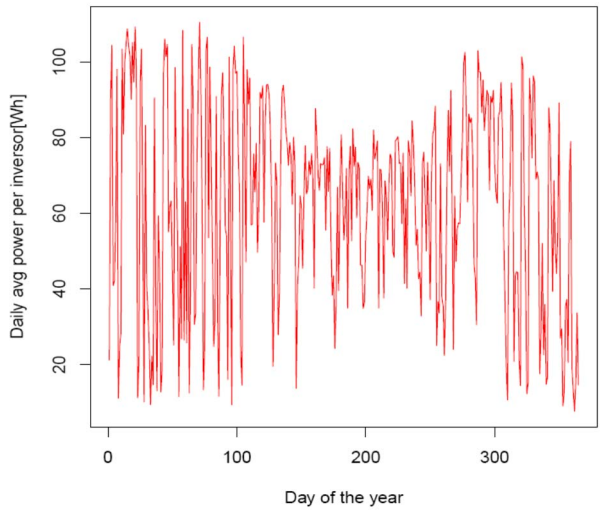


Figure 2: Daily avg. power production per inverter

variables in the power prediction. The bottom of the box represents the first quartile, and the top of the box represents the third quartile. The mean is represented by a bold line. The upper and lower whisker represents the maximum and minimum, respectively. In figure 4, we present a box plot of relative humidity for every month of the year, considering only days with no more that 20% of cloudiness. The months in winter/spring present the lowest relative humidity. The summer

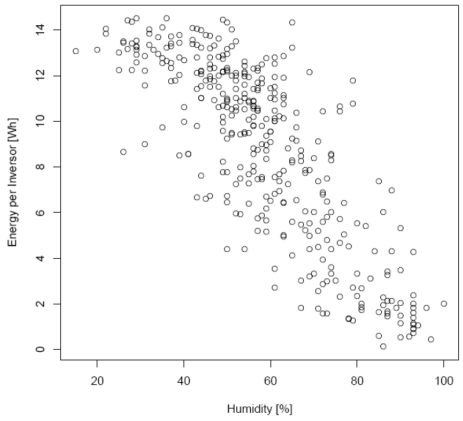


Figure 3: Energy vs. humidity

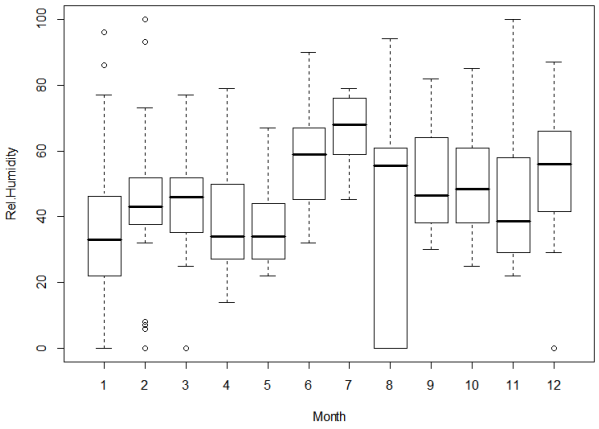


Figure 4: Relative humidity per month

season presents the higher humidity. The humidity during fall months are in the middle of the other seasons. In figure 5, each boxplot let us visualize the variability of the energy generated through the year 2014, grouped by hour of the day. The plot shows clearly that noon time has the best solar power production. We analyzed the power production for three months of the year: February, June and October. Figure 6 shows box plots for the energy produced in every day of the three aforementioned months. In contrast, the box plots in figure 7 shows the energy produced only in sunny days during the same moths. Comparing these two figures, we observe: a) The hourly mean power depends clearly on the time of the days, with a shape similar to a sinusoid, with spike at noon. b) For sunny days, the power production is clearly superior in February.

Storing data in an array database.

Relational databases store the data in the row-column layout. In contrast, an array database stores the data as a set of cell organized as a multi-dimensional array. Each cell contains *p* attributes, a one dependent variable (energy). We store the historical data as a bidimensional array

- Dimension 1: Day of the year
- Dimension 2: Weather variables and Power (dependent variable)

Learning the model

We learn the model based on the preprocessed dataset comprised of hourly solar power observations, along with time and 12 weather variables. We load the data set from SciDB to R. Data is retrieved via http. Data is loaded in R as a dataframe. We noted that splitting the data in two sub data sets improves the results accuracy. Others pre-splitting schemas did not present good results.

- Observations when hr < 13
- Observations when hr >= 13

Therefore, our aim is two obtain two regression trees. The main steps in the process are;

1. Express all the times in CST time.
2. Split the dataset in two: Data before noon, and data afternoon.
3. Apply regression tree algorithm over each dataset and take the two resulting models.

RESULTING MODELS

Model	Splits	Rel. Error	Cross Val. Error
A: hr < 13	9	0.141	0.156
B: hr>=13	10	0.214	0.231

Figure 5 shows the comparison of the predictive model results and the actual solar power generated. The predictive algorithm

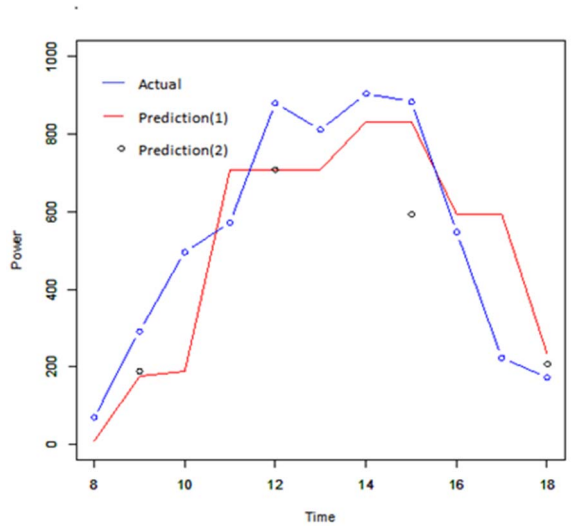
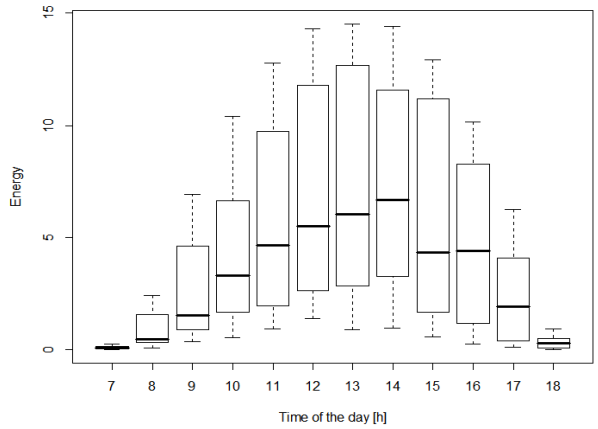


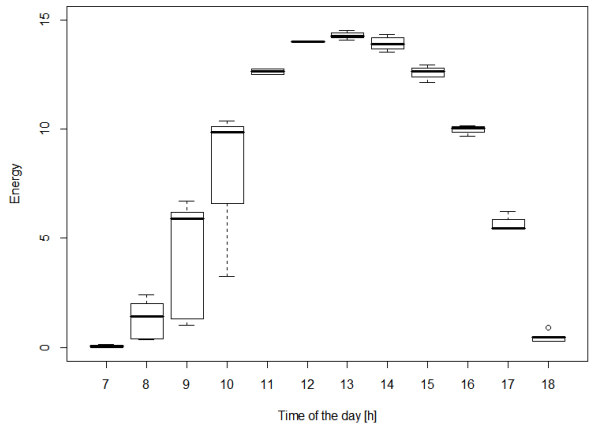
Figure 5: Comparison Actual / Prediction

Hourly energy – any day in the month

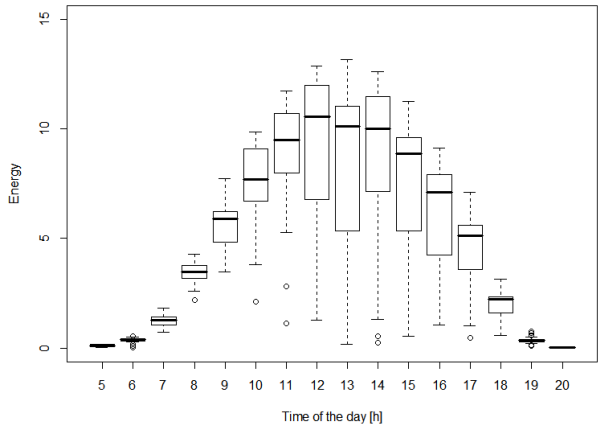
Hourly energy, only sunny days in the month

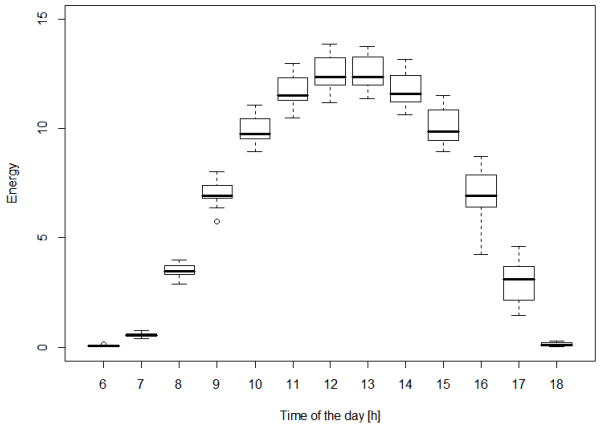
FEBRUARY



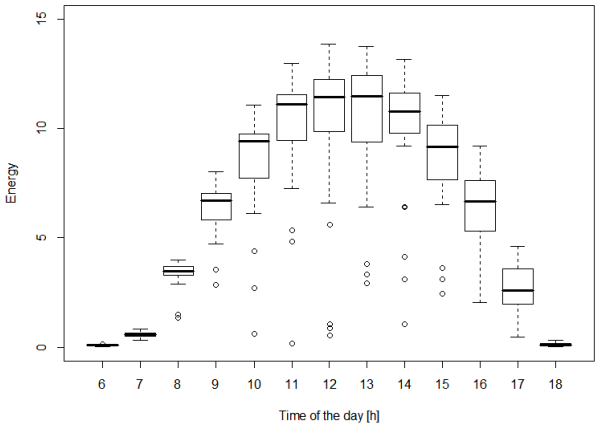


JUNE





OCTOBER



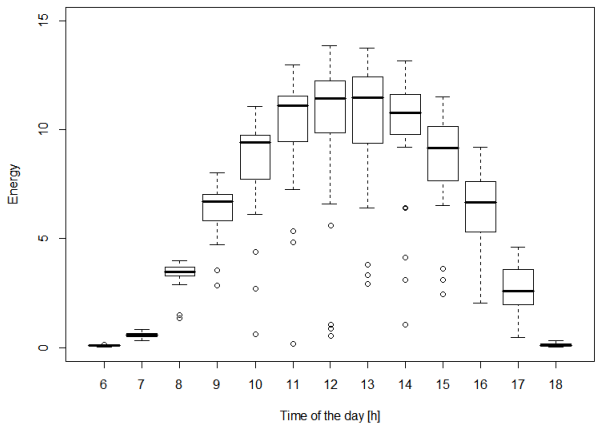


Figure 6: Box plots for hourly solar generation in sunny days,
for February, June and October

was very close to the real power generated with 85% accuracy in the morning and 77% accuracy in the afternoon (see resulting model table above). The model takes just 3 seconds to be computed.

V. CONCLUSIONS

Prediction algorithms play a key role in managing microgrids in both connected and islanded modes. They will enable energy management systems to use predicted information to optimize microgrid operation. In this paper we have applied machine learning techniques to learn a predictive model of solar power generation based on historical weather and power. Although exploratory analysis has been used to understand the data set and its characteristics, regression tree algorithm was able to identify time of the day and humidity as main parameters of the predictive model. Exploratory analysis shows that peak solar power production occurs in winter and late fall, which seems counterintuitive. This observation can be explained because the Houston Area has high humidity during summer; research in PV generation has shown that the power is inversely proportional to the humidity. The results are 77% and 85% accurate. Different regional weather characteristics make unfeasible to apply the same model across diverse regions. Even though, machine learning algorithms are able to learn predictive models that can explain with enough accuracy the solar power generation. Besides, our results suggest that regression tress could be used to identify which are the most important weather variables in a specific region. As future work, we are interested on comparing the accuracy of regression tress to support vector machines and artificial neural networks.

ACKNOWLEDGMENT

The authors would like to acknowledge Mr. Michael Burriello from the University of Houston Central plant for provide access to the data of the solar panel roof tops at the University of Houston.

REFERENCES

[1] M. Wissner, "The Smart Grid – A successful of secrets.," . *Applied Energy,* vol. 88, p. 2509–2518, 2011.

[2] J. A. P. Lopes, C. L. Moreira and A. G. Madureira, "Defining control strategies for microgrids islanded operation.," *IEEE Transactions on Power Systems,* vol. 21, pp. 916-924, 2006.

[3] A. M. Foley, P. G. Leahy, A. Marvuglia and E. J. McKeogh, "Current methods and advances in forecasting of wind power generation," *Renewable Energy,* pp. 1-8, 2012.

[4] P. Bacher , H. Madsen and H. A. Nielsen, "Online short-term solar power forecasting.," *Solar Energy 83.10,* pp. 1772-1783., 2009.

[5] R. Huang, "Solar Generation Prediction using the ARMA Model in a Laboratory-level Micro-grid.," in *IEEE Smart Grid Communications (SmartGridComm)*, 2012.

[6] S. Navin, P. Sharma, D. Irwin and S. Prashant, "Predicting solar generation from weather forecasts using machine learning," in *EE International Conference on Smart Grid Communications (SmartGridComm)*, 2011.

[7] J. Shi, W.-J. Lee, Y. L. Liu, Y. Y. Yang and P. Wang, "Forecasting power output of photovoltaic systems based on weather classification and support vector machines." *IEEE Transactions on Industry Applications 48, no. 3,* pp. 1064-1069, 2012.

[8] C. Chen, S. Duan, T. Cai and B. Liu, "Online 24-h solar power forecasting based on weather type classification using artificial neural network," *Solar Energy,* vol. 85, pp. 2856 - 2870, 2011.

[9] A. Mellit and A. M. Pavan, "A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy," *Solar Enery,* vol. 84, pp. 807-821, 2010.

[10] L. Breiman, J. Friedman, C. Stone and R. Olshen, Classification and regression trees, CRC Press, 1984.

[11] A. Kusiak, H. Zheng and Z. Song, "Wind farm power prediction: a data-mining approach," *Wind Enery,* vol. 12, pp. 275-293, 2009.

[12] https://enlighten.enphaseenergy.com/pv/public_systems/RhzL65901/overview