Towards Statistical Modeling and Machine Learning Based Energy Usage Forecasting in Smart Grid

Wei Yu[*], Dou An[§], David Griffith[†], Qingyu Yang[§], and Guobin Xu[*]

[*]Towson University
Towson, MD 21252
wyu@towson.edu
gxu2@students.towson.edu

[§]Xi'an Jiaotong University
Xi'an, Shaanxi, China
adkaka.an@gmail.com
yangqingyu@mail.xjtu.edu.cn

[†]National Institute of Standards and Technology
Gaithersburg, MD 20899
david.griffith@nist.gov

ABSTRACT

Developing effective energy resource management strategies in the smart grid is challenging due to the entities on both the demand and supply sides experiencing numerous fluctuations. In this paper, we address the issue of quantifying uncertainties on the energy demand side. Specifically, we first develop approaches using statistical modeling analysis to derive a statistical distribution of energy usage. We then utilize several machine learning based approaches such as the Support Vector Machines (SVM) and neural networks to carry out accurate forecasting on energy usage. We perform extensive experiments of our proposed approaches using a real-world meter reading data set. Our experimental data shows that the statistical distribution of meter reading data can be largely approximated with a Gaussian distribution and the two SVM-based machine learning approaches to achieve a high accuracy of forecasting energy usage. Extensions to other smart grid applications (e.g., forecasting energy generation, determining optimal demand response, and anomaly detection of malicious energy usage) are discussed as well.[1]

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Modeling Techniques

General Terms

Measurement, Performance

Keywords

Statistical Modeling Analysis, Energy Usage Forecasting, Machine Learning, Real-world Meter Reading Data, Smart Grid

1. INTRODUCTION

With recent developments in sensing, information, and communication technologies, the smart grid becomes a proposing system that makes the power grid more efficient, reliable, and secure. To efficiently deliver energy resources in the smart grid, an energy resource management strategy needs to be developed to balance the energy demand and supply [28]. Nonetheless, developing effective energy resource management schemes is challenging due to the entities on both the demand and supply sides experiencing numerous fluctuations. For example, on the supply side, fluctuations could come from distributed renewable energy resources due to solar irradiance, wind speed, etc. On the demand side, numerous effects, including natural disasters, plug-in vehicles, personal habits of using energy, weather and temperature, etc., could make it difficult to predict energy usage.

To address these issues, in this paper, we develop techniques to effectively manage energy resources and usage in order to adapt to fluctuations. Particularly, to balance energy demand and supply, we develop effective techniques to accurately model and forecast the amount of energy generation and demand over time. Therefore, the issue of quantifying fluctuations on the energy demand side can be addressed. It is worth noting that the techniques developed in this paper can be applied to the energy generation side as well. We also conduct the modeling analysis to derive a statistical model of energy usage and develop several machine learning based approaches to perform accurate forecasting of energy usage. The extensions to areas, including forecasting energy generation, determining optimal demand response, and anomaly detection of malicious energy usage, are discussed as well.

To summarize, the key contributions of this paper are as follows:

- First, using the real-world meter reading data set from Stanford University that consists of meter readings from houses over 200 days[2] as described in [18], we study the statistical distribution of real-world meter reading data using non-parametric tests, including the Shapiro-Wilk test [31] and the Quantile-Quantile plot normality test [9]. The experimental data shows that the distribution of meter reading data can be approximated with a Gaussian distribution.

- Second, we develop machine learning based approaches to conduct accurate energy usage forecasting. Partic-

[2]The authors would like to acknowledge Mr. Sebastien Houde at Stanford University for his dedicated help on providing the real-world smart meter measurement data set.

ularly, we consider the standard Radial Basis Function (RBF) based SVM, the Least Squares (LS) based SVM, and the Backward Propagation Neural Network (BPNN). In addition, we conduct extensive experiments using the aforementioned real-world meter reading data set to validate the effectiveness of these approaches. The experimental data shows that the two SVM-based approaches achieve a higher prediction accuracy than the BPNN based approach.

- Third, the techniques that we developed in this paper can be expanded to other areas as well, including the modeling and forecasting of energy generation, the optimal demand response, and anomaly detection of malicious energy usage. Using the prediction of wind speed as an example, the use of the SVM machine learning based approach can be used to effectively conduct the forecasting on the distributed energy resources in the energy supply side. In addition, the developed statistical modeling and forecasting results can be applied to derive the upper and lower bounds of energy usage and determine optimal demand response as well as anomaly detection of malicious energy usage.

The remainder of this paper is organized as following: The literature review is conducted in Section 2. The problem of balancing the energy demand supply and the developed approaches to perform the statistical modeling and forecasting of energy usage are presented in Section 3. The experimental results using real-world meter reading data set to validate the effectiveness of the developed approaches are shown in Section 4. The extensions of the work to other areas (e.g., forecasting energy generation, determining optimal demand response, and performing anomaly detection of malicious energy usage) are presented in Section 5. Finally, the conclusion is drawn in Section 6.

2. RELATED WORK

A number of research efforts have been conducted to improve energy transmission and distribution efficiency [6, 10, 25, 5, 20, 11]. For example, Guan *et al.* [10] proposed minimizing the overall cost of electricity and natural gas for a building operation. Chen *et al.* [5] proposed an optimal demand response scheme that could match electricity supply and shape electricity demand accordingly in both competitive and oligopolistic markets.

The challenges associated with the forecasting and demand response associated with energy usage were also discussed in [23]. Broadly speaking, energy usage forecasting can be categorized into short-term, medium-term, and long-term forecasting. For example, Hong *et al.* [13] adopted a multiple linear regression mechanism for conducting short-term forecasting, which provides an interpretability of the behavior of the electricity usage in the service territory. A semi-parametric additive model proposed by Fan *et al.* in [8] used a regression mechanism and investigated the nonlinear relationships between energy usage data and variables in the short-term time period. In addition, a human-machine co-construct intelligence framework was proposed in [14] to determine the horizon year load for a long term load forecasting.

Machine learning methods such as SVM and neural networks have been used in carrying out forecasting [2, 32, 37, 35, 1, 19, 15, 29]. For example, Shi *et al.* [32] developed a SVM-based model for one-day-ahead power output forecasting using the characteristics of weather classification.

Different from the existing research efforts, using the real-world meter reading data set [18], non-parametric tests were used to investigate the statistical distribution of energy usage. To the best of our knowledge, our paper is one of the first to validate that the statistical distribution of meter reading data can be largely approximated with a Gaussian distribution. In addition, two SVM and neural network based approaches were used to systematically perform the energy usage forecasting and the effectiveness of these machine learning approaches was systematically evaluated and compared. The findings from the paper can be extended to other areas, including the energy generation forecasting, the optimal demand response, and anomaly detection of malicious energy usage.

3. OUR APPROACHES

In this section, we first present an overview of the problem and our proposed approaches. We then describe the real-world data set and develop the non-parametric test based approaches to carry out statistical modeling. Finally, we discuss machine learning based approaches to perform energy usage forecasting.

3.1 Overview

In the smart grid, the electric power from generators can be delivered through the power grid to large geographical areas. High efficiency in power production and energy utilization can be realized through monitoring and control of power transmission and distribution processes. How to manage both bulk and distributed energy resources and the consumption levels of consumers to balance energy supply and demand is important. Nonetheless, developing effective management techniques to balance energy supply and demand is a challenging task because both sides experience various fluctuations.

To address this issue, we developed a statistical analysis and model of energy usage in this paper. We also developed machine learning based approaches to conduct accurate forecasting of energy usage. For the statistical modeling, we use two types of non-parametric test approaches to derive the distribution of energy usage based on real-world meter reading data. For forecasting energy usage, we developed several machine learning based approaches to conduct accurate energy usage forecasting. Energy providers can use these techniques to schedule energy generation and to make energy transmission and distribution efficient.

3.2 Real-world Energy Usage Data Set

We now introduce the real-world data set from Stanford University, which consists of meter readings from houses over 200 days (between February 2010 and October 2010) [18]. In this data set, weather information (e.g., mean temperature) for each 24 hour period is taken from archival data at *Weather Underground* website. We use meter readings and weather information for 283 houses in our experiments in Section 4.

An example of meter reading is shown in Table 1. From the table, each house is assigned an ID. The meter reading data for energy usage is measured hourly. The fields contained in the data set are shown in Table 2, which consists of the house ID , time, energy usage, the maximum, mean, and minimum value of temperature, and maximum and mean value of wind speeds. The house size (i.e., the area) is included as well. As an example, the information shown in Table 3 is the data associated with house 1001 that is for a rented townhouse, built in 2004, with $92.90 - 139.35$ sq. meters. In Table 4, we show an example of meter readings for energy usage and weather information at 2 p.m. from days 100 to 102 for house 1001. On day 100, the energy usage is 2.20 kilowatt hours (KWh) and the mean values of temperature and wind speed are 50 Fahrenheit degrees (F) and 20.92 Km Per Hour (KmPH), respectively.

Table 1. Data Range and Time Scale

Data Type	Range
ID of Houses	1-283
Time Interval	Hourly
Time Span	Approximately 200 days
Number of Data Points	Approximately 4800 (one per hour)

Table 2. Data Fields

Max_Temp	Mean_Temp	Min_Temp
Max_WindSpeed	Mean_WindSpeed	ID
Day-of-Year	Hour	Electricity Consumption

Table 3. Sample of House Information

ID	Building	Rent	Year Const.	Size
1001	Townhouse, duplex or row house	Rent	2004	92.90-139.35 sq. meters
1002	Single Family Detached House	Own	1992	185.81-232.37 sq. meters

3.3 Statistical Model of Energy Usage

To establish a statistical model of energy usage, we develop two non-parametric test based approaches to derive the statistical distribution of energy usage based on the aforementioned real-world meter reading data. We use a non-parametric test to carry out the analysis of the energy usage data. For a set of one-dimensional data, common non-parametric test approaches include the Shapiro-Wilk test [31] and the Kolmogorov-Smirnov (K-S) test [12]. It is worth noting that because the K-S test demands the preknowledge of the distribution of the sample data, the test result will not be credible if the population's Cumulative Distribution Function (CDF) is estimated from the sample data. It is worth noting that the predetermined CDF of the meter data is not known, so we consider the Shapiro-Wilk test to test the distribution of the sample data. We also use another non-parametric test approach, which is also called Quantile-Quantile (Q-Q) plot normality test, to confirm the distribution of meter reading data [36]. On the plot, when two data sets are identically distributed, the Q-Q plot will be shown a line. Then, we know that the greater the departure from the reference line, the greater the chance that the two data sets are drawn with different distributions.

Table 4. An Example of Real-World Meter Reading Data

Day	EU	Max_T	Mean_T	Min_T	Max_W	Mean_W
100	2.20	55	50	46	33.80	20.92
101	1.29	57	54	50	22.53	12.87
102	1.58	59	54	50	22.53	11.27

[1] T stands for temperature (Fahrenheit degree (F)), W stands for wind speed (Km per hour (KmPH)), and EU stands for energy usage (kilowatt hour (KWh))

3.4 Machine Learning Based Approaches for Energy Usage Forecasting

To accurately forecast energy usage in the smart grid, we use the following machine learning based approaches: neural network based machine learning, the standard SVM and the least squares SVM.

3.4.1 Neural Network Based Machine Learning

There are a number of research efforts on neural networks [16, 17]. A classic example of one of these neural networks is the Backward Propagation (BP) neural network, which consists of three layers: *input layer*, *hidden layer*, and *output layer*. Note that the error between real value and estimated value will be propagated backward from output layer to hidden layer and from hidden layer to input layer. The error of each layer can be re-estimated and the weights can be assigned correspondingly. Parameters for neural networks are set through a training process that use known data sets as input. After the training process, the trained model can then be used to carry out forecasting.

3.4.2 Standard SVM and LS-SVM

The standard SVM was originally proposed by V. N. Vapnik *et al.* [7]. Generally speaking, the SVM is one of the popular methods to efficiently classify data and to build a classifier, which can be further used to carry out forecasting. In SVM, the data and associated features can be treated as a point and vectors in multi-dimensional space. The basic principle of a standard SVM is to find a hyperplane, which could divide the points into different spaces. By doing so, we can classify data into different categories [27]. In order to minimize the classification error, the proper hyperplane needs to be determined.

The least squares SVM that is also denoted as LS-SVM is an enhanced SVM [33]. In a LS-SVM, there are two major enhancements in comparison with the standard SVM. First, the inequality constraints are substituted by equality constraints. Second, the squared loss function is used in the objective function [34]. In our experiment, we use the radial basis function as the kernel function in LS-SVM due to its wide use.

3.4.3 Workflow for Energy Usage Forecasting

As shown in Figure 1, the main process of machine learning based approaches can be divided into the following three steps: (i) data preprocessing, (ii) input feature selection, and (iii) energy usage forecasting. In the following, we describe these steps in detail.

Step 1: Data Preprocessing. To make our data more suitable for energy forecasting, data preprocessing needs to performed first. Note that the real-world energy usage data

cannot be directly used due to the following reasons: (i) the data is lacking attribute values that could be caused by the measurement noise of meters; and (ii) existing noises or bad data could be deviated from the norm values due to malfunction or unexpected events in the system (e.g., failures, power cuts, and/or natural disasters, etc.).

To address these issues, we introduce an interpolation mechanism to fill the missing values in the experimental data set and smooth incorrect data values with the average value of points around them. The missing data is filled using a linear interpolation mechanism. For bad data, because energy usage has continuity, the data located before and after adjacent time periods should not have a distinct change. Therefore, the average value in a continuous period of time can be considered as a baseline. Then, data beyond the baseline could be treated as bad data. Our experiments on the aforementioned real-world data set shows that the percentages for missing values and bad data are 3.08 % and 3.26 % on average, respectively. Therefore, around 6.34 % of data in the real-world data set used in this paper needs to be reprocessed using the mechanism discussed above.

Step 2: Input Feature Selection. As described in Section 3.2, various factors (e.g., weather and/or user's behaviors) can affect levels of energy usage. To achieve accurate energy usage forecasting, the selection of input features is important. The common way for feature selection is to choose related input variables such as the energy usage data in the past few days, humidity, temperature, and wind speed.

Recall that each component in the training data set is denoted as a feature. Here, the type of feature should be considered to include in input vectors. In this paper, the input features consist of two basic features: (i) hourly historical energy usage, and (ii) weather information. For the historical energy usage data, the measurements of the previous three hours are selected as input elements. The *relieff* [30] algorithm was used to determine the importance of features. In particular, the algorithm appraises features one by one and assigns a weight to each feature to indicate its importance. The larger the weight, the higher the importance of the feature. Table 5 and Table 6 illustrate the weight of all features related to weather in the experimental data and the results of the input features selection, respectively. In the experiments, the top three largest weights are selected and the energy usages in three hour timespans are chosen as input features for each house. To achieve rapid convergence during the training process, the network input data and the corresponding output data for the forecasting models are normalized such that all data is mapped into the range of $[-1, 1]$.

Table 5. Weight of Weather Features

ID	Max_T	Mean_T	Min_T	Max_W	Mean_W
1002	0.2539	0.2367	0.5011	1.4756	0.6090
1035	-0.0004	0.0060	0.0032	-0.0032	0.0090
1044	0.0013	0.0015	0.0013	0.0006	0.0019

[1] T denotes temperature, W denotes wind speed

Step 3: Energy forecasting with SVMs. After the feature selection, the energy usage data should be divided into two parts: (i) training set, and (ii) testing set. The training set is used to train the learning models and the optimal setting for parameters. The important parameters include the width of $\varepsilon-$insensitive tube ε and the error cost C, which are discussed in Section 3.4. After completing the above process, a trained SVM model is complete. Then, the trained model is ready to predict future energy usage.

Note that the training process of SVM can be formulated as solving a quadratic programming (QP) problem, which is optimized by a numerical method. The time complexity of the QP problem is of $O(n^3)$, where n is the number of training examples. For LS-SVM, the QP problem can be transformed into linear equations, thus the time complexity reduces to $O(n^2)$. Therefore, the time complexity of SVMs increases with an increase of training examples, which will not be correlated with a class of energy consumers. It worth noting that the development of distributed computing, parallel computing, and cloud computing can be used to speed up the training and decision process described in the paper.

Table 6. Input Features

ID	Results of Input Features Selection
1002	Max_W, Min_T, Mean_W and data in previous three hours
1035	Mean_T, Min_T, Mean_W and data in previous three hours
1044	Max_T, Mean_T, Min_T and data in previous three hours

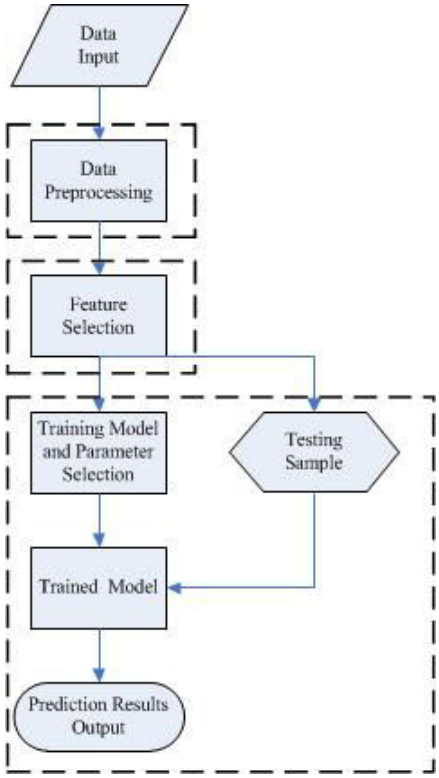


Figure 1. Workflow of Machine Learning Based Energy Usage Forecasting

4. PERFORMANCE EVALUATION

In this section, we introduce the performance evaluation results. We first introduce the experimental setup and then present the results of statistical modeling and energy usage

forecasting.

Based on the real-world meter reading data set described in Section 3.2, we carried out extensive experiments to evaluate the effectiveness of our developed statistical modeling and machine learning based energy usage forecasting approaches. MATLAB R2010b[3] was used to implement our developed approaches and the experiments were performed on a laptop PC (Centrino Duo, 2.3 GHz, 3 GB RAM). The toolkit LIBSVM in Matlab [4], a library for SVMs that includes the implementation of both SVM and LS-SVM, was used in our experiments. For comparison purposes, the neural network toolbox in Matlab was also used to evaluate the performance of the Backward Propagation (BP) neural network based forecasting approach, one of the classical neural networks that consists of three layers: *input layer*, *hidden layer*, and *output layer* [17].

4.1 Results of Statistical Modeling

To perform the statistical modeling of energy usage, the two non-parametric test approaches: Shapiro-Wilk test and Q-Q plot normality test as we discussed in Section 3.3 are used. In our experiments, the meter reading measurements over the following three time windows are aggregated: (i) morning (8:00-12:00), (ii) afternoon (14:00-18:00), and (iii) evening (20:00-24:00). Due to the space limitations, we only show limited scenarios here using the energy usage measurements for house 1002, house 1035, and house 1044 as examples. It is worth noting that two non-parametric tests on 200 houses were performed at a significance level of $\alpha = 0.05$. The experimental data shows that the meter readings of 148 houses could be approximated by a Gaussian distribution. In addition, more than 40 % of the remaining 52 houses contain a number of 0 values and error information, which largely deviate from the normal values, leading to the failures of the tests.

The Shapiro-Wilk test [31] with a significance level ($\alpha = 0.05$) is used for the measurements in individual time windows. It is worth noting that α is defined as the probability that a Gaussian distribution approximation is mistakenly rejected whereas it is actually true. Here, we consider two hypotheses: (i) H_0: the data follows a Gaussian distribution; and (ii) H_1: the data does not follow a Gaussian distribution. The P-value, in contrast to the threshold α, is computed based on the test statistics, which can be denoted as the probability, in the case of the null hypothesis H_0, of sampling results being equal to or being closer to the actual sampling results. As such, when the P-value is less than the predetermined significance level α, the observed results are be highly unlikely under the null hypothesis.

In our experiments, the P-value obtained from the meter reading measurements for the morning, afternoon, and evening windows are illustrated in Table 7. As shown in the table, in addition to the P-value for the morning meter reading measurements at house 1002, the remaining P-values are larger than the threshold of 0.05. Therefore, the energy usage in morning, afternoon, and evening windows of these houses can be approximated with a Gaussian distribution. It is worth noting that the Shapiro-Wilk test on morning data measurements from house 1002, in which the P-value is 0.000813, is an example of a failure case. As the P-value is far less than the threshold of 0.05, the morning data from house 1002 cannot be approximated with a Gaussian distribution.

The Q-Q plot normality test [36] is also used to test the distribution of meter reading measurements. As an example, the energy usage in house 1002 in three time windows is shown in Figure 2. The trend of points in Figures 2(b) and 2(c) has a higher degree of approximation to a straight line than the one in Figure 2(a), which indicates that the energy usages at noon and evening times can be better be approximated with the Gaussian distribution. This is because the closer the points are to a line, the closer the reading is to a Gaussian distribution. In Figure 2, there is significant deviation in the quantiles associated with the tails of the distribution whereas there is close agreement near the median. To summarize, the results of the two statistical test approaches draw the same conclusion, that is, the meter reading measurements for the three time windows at the three houses can be approximated with a Gaussian distribution.

Table 7. Results of Shapiro-Wilk Test

ID	Time Window	P-value	Hypothesis
1002	Morning	0.000813	Reject
	Noon	0.3407	Accept
	Evening	0.3236	Accept
1035	Morning	0.08509	Accept
	Noon	0.6062	Accept
	Evening	0.526	Accept
1044	Morning	0.4816	Accept
	Noon	0.6121	Accept
	Evening	0.6593	Accept

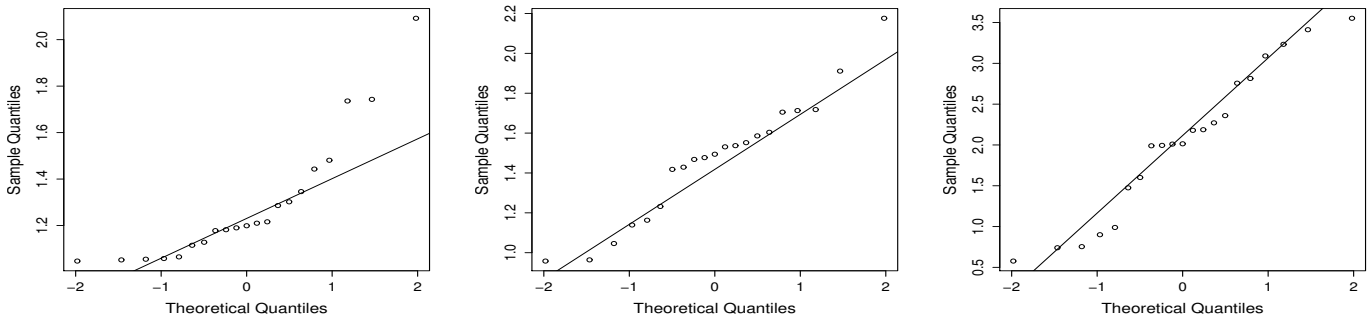
4.2 Results of Energy Usage Forecast

Experiments based on the real-world meter reading data set used in this paper were conducted to validate the effectiveness of two types of SVM presented in Section 3 and BP neural network based approaches in terms of the accuracy of energy usage forecasting. In our experiments, based on the models learned through the training process from the historical energy usage of the past 500 hours, we show the the accuracy of energy usage forecasting in the next 48 hours.

To measure the accuracy of forecasting, the following three metrics are considered: (i) *MAPE* (Mean Absolute Percentage Error), (ii) *MSE* (Mean Square Error), and (iii) *Coefficient of Regression* γ^2, which are used to measure the error between the actual and predicted energy usage. These metrics are defined as follows: $MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y_i}}{y_i}\right|$, $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y_i})^2$, and $\gamma^2 = \frac{\sum_i^n(\widehat{y_i}-\overline{y})^2}{\sum_i^n(y_i-\overline{y})^2}$, where y_i, $\widehat{y_i}$ and $\overline{y}$ are actual value, forecasted value, and mean value of the actual value, respectively.

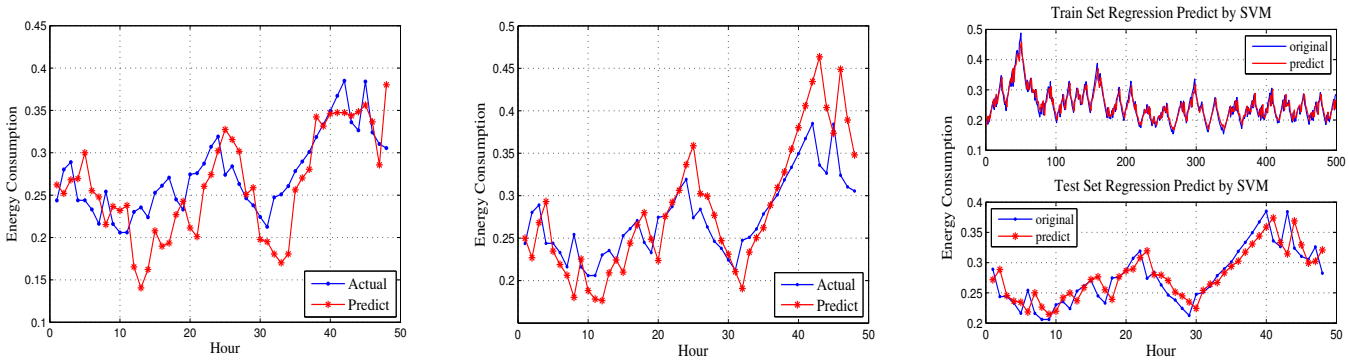
We conducted a large number of experiments on meter reading data for 200 houses. Due to space limitations, only a limited number of results are shown here for demonstration purposes. Based on the workflow showed in Section 3.4.3, the generic optimization mechanism provided by the LIB-

[3]Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology (NIST), nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.



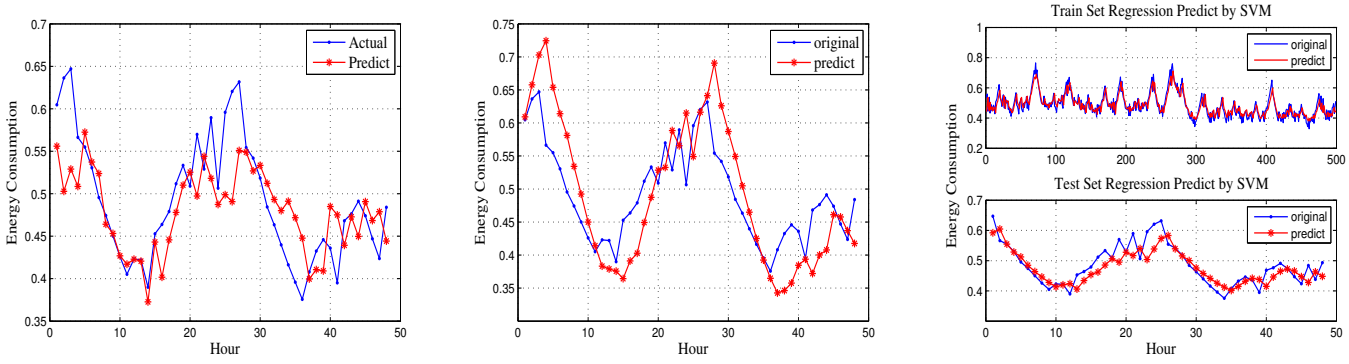
(a) (b) (c)

Figure 2. Q-Q Plot on No. 1002 House (a) Morning, (b) Noon, and (c) Evening



(a) (b) (c)

Figure 3. Forecasting Accuracy of (a) BPNN, (b) LS-SVM and (c) SVM on No. 1002 House



(a) (b) (c)

Figure 4. Forecast Accuracy of (a) BPNN, (b) LS-SVM and (c) SVM on No. 1035 House

SVM toolkit [4] is used to select key parameters for the SVM, including the width of insensitive tube ε and the cost of error C. Table 8 shows the forecasting accuracy of the two SVM based approaches in comparison with the BP neural network based approach (denoted as BPNN). From this table, the standard SVM based approach achieves the MSE at a magnitude of 10^{-4} and the highest coefficient of 0.88 in comparison with the LS-SVM and BPNN based approaches. For the LS-SVM based approach, all its MAPE values are smaller than 10 % whereas the MSE values are around 0.01. Further, the coefficient of regression approaches 0.84, which is better than the one achieved by the BPNN based approach. This can be explained as the neural network can easily fall into a local minimum instead of the global mini-

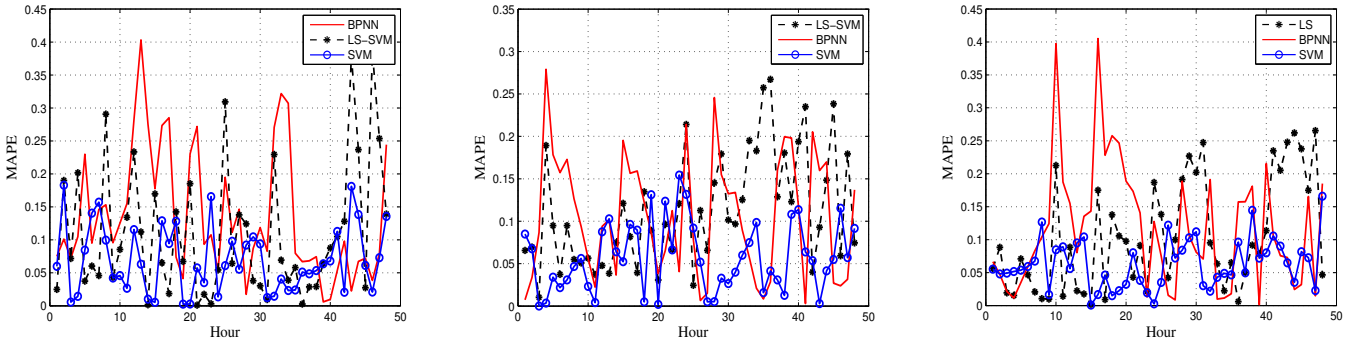


Figure 5. MAPE of (a) No. 1002, (b) 1035, (c) 1044 Houses

mum, leading to a lower accuracy of prediction.

Table 8. Effectiveness of Forecasting Results

ID	Index	SVM	LS-SVM	BPNN
1002	MSE	4.595e-04	0.0015	0.0352
	γ^2	0.8531	0.7482	0.2054
	MAPE(%)	6.9435	9.9221	13.6712
1035	MSE	8.2092e-04	0.0039	0.0278
	γ^2	0.7871	0.6939	0.1037
	MAPE(%)	5.7728	9.5013	11.2417
1044	MSE	6.3689e-04	0.0341	0.0174
	γ^2	0.8819	0.8405	0.3345
	MAPE(%)	4.3568	10.1023	10.6735

Table 9. Overall Forecasting Results

Method	Statistics	MAPE	γ^2	MSE	Time(s)
SVM	Mean	7.1261%	0.7593	0.0037	335.39
	Variance	0.0004	0.0144	0.0009	
LS-SVM	Mean	14.5649%	0.6219	0.0321	36.22
	Variance	0.002	0.0128	0.0014	
BPNN	Mean	16.8356%	0.4338	0.0732	29.28
	Variance	0.007	0.0130	0.0571	

The accuracy of energy usage forecasting for all 283 houses, including the statistical mean and standard derivation of MAPE, γ^2, and MSE for the three machine learning approaches, are demonstrated in Table 9. From this table, we can observe that the SVM achieves the highest forecasting accuracy and the BPNN achieves the worst forecasting accuracy. In addition, the time overhead of these machine learning based forecasting approaches, which is defined as the total time taken for inputting data, preprocessing data, selecting features, conducting training process, and generating forecasting results based on a training model for a single house, is evaluated. The experiments were conducted on a laptop PC (Centrino Duo, 2.3GHz, 3GB RAM). As shown in Table 9, the time overhead for the SVM, LS-SVM, and BPNN are 335.39 s, 36.22 s, and 29.28 s, respectively.

It is worth noting that in order to further improve the time efficiency, more powerful PC, conducting forecast using low level language (instead of using MATLAB), and leveraging techniques (e.g., cloud computing and parallel computing) can be used. For the SVM, time overhead is much greater than that of the LS-SVM and BPNN as the genetic algorithm optimization mechanism is used to select the key parameters for the SVM, including the width of insensitive tube ε and the cost of error C. For the LS-SVM, as explained before, two major enhancements of the LS-SVM in comparison with the standard SVM: (i) using equality constraints instead of inequality constraints, and (ii) using square loss function that can significantly simplify the complexity of the problem solving process, leading to a smaller processing time for carrying our energy usage forecasting.

In Figures 3 and 4, the accuracy of forecasting for the three machine learning approaches on houses 1002 and 1035 is demonstrated. As we can see from these figures, the blue and red curves represent the actual energy usage and forecasted energy usage, respectively. The blue curve and red curve for the SVM based approaches are highly coincidental with each other as shown in Figures 3(b)(c) and 4(b)(c) whereas the results of the BPNN based approach are shown in Figures 3(a) and 4(a). A higher consistency between the real data and forecasted data in the SVM based approaches indicates that the forecast of the SVM based approaches are more accurate than the BPNN based approach. In addition, note that the blue and red curves in the SVM based approaches almost follow the same trend, indicating the forecasted results of these approaches are accurate. For the two SVM based approaches, because a generic algorithm is used to obtain the optimal ε and C, the standard SVM based approach actually achieves a higher accuracy than the experimental data shown in Figure 5.

5. EXTENSION

In this section, the extensions are made from the following aspects: the modeling of energy generation, the optimal demand response, and anomaly detection of malicious energy usage.

5.1 Modeling of Energy Generation

The distributed energy resources are inherently stochastic. Using wind energy as an example, the total wind energy flowing through an imaginary area A at time t can be formalized as: $E = \frac{1}{2}At\varrho v^3$ [24], where ϱ is the density of air and v is the wind speed. Here, the wind energy E is highly

correlated with the wind speed v. Therefore, the forecasting of wind speed is one critical issue before wind energy resources can be broadly integrated in the smart grid.

Using the prediction of wind speeds as an example and applying it to the modeling approach developed in this paper, we are able to improve the ability of forecasting distributed energy resources at the energy supply side. Similar to the prediction of energy usage shown in Sections 3.4.2 and 4, the standard SVM machine learning approach can be used to carry out the prediction of wind speeds. We conducted experiments on the wind speed data of 193 days at No. 1002 house. Recall that as shown in Table 2, the wind data in the real-world data set used in this paper consists of both the max wind speed and mean wind speed for a day. The maximum and mean value of wind speeds in three days are selected as the input features in the standard SVM machine learning approach. The learning and forecasting process follow the same workflow as we described in Section 3.4.3. The mean value of wind speeds of the next two weeks is used to test the accuracy of forecasting. Figure 6 illustrates the accuracy of wind speed forecasting. The error metrics defined in Section 4 are also used to evaluate the accuracy of forecasting. The results of error metrics are 9.3876, 1.0339, and 0.5392, respectively, showing that the standard SVM machine learning approach could achieve a high accuracy of predicting wind speeds.

5.2 Optimal Demand Response

The results developed in this paper can also be used to determine optimal demand response, which allows customers to obtain real time energy prices and enables load shifting and reduction. In the following, we show an example of how to integrate our developed modeling results into the optimization model originally proposed in [5] for conducting optimal demand response. In [5], Chen *et al.* derived an efficient equilibrium based on the upper and lower bounds of customer's energy usage in a competitive market. Nonetheless, their original work did not show how to derive those bounds.

In the following, we briefly show how to apply the results developed in Section 3.2 to determine the optimal demand response. Without loss of generality, we assume that a power grid system consists of N customers, who are served by a power generator. On the demand side, let the power load of each customer be $q_i(t)$ at time t. Then, in a time window $[1:T]$, the bounds for minimum and maximum total energy usages, denoted as, $\underline{Q}_i$ and $\overline{Q}_i$, can be derived. On one hand, based on the results of the energy usage forecasting, the $\underline{Q}_i$ and $\overline{Q}_i$ in a near future time window are derived. In this way, the bound can be precise and is suitable for a short-time demand response process. On the other hand, based on the result of the developed statistical modeling analysis, the bounds in each time window can be derived as well. It is worth noting that bounds based on statistical modeling analysis are more general and suitable for a long-term demand response process. Choosing either the long-term bound or short-term bound can be determined by the time scope of demand response process. In the following, the bounds based on the statistical modeling are used as an example to demonstrate our idea.

Denote the mean and the standard deviation of energy usage as: $\overline{X} = \frac{1}{T}\sum_{t=1}^{T} q_i(t)$ and $S_n = \frac{1}{T-1}\sum_{t=1}^{T}(q_i(t) - \overline{X})^2$, respectively. Based on the statistical modeling results developed by this paper, $\underline{Q}_i$ and $\overline{Q}_i$ can be derived through the interval estimation mechanism [26] and are given by, $\underline{Q}_i = \overline{X} - t_{\frac{\alpha}{2}}(T-1)\frac{S_T}{\sqrt{T}}$, and $\overline{Q}_i = \overline{X} + t_{\frac{\alpha}{2}}(T-1)\frac{S_T}{\sqrt{T}}$, where $t_{\frac{\alpha}{2}}(T-1)$, $\overline{X}$ and S_T are the upper quantile fractile of student t distribution at the confidence level of α, mean value and standard deviation, respectively. Then, assume that each user i satisfies the following constraints in $[1:T]$, $\sum_{t=1}^{T} q_i(t) \geq \underline{Q}_i$, where $i \in N$, and $\sum_{t=1}^{T} q_i(t) \geq \overline{Q}_i$, where $i \in N$. For each user i, a utility function: $U_i(q_i, t)$ is defined to measure its satisfaction for the energy service, supplied by the energy generator, where q_i is the energy usage at time t. We also assume that $U_i(q_i, t)$ is continuously differentiable and increasing with respect to t monotonically.

On the supply side, depending on the state of the power grid, the energy price will be dynamic over time. Assume that the energy generator has a cost of $C(Q,\ t)$ when it supplies energy Q at time t. We also assume that $C(Q,\ t)$ increases with respect to Q and the marginal cost increases with respect to Q.

5.3 Anomaly Detection of Malicious Energy Usage

In an energy resource management system, it is important to report energy usage information from consumers to the utility supply. Nonetheless, this decision process could be impacted by an adversary, who might compromise meters and launch false data injection attacks to disrupt the smart grid operations [22, 21, 3]. Therefore, the detection of false data injections attacks becomes a critical issue. Note that our developed energy usage forecasting can be leveraged to carry out anomaly detection. To be specific, we can compute the lower and upper bounds of energy usage in a near future time window and use them as the baseline profile for conducting anomaly detection. In the following, we briefly demonstrate how to use our statistical modeling analysis results to detect malicious energy usages.

Based on the results in Section 3.2, we now present a hypothesis testing based detection scheme. We consider two hypotheses: (i) H_0: the measurement is valid, and (ii) H_1: the measurement is under attack. Based on our statistical modeling results, we assume that the energy usage measurements $X = (X_1, X_2, \ldots, X_n)$ in the three time windows (i.e., morning, noon, and evening) follow the Gaussian distribution $N(\mu, \sigma^2)$, in which μ and σ are all unknown parameters and n is the total number of measurements.

It is worth noting that the malicious measurement's deviation from the mean value can be treated as noise and the value of μ and σ are unknown to the detection system. Therefore, we consider that the standard deviation of samples, denoted as $S_n = \frac{1}{n-1}\sum(X_i - \overline{X})^2$, can reflect the dispersion of difference between the compromised measurement and the normal one. After letting $\mathbf{T} = \frac{\overline{X}-\mu_0}{S_n/\sqrt{n}}$, we have $\mathbf{T} \sim t(n-1)$. Based on this, the hypothesis test can be formalized as, $\mathbf{T} \underset{H_0}{\overset{H_1}{\gtrless}} \tau$, where $\tau = t_{\frac{\alpha}{2}}(n-1)$ is the threshold determined by considering the null hypothesis given a certain false positive rate α.

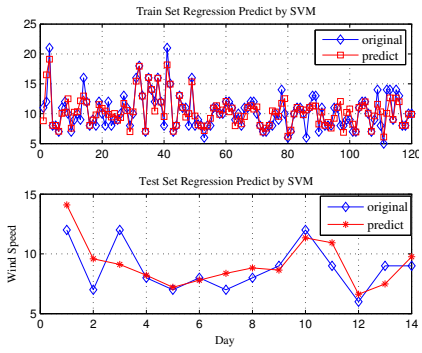


Figure 6. SVM Prediction Accuracy on Wind Speed

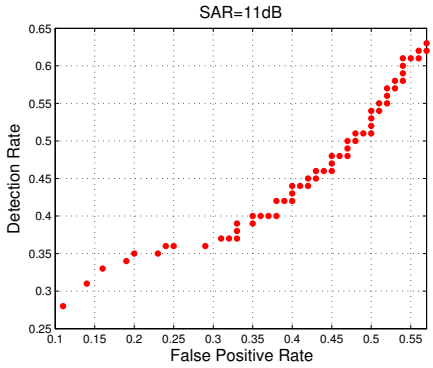


Figure 7. ROC Curve when SAR=11dB

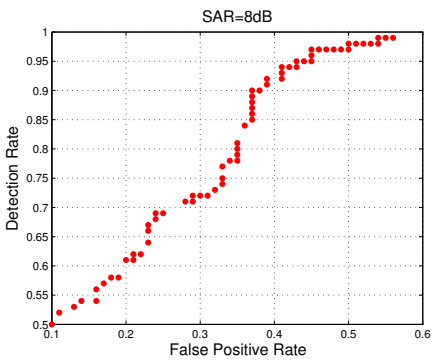


Figure 8. ROC Curve when SAR=8dB

To evaluate the effectiveness of anomaly detection based on hypothesis test, we choose the following two metrics, which are detection rate (same as the true positive) and false positive rate. Detection rate P_D is defined as the probability that the attack is correctly recognized and false positive rate P_F is defined as the probability that a normal measurement vector is misclassified as malicious. We use *Receiver Operating Characteristic* (ROC) curve to show the relationship between P_D and P_F and measure tradeoffs between detection rate and false positive rate. We run simulations based on measurements (e.g., measurements in the morning of 100 days on house No. 1002) to collect enough samples and estimate the mean value μ_0. Then, we set detection threshold τ based on the false positive rate $\alpha = 0.05$. We use the measurements of 100 days to present the normal measurements, which are not manipulated by the adversary and derive P_F with the detection threshold. After that, we simulate the malicious measurements in the following way. Similar to the signal-to-noise ratio (SNR), we first define signal-to-attack ratio (SAR) that is defined as $SAR = 10log_{10}\frac{X_i}{c_i}$ to quantify the strength of attacks, where X_i and c_i are the maliciously manipulated measurement and true measurement, respectively. We then apply the anomaly detection discussed above to derive detection accuracy P_D. Note that $SAR = 11$dB and $SAR = 8$dB represent that the adversary could change 8% and 12% of measurement values, respectively.

Figures 7 and 8 show the ROC curve of our detection algorithm. As we can see, when $SAR = 11dB$, the detection algorithm achieves an accuracy of 60% with a false positive rate of 55%, while the adversary could only change up to 8% of the true value of measurements. When $SAR = 8dB$, the detection rate approaches almost 100% with a false positive rate of 55% when the adversary can manipulate up to 12% of the true value of measurements. Here, we can obtain 90% detection rate with a false positive rate of less than 40%. As we can see from these figures, detection rate becomes higher when the attack strength increases. This is as expected, the standard deviation of malicious measurements is higher when the attack becomes stronger.

6. CONCLUSION

In this paper, the critical issue of quantifying uncertainty on the energy usage was addressed. Particularly, the Shapiro-Wilk test and Quantile-Quantile plot normality test were adopted to investigate the statistical distribution of energy usage and the machine learning based approaches (e.g., SVM and neural network) were developed to conduct the accurate forecasting of energy usage. Extensive experiments on a real-world meter reading data set were conducted to validate the effectiveness of the developed approaches. The experimental data shows that the energy usage can be largely approximated with a Gaussian distribution and the SVM-based machine learning approaches can accurately predict the energy usage. The extensions to other areas (e.g., forecasting energy generation, determining optimal demand response, and anomaly detection of malicious energy usage) were discussed as well.

7. REFERENCES

[1] M. Afshin, A. Sadeghian, and K. Raahemifar. On efficient tuning of ls-svm hyper-parameters in short-term load forecasting: A comparative study. In *Proceedings of IEEE Power Engineering Society General Meeting*, pages 1–6, Jun. 2007.

[2] Z. A. Bashir and M. E. El-Hawary. Applying wavelets to short-term load forecasting using pso-based neural networks. *IEEE Transactions on Power Systems*, 24(1):20–27, Feb. 2009.

[3] A. A. Cardenas, S. Amin, and S. Sastry. Secure control: Towards survivable cyber-physical systems. In *Proceedings of the First International Workshop on Cyber-Physical Systems*, pages 495–500, Jun. 2008.

[4] C.-C. Chang and C.-J. Li. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, Apr. 2011.

[5] L. Chen, N. Li, H. Low, and J. C.Doyle. Two market models for demand response in power networks. In *Proceedings of IEEE International Conference on Smart Grid Communications*, pages 397–402, Oct. 2010.

[6] A. J. Conejo, J. M. Morales, and L. Baringo. Real-time demand response model. *IEEE Transactions on Smart Grid*, 1(3):236–242, Dec. 2010.

[7] C. Corinna and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep. 1995.

[8] S. Fan and R. J. Hyndman. Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, 27(1):134–141, Feb. 2012.

[9] R. Gnanadesikan. *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley-Interscience, 2011.

[10] X. Guan, Z. Xu, and Q.-S. Jia. Energy-efficient buildings facilitated by microgrid. *IEEE Transactions on Smart Grid*, 1(3):243–252, Dec. 2010.

[11] Y. Guo, M. Pan, and Y. Fang. Optimal power management of residential customers in the smart grid. *IEEE Transactions on Parallel and Distributed Systems*, 23(9):1593–1606, Sep. 2012.

[12] Hazewinkel and Michiel. *Kolmogorov-Simirnov test*. Springer, 2001.

[13] T. Hong, M. Gui, M. Baran, and H. Willis. Modeling and forecasting hourly electric load by multiple linear regression with interactions. In *Proceedings of IEEE Power and Energy Society General Meeting*, pages 1–8, Jul. 2010.

[14] T. Hong, S. Hisiang, and L. Xu. Human-machine co-construct intelligence on horizon year load in long term spatial load forecasting. In *Proceedings of IEEE Power and Energy Society General Meeting*, pages 1–6, Jul. 2009.

[15] T. Hong, P. Wang, A. Pahwa, M. Gui, and S. M. Hsiang. Cost of temperature history data uncertainties in short term electric load forecasting. In *Proceedings of International Conference on Probabilistic Methods Applied to Power Systems*, pages 212–217, Jun. 2010.

[16] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of The National Academy of Sciences*, 79(8):2554–2558, Apr. 1982.

[17] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. NatL Acad. Sci.*, 79:2554–2558, 1982.

[18] S. Houdea, A. Todd, A. Sudarshan, J. A. Flora, and K. C. Armel. *Real-time Feedback and Electricity Consumption: A Field Experiment Assessing the Potential for Savings and Persistence*. `http://www.stanford.edu/~shoude/FieldExperimentPowermeter_vfinal_July2011.pdf`.

[19] W. Li and P. Choudhury. Probabilistic planning of transmission systems: Why, how and an actual example. In *Proceedings of IEEE Power and Energy Society General Meeting Conversion and Delivery of Electrical Energy in the 21st Century*, pages 1–8, Jul. 2008.

[20] R.-H. Liang and J.-H. Liao. A fuzzy-optimization approach for generation scheduling with wind and solar energy systems. *IEEE Transactions on Power Systems*, 22(4):1665–1674, Nov. 2007.

[21] J. Lin, W. Yu, X. Yang, G. Xu, and W. Zhao. On false data injection attacks against distributed energy routing in smart grid. In *Proceedings of ACM/IEEE Third International Conference on Cyber-Physical Systems (ICCPS)*, pages 183–192, Apr. 2012.

[22] Y. Liu, M. K. Reiter, and P. Ning. False data injection attacks against state estimation in electric power grids. In *Proceedings of the 16th ACM conference on Computer and communications security*, pages 21–32, Nov. 2009.

[23] P. Luh, L. Michel, P. Frieland, C. Guan, and Y. Wang. Load forecasting and demand response. In *Proceedings of IEEE Power and Energy Society General Meeting*, pages 1–3, Jul. 2010.

[24] G. Masters. Wind power systems. *Renewable and Efficient Electric Power Systems*, pages 307–383, 2004.

[25] J. Medina, N. Muller, and I. Roytelman. Demand response and distribution grid operations: Opportunities and challenges. *IEEE Transactions on Smart Grid*, 1(2):193–198, Sep. 2010.

[26] J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London*, 236(767):333–380, 1937.

[27] W. S. Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

[28] D. of Energy. *Demand Response*. `http://energy.gov/oe/technology-development/smart-grid/demand-response`.

[29] P. Pinson, C. Chevallier, and G. N. Kariniotakis. Trading wind generation from short-term probabilistic forecasts of wind power. *IEEE Transactions on Power Systems*, 22(3):1148–1156, Aug. 2007.

[30] M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(1-2):23–69, Oct.-Nov. 2003.

[31] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

[32] J. Shi, W.-J. Lee, Y. Liu, and Y. Yang. Forecasting power output of photovoltaic systems based on weather classification and support vector machines. *IEEE Transactions on Industry Applications*, 48(3):1064–1069, May 2012.

[33] J. A. Suykens and V. J. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, Jun. 1999.

[34] J. A. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, J. Suykens, and T. Van Gestel. *Least squares support vector machines*, volume 4. World Scientific, 2002.

[35] Y. Wang, Q. Xia, and C. Kang. Secondary forecasting based on deviation analysis for short-term load forecasting. *IEEE Transcation On Power Systems*, 26(2):500–507, 2011.

[36] M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17, 1968.

[37] Z. Yun, Z. Quan, and S. Caixin. Rbf neural network and anfis-based short-term load forecasting approach in real-time price environment. *IEEE Transactions on Power Systems*, 23(3):853–858, Aug. 2008.

ABOUT THE AUTHORS:

Wei Yu received the B.S. degree in electrical engineering from Nanjing University of Technology, Nanjing, China, in 1992, the M.S. degree in electrical engineering from Tongji University, Shanghai, China in 1995, and the Ph.D. degree in computer engineering from Texas A&M University in 2008. He is currently an Associate Professor with the Department of Computer and Information Sciences, Towson University. Before joining Towson, he was with Cisco Systems Inc. for nine years. His research interests include cyberspace security, computer networks, and cyber-physical systems. He received the NSF Faculty Early Career Development (CAREER) award in 2014 and the Best Paper Award at the 2013 and 2008 IEEE International Conference on Communications (ICC), respectively.

Dou An received the B.S. degree in applied mathematics from Northwestern Polytechnical University, China, in 2011. He is currently working toward the Ph.D. degree in the Department of Automation Science and Technology, Xi'an Jiaotong University. His current research interests include cyber-physical systems, power grid security.

Qingyu Yang received the B.S. and M.S. degrees both in mechatronics engineering from Xi'an Jiaotong University, China, in 1996 and 1999, respectively, the PhD degree in control science and engineering from Xi'an Jiaotong University, China, in 2003. He is a professor in School of Electronics & Information Engineering at Xi'an Jiaotong University. He is also with the State Key Laboratory for Manufacturing System Engineering, Xi'an Jiaotong University. His current research interests include cyber-physical systems, power grid security, control and diagnosis of power system, and intelligent control of industrial process.

David Griffith received the PhD in electrical engineering from the University of Delaware. He worked on satellite communications systems at Stanford Telecommunications and Raytheon, and is currently with the Communication Technology Laboratory at the National Institute of Standards and Technology (NIST). His research interests include mathematical modeling and simulation of wireless communications networks, including public safety broadband networks and communications for the smart grid. He is also working on spectrum monitoring and spectrum sharing.

Guobin Xu received the B.S. degrees in mathematics and economics from Qingdao University, in 2009. He is currently a Ph.D. candidate in the Department of Computer and Information Sciences at Towson University. His research interests include cyber-physical systems, computer networks, and cyber security.