# Wind Power and Electric Load Forecasting

Jie Wu[1], Mengwei Liu[2], Xuhua Gao[3]
Department of [1]Civil and Environmental Engineering, [2]Electrical Engineering, [3]Energy Resources Engineering, Stanford University, USA

***Abstract* - As renewable energy increasingly integrates into the electric power system, electric load forecasting and renewable energy power generation forecasting become more important. In this project, ARIMA and NARX are applied to build load forecasting model focusing on improving statistical and computational efficiency without losing accuracy. ARIMA turns out to be better for short term forecasting while NARX is more stable and efficient. Meanwhile, forecasting power generation by wind farm is implemented by using a hybrid k nearest neighbors and least square boosting algorithm. A supply shortage measured by the difference between load and generation is forecasted with a 95% confidence that the error is less than 24.73%.**

## INTRODUCTION

Renewable energy integration is an essential branch of modern electric power system. One of the biggest challenges that renewable energy integration faces is to match generation and load, which in fact is crucial in any electric power systems since electric energy is hard to store. Energy forecasting has been a fundamental task in electric power system research for more than half a century. With the enormous advantages of renewable energy, it's easy to see that modern energy strategy should be to use electricity generated by renewable resource as much as possible, while employing traditional power plant to make up for the shortage. In the case of extra energy, there should be appropriate size facility to store it. To achieve this, a relatively accurate prediction of both load and amount of electricity generated by renewable source is indispensable, which unfortunately is very difficult. The load is hard to predict since it depends on many factors including weather and terminal users' behavior pattern. Furthermore, what makes it even more challenging is that electricity generated by renewable source like wind is highly stochastic and very difficult to predict considering the complexity of weather conditions. Therefore, the need to simultaneously predict two stochastic variables presents a unique task for renewable energy integration. In this project, the behavior of both the load and electric power generation are modeled, and the difference value between them is further explored to estimate energy supply shortage, which can provide guidance on energy strategy.

The main objectives include: 1) Efficiency improvement: Reduce the amount of training data without trade-off accuracy. 2) Maximize performance: Focus on the points where the data show more complexity and are harder to predict. 3) Overall forecasting: Combine the load and wind forecasting three day ahead to provide energy supply guidance.

## DATASET AND FEATURES

The hourly electric load and temperature data are collected from 2004 to 2014 (96432 data samples in total). The data are publicly available from ISO New England website. The test data are selected in year 2014. Data from 2004 to 2013 are treated as training data base. The electric load data has two major characteristics: 1) Periodicity: The electric load data has obvious daily cycles, weekly cycles, monthly cycles and yearly cycles. Figure 1 shows the probabilistic distribution of daily electric load data generated by 10 years' data. 2) Electric load data is highly correlated with the temperature, as shown by Figure 2.

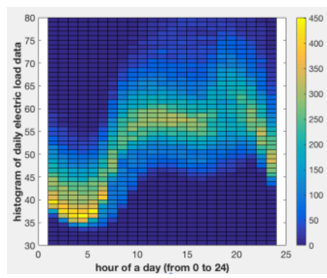


*Figure 1 Probabilistic distribution of daily electric load data*

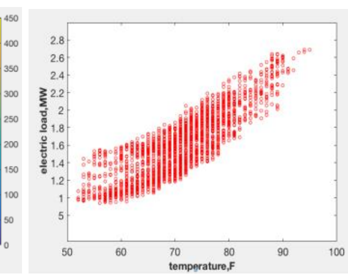


*Figure 2 Relationship between electric load and temperature*

Hourly based meridional and zonal components of wind forecast and power generation data from 2012 and 2013 for a wind farm are collected. The wind data is publicly availbale from the Global Energy Competition 2014. There exist three major differences between wind data and load data : 1)wind data have weaker seasonal change. 2) wind data almost have no fixed periodicity. 3)wind data are more stochastic, with weak correlation with single atmospheric variable such as wind speed or wind direction as shown by Figure 3. Outlier removal is implemented by validation.

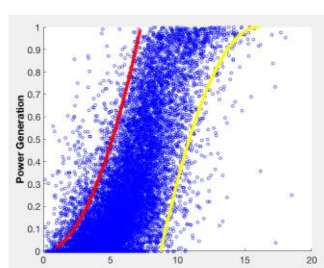


*Figure 3 Wind power generation is not strongly correlated with predictors.*

## METHODS

*1. Wind power generation forecasting methods*

1.1 Multivariate Linear Regression

Multivariate linear regression algorithm (MLR) is a simple, straightforward and powerful method for many problems including wind power forecasting, which is also a good starting point to help understand predictors. Some reasonable predictors are created based on given features including wind speed and direction 10 m and 100 m above ground, ratios of these values, responses 1, 2, 3 and 24 hours ago, hour of the day, day of the month and month. An interaction term, multiplication of wind speed and hour of the day is added as well. A forward feature selection is also applied.

1.2 k-Nearest Neighbors algorithm

A nonparametric algorithm, k-Nearest Neighbors (kNN) is also used for wind power generation forecast in this project. The basic idea is to find k closest samples in training data and use the average response of these k samples as predicted response. Total distance between current sample and a training sample is calculated by following equation [1].

$$D(x_i, x_j) = \sum_{p=1}^{n_p} w_p * d(x_i^p, x_j^p) \tag{1}$$

$n_p$ is total number of features selected and $w_p$ is the weight for distance, $d(x_i^p, x_j^p)$, calculated based on each feature. All features are normalized to avoid scaling issue. Features are selected based on regression results from MLR. Validation set is then used to select the value of k and weights.

1.3 Least Square Boosting

Least Square Boosting (LSB) is another nonparametric algorithm applied to forecast wind power generation [2]. At each step, a week learner, $h_{i+1}$, is fitted towards residual between predicted response of previous step $F_i$ and observed response, y, as shown in equation (2). λ is the learning rate. Regression trees are used as weak learners in this project and each step minimizes mean-square error.

.

$$h_{i+1} \to y - \lambda * F_i \tag{2}$$

*2. Electric load forecasting methods*

2.1 Nonlinear Autoregressive Neural Network

A process is called autoregressive process of order n, if it can be expressed using equation (3) [3]:

$$x_t = F(x_{t\text{-}1}, x_{t\text{-}2}, \dots, x_{t\text{-}n}) + \varepsilon_t \tag{3}$$

where $F$: $R^n \to R$, and $\varepsilon_t$ is a i.i.d. $N(0,\sigma^2)$ random variable. In the case of Nonlinear Autoregressive Neural Network, $F$ is a nonlinear function. The hyperbolic tangent function is commonly used. If there exist no external input, the neural network is simply nonlinear autoregressive (NAR). However, if external input is considered, the neural network becomes nonlinear autoregressive with exogenous input (NARX). Temperature is the external input in this project. Figure 4 shows the structure of NAR and NARX.

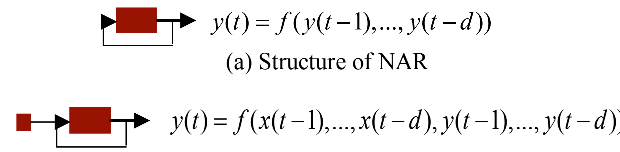


*Figure 4 The structure of NAR and NARX*

Levenberg-Marquardt backpropagation is implemented in the training process, because it has higher computational efficiency and tend to produce smaller error than other training algorithms. Effect of feedback delay steps (d) and hidden layer size (h) on the model performance and the risk of over fitting are studied, as shown in Figure 5.

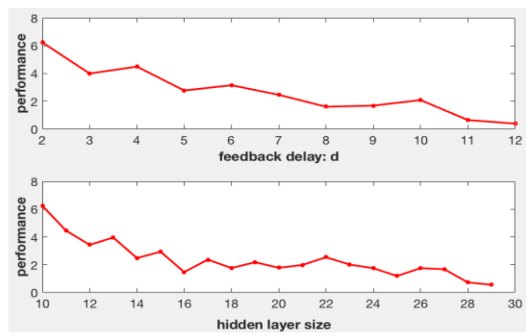


*Figure 5 Impact of feedback delay(d) and hidden layer size (h) on performance.*

As expected, increasing h or d indicates increasing model complexity and reducing error, which also increases the chance of overfitting. When h = 25 and d = 12, the training error is much lower than test error, indicating overfitting. In this project, hidden layer size and d are set to be 10 and 4 respectively.

2.2 ARIMA

An ARIMA model consists of three parts: autoregression (AR), moving average (MA) and integration (I). AR and MA models are used to describe stationary processes, and I is for differencing a non-stationary process to a stationary one [4]. The hourly load data have two significant characteristics: seasonality, as mentioned above, and non-stationarity [5]. Therefore, a seasonal ARIMA model could be a proper method to forecast the non-stationary load behavior. Highly non-stationary and daily as well as weekly seasonality indicated by the autocorrelation function (ACF) shown in Figure 6 allow appropriate setting of ARIMA model. Furthermore, log transformation is implemented considering multiplicative seasonal pattern of data [6].

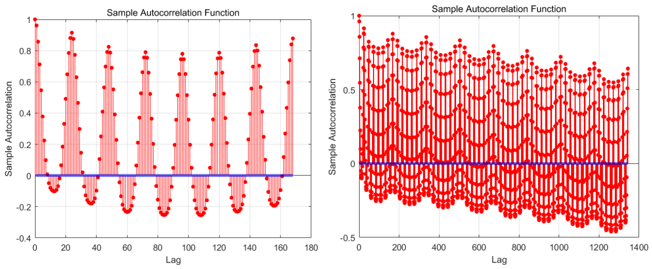

*Figure 6 Weekly and monthly autocorrelation of load*

However, external feature's impact cannot be fully captured by ARIMA. Figure 7 shows a preliminary result of a two-week forecast, yielding a high mean absolute percentage error (MAPE) of 0.6177. Although it correctly captures the overall trends, forecast for day 8 and day 9 is inaccurate. The major reason is that an abrupt temperature change occurred during these two days.

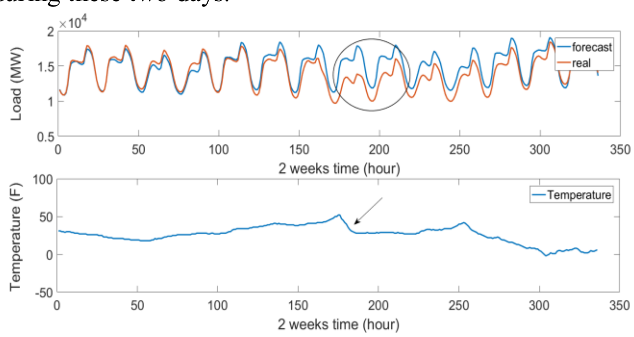

*Figure 7 Temperature and load variations in the two weeks of interest. Note that a temperature drop coincides with the sudden change of load in the circle.*

## RESULTS AND DISCUSSION

*1. Wind power generation forecasting*

Since the objective of this project for wind power generation analysis is to forecast power generation three days ahead. 10 three-consecutive-day are randomly selected from second half of 2013 as testing space for all methods. Root-mean-square error (RMSE) is used as measurement method for error.

MLR is applied first. Data of 2012 is used as training samples, data between January and June of 2013 is used as validation samples. Forward feature selection selected following predictors: wind speed and direction 100 m above the ground, ratio of wind speed between values of 100m and 10m, ratio of wind direction between values of 100m and 10m, responses 1 ,2 ,3 and 24 hours ago. All data before June of 2013 are used as training samples to fit model. Test RMSE is 0.2176.

kNN is then applied. Wind speed and direction 10 m and 100 m above the ground as well as hour of the day are selected as distance measuring features. The same validation set is used. The values of k are selected as 100. Test RMSE turns out to be 0.1988 which is significant better than that of MLR. Considering kNN and MLR may catch different pattern of wind power generation, a hybrid method is then created. It linearly combines MLR and kNN using an optimal weight (0.31 and 0.69) selected from validation set. Then test RMSE is improved to be 0.1938, which is only slightly better than error of kNN. A random three-consecutive-day forecasting is shown in Figure 8. Generally, kNN can capture local fluctuation better.

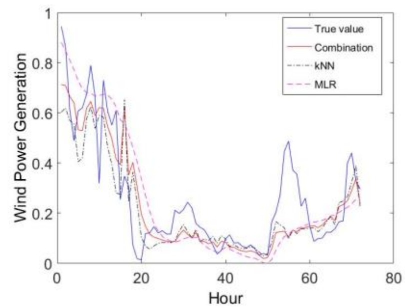

*Figure 8 Predicted Wind Power Generation for a random three-consecutive-day*

To further improve forecasting, LSB is applied. 5-fold cross validation is applied to select optimal parameter for LSB. All data before July of 2013 serves as training data. Maximum number of splits of regression tree is selected, and learning rate is selected to be 0.1 and optimal number of regression trees is 132.

Using LSB trained with optimal parameter, test RMSE turns out to be 0.1970. It is better than result of MLR and kNN but not the hybrid. So, several hybrid methods are tried and the result is summarized in following Table 1. The combination of kNN and LSB has lowest error. Figure 9 shows the updated result.

Table 1: Result of all method

| Method | Test RMSE |
|---|---|
| MLR | 0.2176 |
| kNN | 0.1988 |
| LSB | 0.1970 |
| kNN+MLR | 0.1938 |
| kNN+LSB | 0.1850 |
| LSB+MLR | 0.1970 |
| MLR+kNN+LSB | 0.1880 |

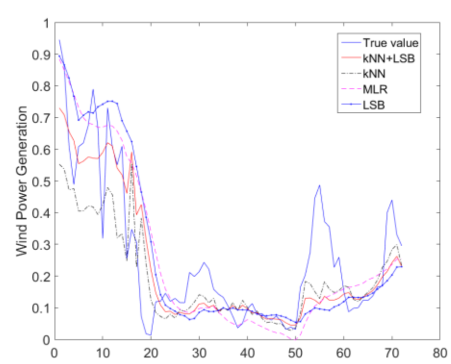

*Figure 9 Predicted Wind Power Generation using LSB*

*2. Electric load forecasting*

Both NARX and ARIMA are used to model the behavior of electric load. Figure 10 shows an example of load forecasting with both models. In this case, three days from 2014 is selected as the target .

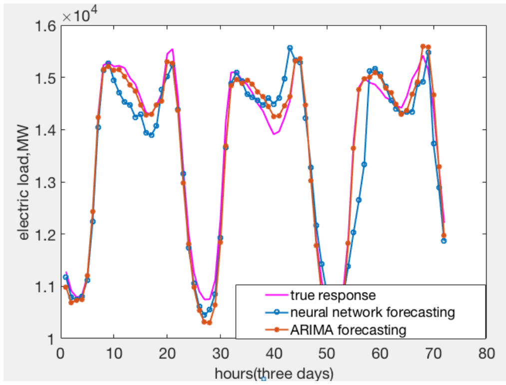


*Figure 10 Forecasting electric load for three days using NARX and ARIMA*

From Figure 10 it can be seen that both ARIMA and NARX provide a relatively good accuracy, and ARIMA is slightly better for short-term forecasting. However, for long term forecasting, ARIMA performs much worse than NARX as shown in Figure 11. The reason is that ARIMA is a strictly time series model, so later day's forecasting will use previous days' prediction result. The error of former days will be accumulated to later days. Here MAPE is used as error measurement.

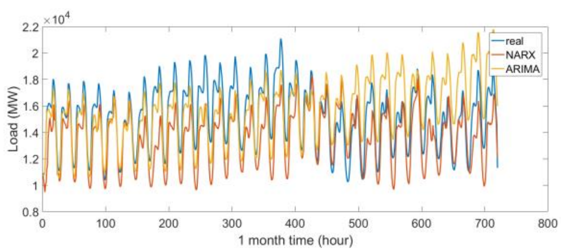


*Figure 11 Comparision between ARIMA and NARX*

Furthermore, it is necessary to consider whether reducing the training data size could still provide the same level of forecasting accuracy. The test data used here are selected carefully so that they contain weekdays (5/7 of the test data), weekends (2/7 of the test data), holidays, and days from four seasons. It is found that hourly data of one year is sufficient to produce relatively small error (MAPE around 6), while if the training data size is smaller than one year, error obviously increases as the training data size decreases. Figure 12 also indicates that ARIMA is more sentsitive to training data size, because it produces larger error than NARX when the training data size is limited. So NARX is more statiscically efficient than ARIMA.

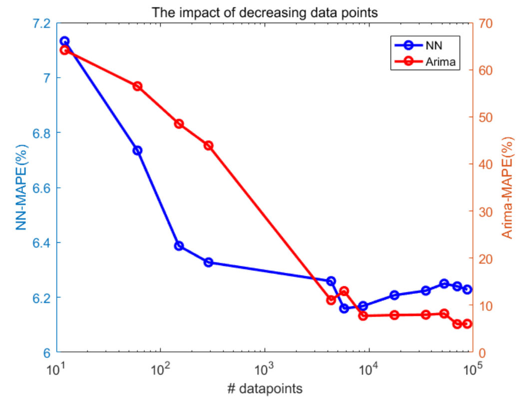


*Figure 12 Effect of the training data size on the error (MAPE)*

Current models, including NARX and ARIMA, perform unsatisfactorily on days where the data display more complexity. Electric load on Christmas, for example, is particularly difficult to predict due to the usual usage of electricity on that day. The forecasting on days like Christmas or other special days need to be improved to achieve better general forecasting accuracy. This can be done by appropriate feature engineering. Important factors to be considered include outlier removal, temperature, holiday, weekday or weekend, circular days (same day from different years) and seasonal periodicity. The test data used for featuring engineering contain weekdays, weekends, holidays, and days from different seasons, so that they can represent all considerations mentioned above. Features are added step by step, and the change of overall generalization error after traning with NARX is shown in Figure 13. MAPE after feature engineering is reduced to 4.9.

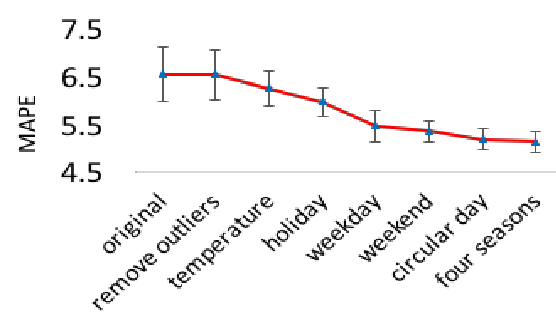


*Figure 13 Effect of Feature Engineering*

From Figure 13 it can be observed that temperature, distinction between holiday and non-holiday and distinction between weekday and weekend have considerable influence on the error, while outlier removal exerts less impact. Model performance is significantly enhanced after feature engineering. Figure 14 demonstrates that load forecasting on Christmas using NARX is much more accurate after feature engineering.

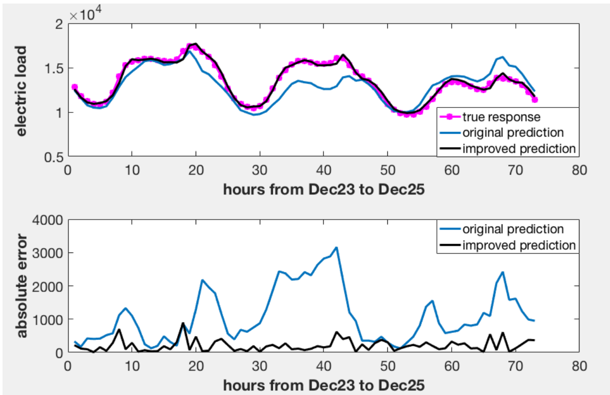


*Figure 14 Improved load forecasting during Christmas*

*3. Combination: Energy shortage supply estimation*

Based on load forecasting and wind power generation forecasting, energy supply shortage can be estimated, which is of great practical significance. From empirical data and literature, it is assumed that wind power generation can satisfy 20% of the energy demand [7]. The difference between normalized electric load estimation and 20% of normalized wind power generation estimation is considered to be energy supply shortage. The error of energy supply shortage estimation depends on both load forecasting error and wind power generation forecasting error. Energy supply shortage estimation for single day is shown in Figure 15. Estimation at 3am has higher level of uncertainty because both load forecasting and wind power generation forecasting have large error at that point.

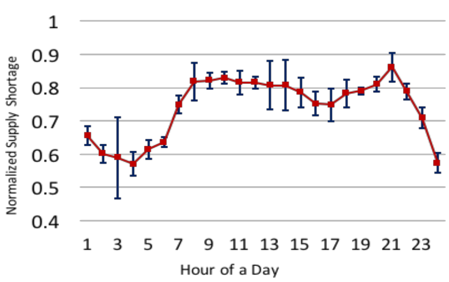


*Figure 15 Energy supply shortage estimation for singe day*

A 95% confident level is achieved that the error of energy supply shortage forecasting three days ahead is less than 24.73%. Wind power forecasting error has more impact on this confidence level, because generally wind power forecasting error is larger than electric load forecasting error.

The combination of electric load forecasting and wind power forecasting can give strategies to energy management system. When supply is predicted to be more than consuming, more storage unit could be prepared. On the other hand, if the supply cannot meet the customers' requirement, energy should be dispatched from other places. Because electric energy storage is still very expensive and limited, the forecasting can provide time to prepare and thus reduce the economic lost. This project gives a method of predating the energy supply shortage.

## CONCLUSION

This project finished both electric load and wind forecasting in several different methods. For electric load, ARIMA is better for very short term forecasting (3 days), but the computational cost is higher than that of NARX. NARX is relatively stable for both long and short term forecasting with minimum statistical or domain knowledge required. Although performance of current models on days where the data are more complex is less satisfactory, proper feature engineering can solve this problem and significantly improve the overall forecasting accuracy. As for wind forecast, the overall results are not as good as load forecast since response is more stochastic. But a combination of kNN and LSB can capture different pattern of the response and lead to a better result. For energy supply shortage, a 95% confident level is achieved that the error of forecasting three days ahead is less than 24.73%. Further research can focus on the wind power generation forecasting to further reduce the error. More data including other necessary features like air density is needed for better forecasting accuracy.


## ACKNOWLEDGEMENT

This report is part of final project in CS229. We would like to thank Professor John Duchi and Professor Andrew Ng for teaching this great class. We also want to thank all the TAs especially Francois Germain who provide tremendous help for this project.


## REFERENCES


[1] Mangalova, E. and Agafonov, E., 2014. Wind power forecasting using the k-nearest neighbors algorithm.International Journal of Forecasting, 30(2), pp.402-406.

[2] Matlab. Ensemble method. Retrieved November 20, 2016, from http://www.mathworks.com/help/stats/ensemble-methods.html#bsw8av_

[3] Dietz, S., 2011. Autoregressive Neural Network Processes-Univariate, Multivariate and Cointegrated Models with Application to the German Automobile Industry.

[4] Hong, W.C., 2013. Intelligent energy demand forecasting. New York: Springer.

[5] Hong, T. and Fan, S., 2016. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting, 32(3), pp.914-938.*

[6] Business, F. S. of. Statistical forecasting: notes on regression and time series analysis. Retrieved November 20, 2016, from https://people.duke.edu/~rnau/411home.htm

[7] Saidur, R., Islam, M.R., Rahim, N.A. and Solangi, K.H., 2010. A review on global wind energy policy. Renewable and Sustainable Energy Reviews,14(7), pp.1744-1762.