PREDICTIVE ANALYSIS IN ELECTRIC DISTRIBUTION GRID: CASE STUDIES ON SOLAR AND ENERGY PRICE FORECAST

A Thesis

by

YUANYUAN LI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Le Xie
Committee Members,	Robert S. Balog
	Tie Liu
	Steven L. Puller
Head of Department,	Miroslav M. Begovic

May 2017

Major Subject: Electrical Engineering

ABSTRACT

The electricity distribution system is undergoing profound changes as the society moves towards more sustainable utilization of energy resources. A common challenge in both supply and demand sides is how to provide accurate near term (within a day) forecast of the uncertainties to enable the distribution grid operation to modernize their decision making and deliver clean, affordable, and reliable electricity services.

This thesis focuses on the common challenge mentioned above, namely, how to improve the predictive capability for distribution system operators and load serving entities (LSEs). In particular, this thesis focuses on two of the major uncertain variables in future distribution grid: solar and electricity price forecast. Series of data-driven analysis are applied to develop efficient prediction models of these two variables. For the solar power generation prediction, the spatial temporal autoregressive model (ST ARX) is applied to the distribution system by including the neighboring data at nearby locations. Comparing to the benchmark models, the proposed model results in a better prediction accuracy and indicates the strong correlation between optimal neighboring distance and prediction time scale. As for the electricity price prediction, a comprehensive classification model based on decision tree algorithm is developed for the EnergyCoupon system. This algorithm is tested in Houston area with 10 customers and results in a good accuracy.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

Part 1, faculty committee recognition

This work was supervised by a thesis committee consisting of Professor Le Xie, Professor Tie Liu and Professor Robert Balog of the Department of Electrical and Computer Engineering and Professor Steven Puller of the Department of Economics.

Part 2, student/collaborator contributions

The related work depicted in Chapter I were conducted in part by EnergyCoupon Group of the Department of Electrical and Computer Engineering and were published in 2017.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

This work was made possible in part by National Science Foundation (NSF) under Grant Number ECCS-1546682 and CCF-1331863.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

The electricity distribution system is undergoing profound changes as the society moves towards more sustainable utilization of energy resources. There are two major paradigm changes that are taking place. First, from the supply side, much of the new generation resources are directly integrated at distribution level, such as solar and wind. Second, from the consumption side, passive end users are being transformed and incentivized to become active decision maker in the energy balance ecosystem. A common challenge in both supply and demand sides is how to provide accurate near term (within a day) forecast of the uncertainties to enable the distribution grid operation to modernize their decision making and deliver clean, affordable, and reliable electricity services. In particular, this thesis focuses on two of the major uncertain variables in future distribution grid: solar and power price forecast.

Enabled by advances in communication and computation technologies, enhanced forecast can be potentially obtained through more advanced statistical methods. For the power price forecast, an algorithm based on classification is proposed to predict the whole sale level electricity market prices. Such algorithm is an integral part of the larger context of engaging end users for demand response. As for the solar power generation forecast, an autoregressive model with exogenous input (ARX) model is proposed to improve the quality of forecast of distribution level solar generation.

CHAPTER II

ON-LINE ELECTRICITY PRICE PREDICTION ALGORITHM DESIGN FOR ENERGYCOUPON SYSTEM*

Introduction of EnergyCoupon System

Integration of heterogeneous energy sources in the US, as well as in many other countries in the world, is achieved using a wholesale level electricity market. In this market, aggregators (such as load serving entities (LSEs)) and generating companies trade their demand and supply under the supervision of Independent System Operators (ISOs), which results in real-time electricity price changes that temporally follow the diurnal variations. In the Texas market, which is ruled by the Electric Reliability Council of Texas (ERCOT), electricity can be traded one day in advance based on predictions of demand and supply, or in near-real-time with the fine grain data available every fifteen minutes. Usually, market participants take advantage of both day-ahead and real-time market rates to balance their demand and supply, and to optimize their profits.

The aggregators also participate in retail level markets to provide individual customers with different options of electricity services. An example of this is the retail market in several cities in Texas, in which over a hundred LSEs compete for residential customers via (slightly) different pricing plans. Retail customers typically pay a fixed or

* Chapter 2 includes parts of "EnergyCoupon: A Case Study on Incentive-based Demand Response in Smart Grid" which is under review for the Eighth International Conference on Future Energy Systems (ASM E-Energy) Hong Kong, 2017.

tiered fee, which typically does not allow real-time variations with respect to the wholesale price of electricity. Consequently, LSEs would benefit from an energy consumption shift from high-wholesale-price hours to low-wholesale-price hours of end users.

In order to induce savings for LSEs, we design and implement the EnergyCoupon system, which provides electricity usage targets to customers in real-time, measures their responses, awards coupons accordingly, and conducts a periodic lottery to reward users. In addition, the system also encourages the users to make more efficient use of energy by providing direct feedback of their total consumptions [1-3]. Figure 1 illustrates the architecture of the system, which consists of five parts, including three classes of functionalities (shown using different colors/shades), an SQL Database, and an Android/iOS App that forms the user interface.

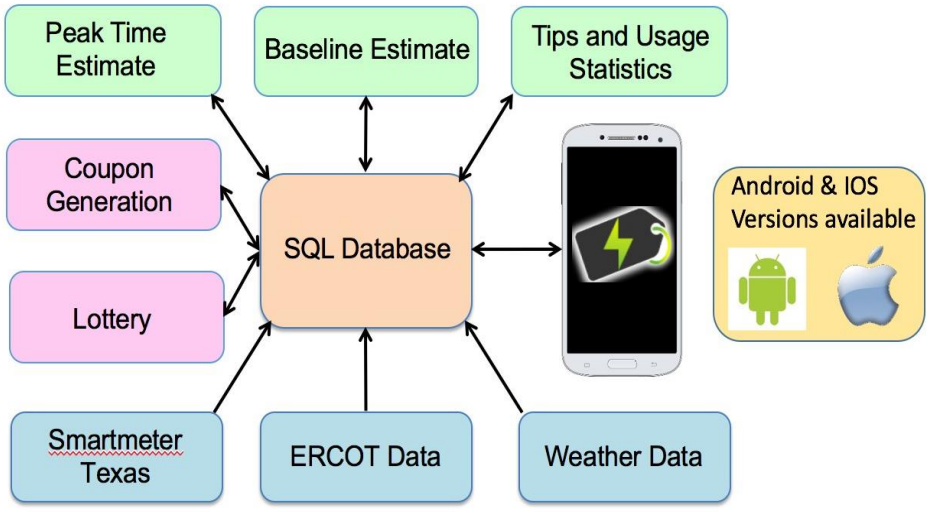


Figure 1. The architectural design of the EnergyCoupon system

The ultimate goal of the EnergyCoupon system was to incentivize end-consumers to shift loads from high-price hours to low-price hours in the wholesale market. The market clearing prices in the real-time market are only published in a real-time fashion which is not sufficient for the end-consumers in our system to react. Hence, we decided that the system should be capable of accurately predicting the electricity price at the near future, i.e. two-hours ahead. In addition, our system has several slightly different requirements for price prediction. For this project, we were more interested in identifying the electricity prices as being either 'high' or 'low', rather than obtaining or providing an exact-value prediction. Based on our proposed scheme, whenever the end-consumer shifted load from time period A to time period B, if the price in A was higher than it was have been in B, then it would yield extra savings for the LSE. From this perspective, we formulated the price-prediction issue in our system into a classification problem. We aimed to design a classifier that would determine the time periods in which the price is higher than at other times throughout the day. Since the price-prediction aspect of our system is an on-line algorithm that needed to run in real time, low-complexity-computing was essential. One of the simplest and widely used classifiers that fit our situation was the decision tree. In the following subsections, we will show the process of how we built the tree structure and include a series of analyses that will verify the performance.

Literature Review

There are many studies of price prediction published in recent years. Time series models are often used to forecast the numerical values of price. In [4] and [5], a time

series analysis is used to build autoregressive integrated moving average (ARIMA) models for forecasting next-day prices. The wavelet transform process is employed to decompose historical price series in [6] and the ARIMA model with an inverse wavelet transform is applied to forecast day-ahead electricity prices. A hybrid model of time series analysis followed by an adaptive wavelet neural network is employed in [7] to forecast PJM day-ahead market prices. Two optimized time series models, based on weighted least squares residuals, are applied in [8] to forecast the spot market price and the impact of predictive load and wind power generation are considered. Univariate and multivariate time series models are applied in [9] to forecast short- and mid-term base load prices, and the accuracy of models is considered to evaluate the impact of electricity prices and demand.

In addition to time series models, advanced machine-learning forecasting techniques are employed in the existing literature. The support vector machine (SVM) is employed in [10] to forecast the price values in Australia. In the work presented in [11], data series are separated by each trading interval and the genetic algorithm has been used to optimize parameters for the SVM based forecasting models. Two alternative SVM based models are proposed in [12] for classifying next-day electricity market prices in the market of Ontario and Alberta, Canada, according to pre-determined price thresholds. In order to optimize the SVM model parameters, a particle swarm technique is proposed in [13] to minimize a modified-prediction-intervals-based objective function. The support vector classification is applied in [14] to predict the occurrence of spike

price, followed by a forecast of the magnitude of the non-spike and spike prices in the Texas wholesale market by support vector regression.

The neural network (NN) method is also a commonly used forecasting method. A cascaded NN model and a NN model with an extended Kalman filter are employed respectively in [15] and [16] to predict the market clearing price. Moreover, an NN model with an adaptive wavelet transform has been proposed in [17] to improve the prediction accuracy of market clearing price. The extreme learning machine based single hidden layer feed-forward NNs are applied in [18] to improve the training speed of forecasting electricity prices. And, in [19], a recurrent NN realized by the Elman network model is proposed to forecast the electricity prices in Spain and New York.

Decision trees models are also useful data-mining tools designed to solve prediction problems. In [20], four decision tree models are employed to predict electricity prices for each submarket in the Brazilian market. In another study [21], decision trees are applied to classify future prices for the New York electricity market. In addition, regression tree models, including classification and regression trees, bagging, and random forests, are reported in [22] to predict prices in the Spanish market.

Electricity Price Prediction Algorithm Design Process

In this section, we introduce the design process for building the decision tree structure of price prediction for the EnergyCoupon system. Figure 2 shows the flowchart of the design process. Firstly, we conducted three data-driven analyses based on the historical market data, including a price threshold analyses, feature selection, and cost analyses in order to grow a decision tree model that meets the objectives of the

EnergyCoupon system and that will adapt well to the local market. Then, the price prediction model was trained using whole-year data, which was validated in the experiment. An analysis is provided at the end of this section to compare and evaluate the performance of the on-line price prediction model.

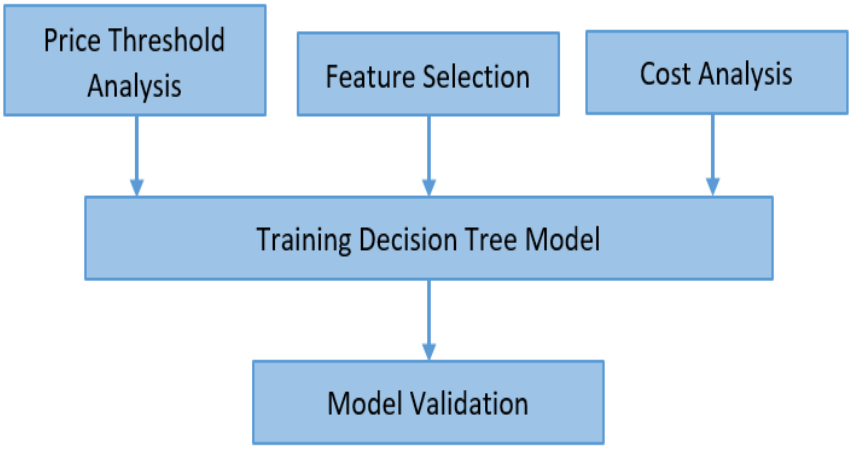


Figure 2. The flowchart of price prediction algorithm design

The historical data used for the data-driven analysis and model training is of the whole year of 2015 with a 15-minute resolution. For the model validation, we used the data from 1st July to 31st August in 2016. All the data were based on the Texas market in the Houston area.

A. Price threshold analysis

The price threshold analysis is aimed at finding a rational price boundary differencing the high price and low price and balancing the profits of both LSEs and household customers. A rigorous study of price threshold should involve an optimization

analysis of profits for both LSEs and individual customers, which requires some market data that was not available. Therefore, we have simplified the analysis by checking the average high price frequency in the summer days; the calculation formula is shown as follows:

$$Average\ Frequency\ per\ Day = \frac{Total\ Number\ of\ High\ Price}{Total\ Days} \quad (1)$$

Since the high price frequency is higher in the summer days than winter days due to the high electricity demand in the hot weather in Texas, the average high price frequency from May to October was used to estimate the coupon numbers received by the customers in the busy season. We used the real-time price data from May to October of 2015 for the analysis, and the results are shown in the Table 1. In Table 1, it indicates that the high price appears many times a day with a threshold of below 40 and becomes very rare at a threshold above 50. Meanwhile, based on the observations in the training stage, a lower threshold gives a higher fitting error. Thus, we chose the high price threshold to be 50 in the following study.

Table 1 Daily Average High Price Appearance by Threshold

Price Threshold ($/MWh)	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**
Average High Price	94	57	13	4	3	2	2	1

B. Feature selection

Considering the possible impact of weather and changes in demand to the real-time electricity price, it is necessary to include extra features to improve the accuracy of

price prediction. In this section, we check the cross-correlation of electricity price and four features (demand, temperature, humidity, and wind speed) to select the most significant features as the extra variables for the prediction algorithm. The analysis is based on the historical data from 2015. Since the data resolution is 15 minutes, one lag equals 15 minutes in the following plots. Figure 3 shows the self-correlation of real-time price data. The x-axis is the time lag from 0 to 96 corresponding to 0 to 24 hours. The y-axis identifies the correlation values. It is worth mentioning that the first point on the plot is P_{t-2h}, which indicates that the latest the price value can be included in the prediction is two hours prior to the predicted one. This is because the prediction time horizon is 2 hours ahead. The plot shows that the late values have stronger correlations to the predicted price than the early values. Therefore, in the price prediction algorithm, we include the last four features ($P_{t-8\,lag}$, $P_{t-9\,lag}$, $P_{t-10\,lag}$, $P_{t-11\,lag}$) in the algorithm.

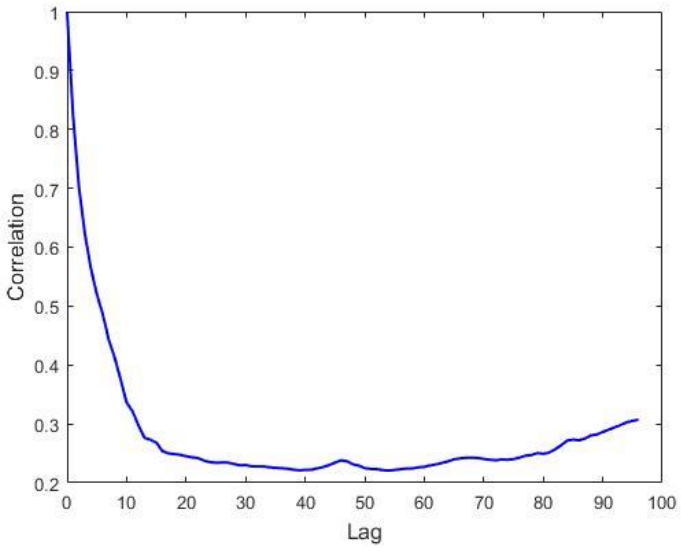


Figure 3. Self-correlation of electricity price

Figure 4 to Figure 7 show the cross-correlation between price and other features. The highest cross-correlation values in Figure 4 (cross-correlation between price and demand) and Figure 5 (cross-correlation between price and humidity) are lower than 0.5 and the correlations between price and temperature and wind speed features are even smaller. Although the correlations are not very strong between price and features, it is still reasonable to use them as extra variables in the prediction algorithm for two reasons. Firstly, the cross-correlations are calculated based on the values of the electricity price, while in the EnergyCoupon system the exact values of price prediction are not necessary. Thus, the relaxed prediction requirement makes the features significant to the price label prediction. Secondly, the prediction technique is pre-designed as a decision tree model, therefore the superfluous features added in the prediction algorithm will not affect the prediction results negatively, since they will not be selected as the nodes in the model training process. Based on the analysis above, we included the four most significant features of demand, temperature, humidity, and wind speed as the extra variables into the prediction algorithm. They are: Q_t, $Q_{t-1\ lag}$, $Q_{t-2\ lag}$, $Q_{t-3\ lag}$, T_t, $T_{t-1\ lag}$, $T_{t-2\ lag}$, $T_{t-3\ lag}$, $H_{t-47\ lag}$, $H_{t-48\ lag}$, $H_{t-49\ lag}$, $H_{t-50\ lag}$, $W_{t-88\ lag}$, $W_{t-89\ lag}$, $W_{t-90\ lag}$, $W_{t-91\ lag}$.

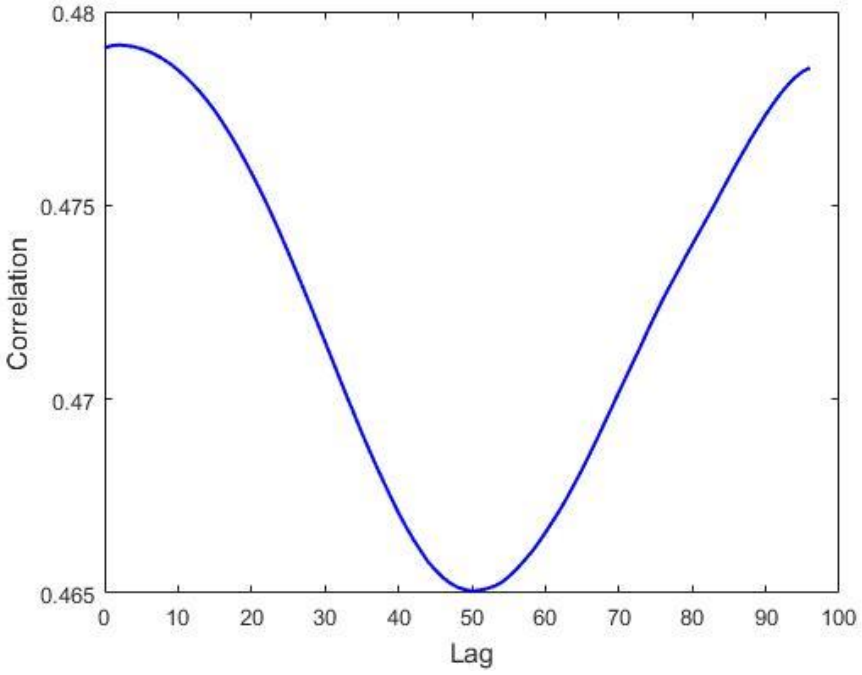


Figure 4. Cross-correlation between electricity price and demand

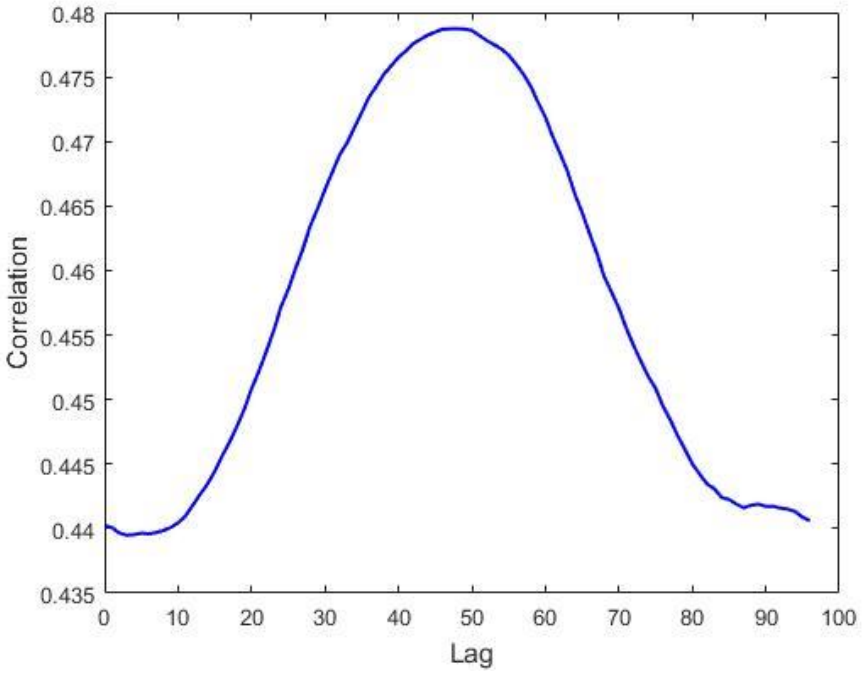


Figure 5. Cross-correlation between electricity price and humidity

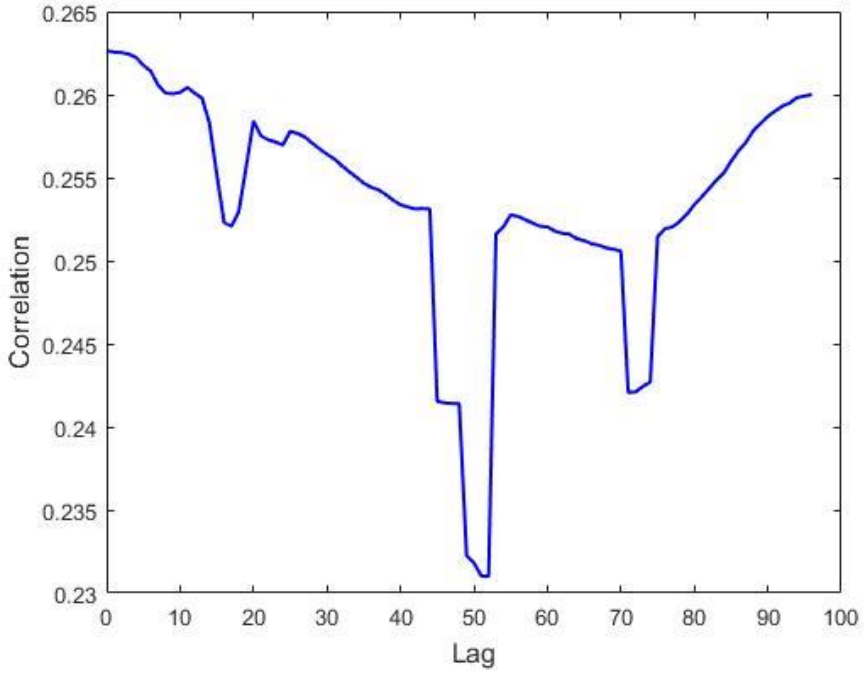


Figure 6. Cross-correlation between electricity price and temperature

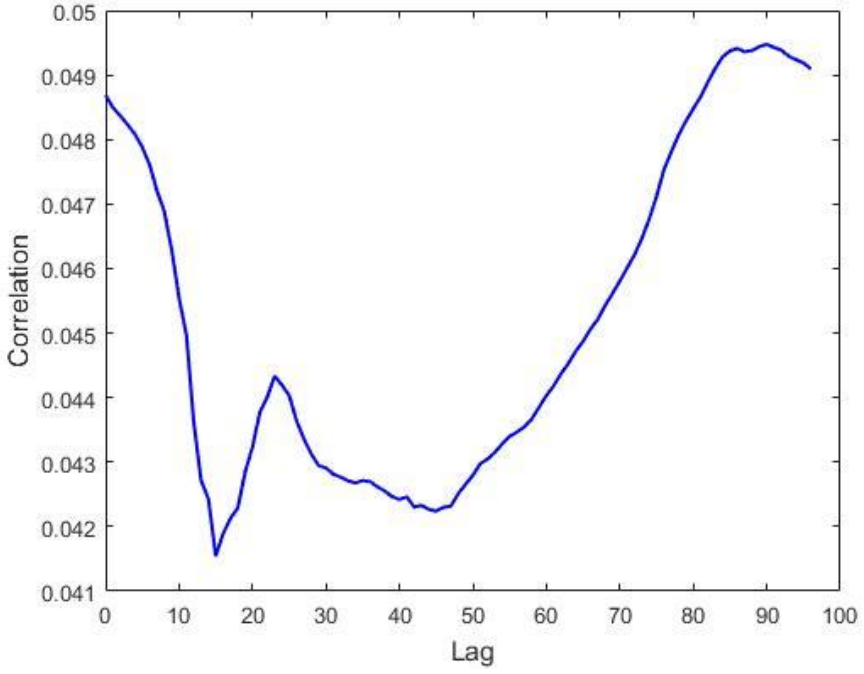


Figure 7. Cross-correlation between electricity price and wind speed

C. Cost analysis

Given the selected feature set, we construct a training data set containing around 30,000 data samples based on figures for 2015. Each sample is a 21-dimension vector including 20 features and one label. Before we trained the decision tree, two parameters, penalty ratio and minimal leaf size, needed to be determined. Penalty ratio is defined as the weight ratio of two kinds of errors: False Positive (FP) and False Negative (FN). In our EnergyCoupon system, FP error means the actual price is low but our prediction is high, and the FN error defines in the opposite way. Similarly, we define the terms True Positive (TP) and True Negative (TN) to capture the correct predictions respectively. The relationship between the four terms are as following:

$$FP + FN = Total\ Prediction\ Errors$$

$$TP + TN = Total\ Correct\ Predictions$$

$$FP + TN = Total\ Negative\ (Low\ Price)\ Samples$$

$$TP + FN = Total\ Positive\ (High\ Price)\ Samples$$

FP errors cause more coupons to be issued, and the FN errors cause misses in detecting the high-price events, which may cause a loss for the LSEs. Due to the natural properties of a lottery scheme, issuing more coupons to customers hurts the system only in a minimal way. Thus, it appears that we should put more weight on FN errors in order to detect more high price events. However, as we increase the weight of FN errors, the total error rate resulting from the cross-validation, which is a classic fitting performance check, increases as well. Thus, the determination of a penalty ratio results in a trade-off between the overall fitting performance and sensitivity, which is defined by the ratio:

$$Sensitivity = \frac{TP}{TP+TN} \quad (2)$$

When building decision trees, the minimal leaf size is a specific parameter that defines the minimal number of samples required for each leaf of the final tree. A larger minimal leaf size yields higher fitting errors. However, a smaller minimal leaf size has a higher risk of over-fitting. For example, an extreme case would be to build a tree that has the same number of leaves as the training data samples, which would give a zero fitting-error but would only fit the training data. The determination of a minimal leaf size requires a trade-off between fitting errors and a risk of over fitting.

In order to overcome the limitations described above, we wanted to choose a proper pair of the two parameters so that the sensitivity is high and the cross-validation error is low. Figure 8 illustrates the scatter plot of the pairs of parameters on sensitivity and cross-validation error plane. Each marker is a pair of the two parameters. We applied the filters in two dimensions: sensitivity ≥ 0.7 and cross-validation error ≤ 0.12, then chose the largest minimal leaf size over the resulting set to finalize the two parameters to be $minimal\ leaf\ size = 70$ and $penalty\ ratio = 1:8$.

D. Training and validation

After a series analysis of the history market data, the decision tree model is trained by the history data and validated by the on-line test. Table 2 shows the overall performance statistics for price prediction in both the fitting stage and the experiment stage.

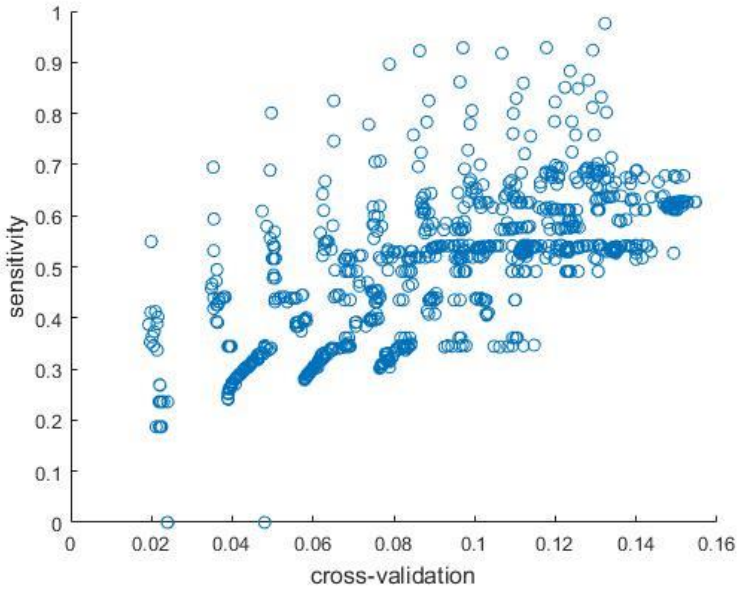


Figure 8. Scatter plot of (minimal leaf size and penalty ratio) pairs on sensitivity vs cross-validation error

Table 2. Price Prediction Performance

	TP	TN	FP	FN	Sensitivity	Error
Training	605	32,838	1,273	234	72.11%	4.31%
Validation	126	4,809	472	71	63.96%	9.91%

Note that in the training stage, there were approximately 30,000 samples, which covers the whole year of 2015, while in the testing stage, only approximately 5,000 samples were available, which cover the two-months period of the experiment. Observe that the error rate doubles in the testing stage and sensitivity decreases by about 8 %. Various issues could cause a performance drop of a classifier in testing. In our situation, one critical issue was that the electricity price could be affected by reasons other than the features we considered in our feature set, such as network congestions and gas price

drifts. Furthermore, we used the historical values of the features in the training stage, but used only the predicted values in the online testing, where the realizations are only revealed after the time had passed. The errors in the feature-value predictions are involved in the prediction errors as well.

Conclusions

In this chapter, we introduced the design process of our on-line price prediction algorithm based on a series data-driven analysis for the EnergyCoupon system, which is intended to incentivize the demand response of household users to shift their electricity usage from high price time periods to low price time periods. The designed price prediction algorithm was tested in the Houston area during the summer of 2016 with good performance.

In a future study, the accuracy of the price prediction algorithm could be improved by importing more related features or updating the algorithm technique. Additionally, more price categories will be considered as our next goal in order to provide more coupon options for customers.

CHAPTER III

SPATIO-TEMPORAL PREDICTION OF SOLAR IRRADIANCE FOR DISTRIBUTION GRID OPERATIONS*

Introduction of Solar Power Generation Prediction

Solar energy is one of the fast growing sources of renewable generation in the power system. In the United States, the cumulative photovoltaic (PV) installations surpassed 25GW by the end of 2015, due in part to continuing reduction in the installation cost [23]. However, integrating greater amounts of solar energy into the power system poses significant challenges due to its intermittent and variable nature. The integration of distributed solar energy such as rooftop PV in the distribution grid presents qualitatively different challenges compared to that at the transmission system level. In the transmission system, large scale solar power plants can be considered as non-controllable generation; while in the distribution system, the large number of rooftop PV installations impacts voltage management, protection coordination, transformer loading and other distribution system operations [24]. Recently, some sophisticated prediction models and methods have been proposed, which have a good prediction performance for large capacity solar power plants. However, rooftop PV power generation still cannot participate in the system planning and operations, due to the limited accuracy of the predictions of distributed PV power outputs. The power

* Chapter 3 is a slightly amended version of "Spatio-Temporal Prediction of Solar Irradiance for Distribution Grid Operations" which is under review for the IEEE Transactions on Power Systems.

generated by solar energy has a non-linear relation with the solar irradiance, temperature and some other weather factors, which have been widely utilized for both system models and prediction models [25-28]. However, at the distribution system level, since the change of weather factors is very small and predictable within a short time period and small geographical area, solar irradiance is the dominant factor for PV power generation, and thus an accurate solar irradiance prediction can contribute to a precise estimation of generated power.

Most prediction models in the previous studies rely on the meteorological features only, while it is worth noting that neighboring locations at the transmission system level show significant spatial correlations in the solar irradiance patterns [29]. This correlation suggests a good prediction model with spatial neighboring inputs, named the Spatio-temporal (ST) ARX model, which improves the accuracy of PV generation prediction at the transmission system level. With increasing installations of distributed PV by utilities and residents [23], the good performance of the ST ARX model shown in [29] inspired us to apply this model to the solar irradiance prediction for distribution system. Compared to the transmission system, in the distribution system the distance between accessible spatial neighboring data, which is the critical feature added in the ST ARX model, and the target prediction location cannot be very large. Due to this characteristic the impact of weather on the solar power generation at different locations can be excluded, since the weather conditions usually remain consistent within a narrow geographical region. Thus, variation in irradiance becomes the major cause of variations in power generation between different locations, which simplifies the transfer

function between irradiance data and solar power generation. Thus, in the distribution system, the distance of included spatial neighboring inputs should be limited to a much smaller range to produce a good prediction performance. Another significant difference between distribution system and transmission system is the impact of the local environment. Unlike the large scale solar PV power plant, which is usually located in an isolated area, the irradiance received by roof-top PV is easily affected by surroundings (i.e. the shade of the trees in the yard). Therefore, the averaged spatial neighboring data of several individual locations is added in the ST ARX model in order to minimize the effects of the individual local patterns.

In this chapter, I applied the spatial-temporal (ST ARX) prediction model proposed in [29-30] to solar irradiance prediction in distribution system with two improvements. Firstly, I assume that in the distribution system the farthest accessible distance of neighboring data is 30 km from the target location, which is about 0.7 times of the total distance from the northern-end to the southern-end of Austin area (location of our simulation data sets). Thus, spatial neighboring data included in the ST ARX model is limited within this distance. Secondly, the neighboring input in the prediction model is the averaged neighboring data from different directions, which may reduce the error caused by individual local irradiance pattern. I evaluated the prediction performance of the proposed ST ARX model by comparing it to a basic auto regression (AR) model, in order to verify the contribution of spatial neighboring data. In addition, a contribution analysis and discuss the optimal neighboring data distance for multi-time-scale prediction is provided. In addition, I evaluated spatial neighboring data from six

distance ranges for multi-time-scale predictions to show the relation between optimal neighboring distance and prediction time scale, which can help with optimizing the prediction performance of the ST ARX model in the application.

Literature Review

A number of approaches of solar irradiance prediction based on different time scales and applications have been discussed in the published literatures [31]. In recent years, many statistical models have been proposed for irradiance prediction based on the deep learning of historical data. These statistical models include linear models (time series models) and some non-linear models, such as Artificial Neural Network (ANN) and Support Vector Machine (SVM). Due to the intermittent nature of solar irradiance, linear models are not capable of providing a good prediction performance [32]. Comparing to linear models, non-linear models are more powerful in predicting the stochastic process and hence are widely discussed in the publications [33-37]. The boom of public data availability provides a possibility of promising development of statistical models. Another type of model is based on the cloud imagery and satellite data, which are applied for the short term prediction up to 5-hours-ahead. Physical models applied to the Satellite data were very popular in past several decades [38-39]. However, its unsatisfactory performance under certain weather conditions makes it unreliable for practical applications. Although the Satellite physical models are seldom discussed in the recent publications, the satellite data are valuable for prediction and were widely used by some statistical models [40-41]. Numerical Weather Prediction (NWP) models usually are operationally used for system planning [42-43]. These models are based on

the reproduction of physical phenomenon which can be very complicated and requires a high level of knowledge about the local environment.

Data Processing

In this section, I present the selection criteria details of solar irradiance raw data sets and check the validity of using averaged neighboring inputs after processing.

A. Data selection

The solar irradiance data used in this paper are taken from the National Solar Radiation Data Base [44]. All collected data sets are from locations in the Austin area from January 1st to December 31st 2014, and are at a 30-minute resolution. The spatial relations between the target location and its neighboring locations are shown in Figure 9. The red dot indicates the target location, and the stars are spatial neighboring data sets. A total of 30 neighboring locations from within 30 km of the target location are selected. The radius of 30 km is evenly divided into 6 segments of 5 km each, and 5 locations are selected in each segment. The details of physical locations of each data set can be found in the Appendix Table 9. In order to minimize the error caused by individual neighboring data sets, I averaged all data sets in the same distance segment, and considered this averaged input as the spatial neighboring input in the *ST ARX* model.

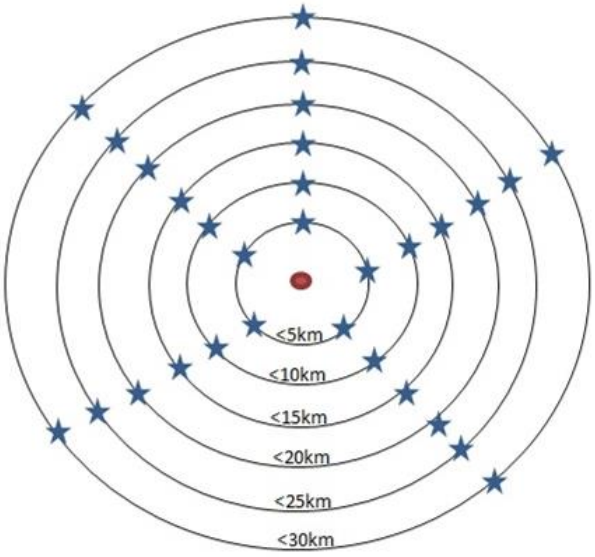


Figure 9. Target location and its neighboring data set

Because of the intermittent nature and impact of environmental factors, the solar irradiance data are not stationary, which makes it difficult to predict. Figure 10 shows the solar irradiance data from the target location in January 2014. The x-axis is the time stamp at 30-minute intervals and the y-axis is the Diffuse Horizontal Irradiance (DHI) values in W/m^2. I observe that the "normal days", during which solar irradiance data exhibit a good periodic pattern, can be modeled by time-series models. While the unexpected spikes on the "ramping days", which have no strong periodic solar irradiance patterns, may require extra meteorological features and stochastic models to estimate. In this paper, our simulations only consider the data from normal days, due to their strong periodic pattern. Accordingly, I designate the normal days as those where no DHI value higher than 150 W/m^2 occurs during the day which incidentally is about 1.1 times the averaged DHI value (130 W/m^2) during the daytime in whole year 2014.

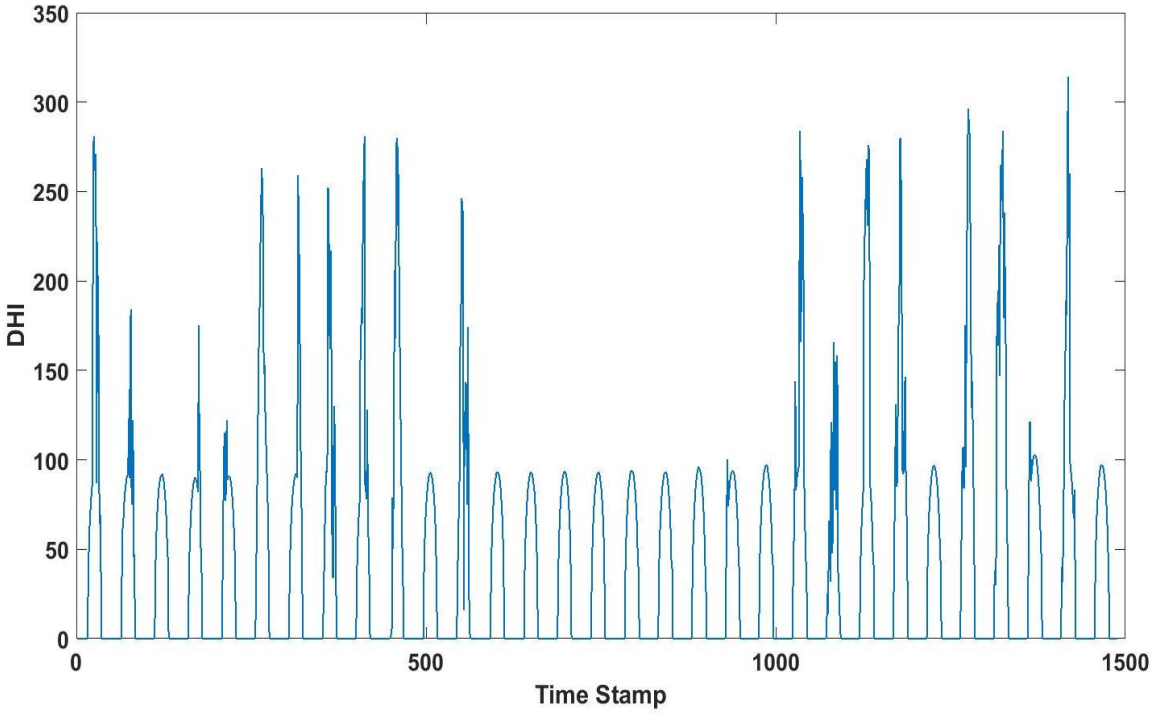


Figure 10. Solar irradiance data of target location in January 2014

From the supporting study I found out that classification of normal and ramping days can be very accurate based on the weather features. So a complete irradiance prediction process is designed as shown in Figure 11. According to the resulting different identifications, different solar irradiance prediction models will be applied. In this paper, our contribution and analysis are mainly limited to the normal days, which is part 2 in Figure 11.

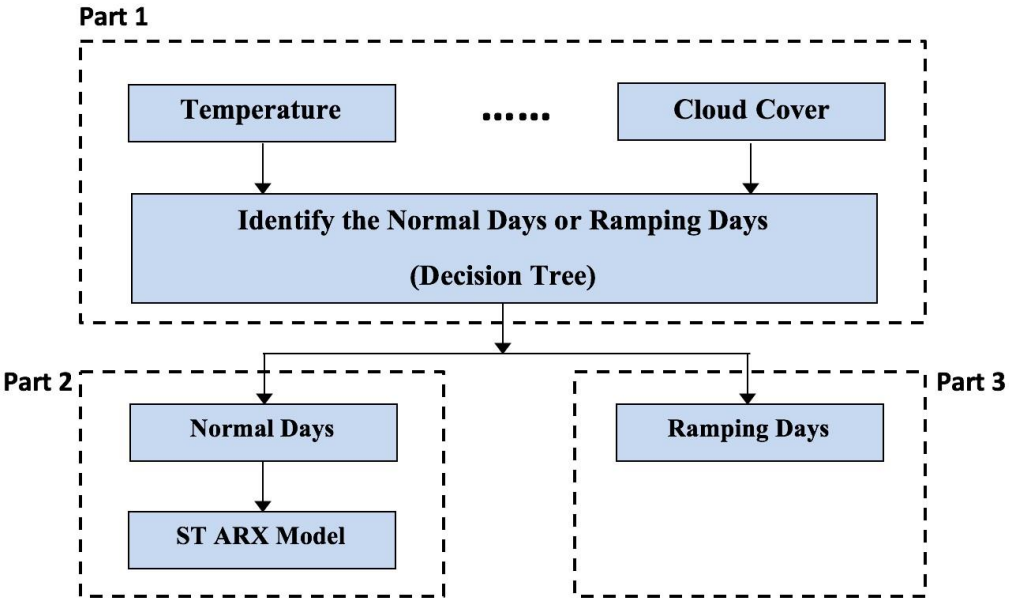


Figure 11. Flowchart of solar irradiance prediction process

B. Cross correlation check

The proposed *ST ARX* model is inspired by the strong cross correlation of solar irradiance data between neighboring locations, which is the critical condition to obtain good simulation results of the *ST ARX* model. Therefore, a cross correlation check of solar irradiance between averaged neighboring inputs from six distance ranges and target location are provided in Figure 12. The time lag in the figure is from 30 minutes to 2 hours corresponding to our multi-time-scale prediction study in Section 5. And all cross correlation values shown in Figure 12 are higher than 0.80 (details can be found in Table 7) which indicates strong correlations between our averaged neighboring inputs and target data.

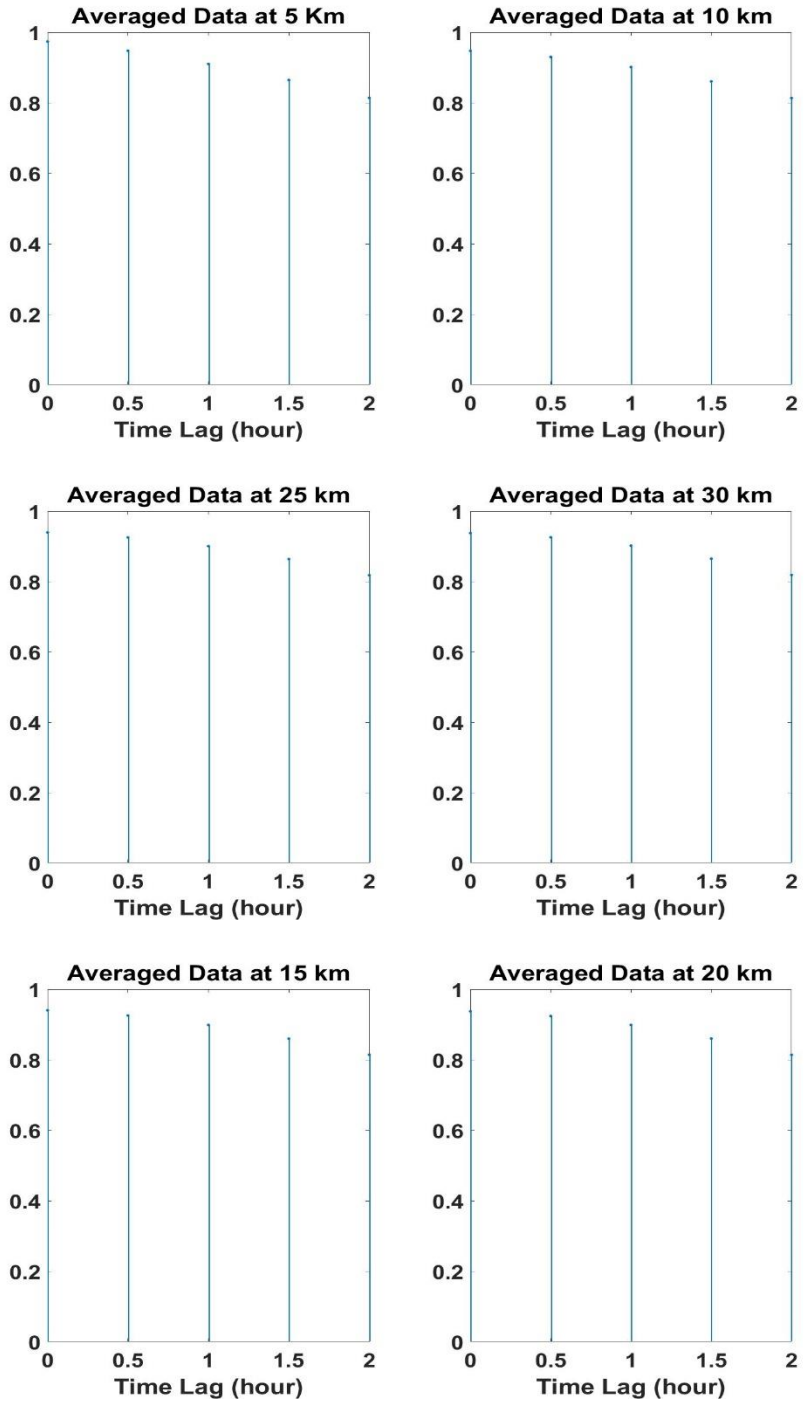


Figure 12. Solar irradiance cross-correlation between target location and averaged neighboring inputs

C. Benchmark model selection

In order to verify the contribution of spatial neighboring data to the prediction improvement and the superiority of the proposed *ST ARX* model, a basic time series model is chosen as the benchmark for the proposed model. Figure 13 shows the Autocorrelation Function (*ACF*) and Partial Autocorrelation Function (*PACF*) plots of the data at target location and each time lag corresponds to a 30-minute interval. The *ACF* decays with the lags and shows a periodic pattern; while *PACF* values sharply decrease to a very small number after lag 1. Table 3 shows the identification of time series model including autoregressive model (*AR*), moving average models (*MA*), and mixed autoregressive-moving average model (*ARMA*). According to the Table 3, the gradual decay of the *ACF* combined with the distinct cutoff of the *PACF* suggests that the *AR* model might be an appropriate benchmark for this data set. The mathematical details of the benchmark model are shown in next section.

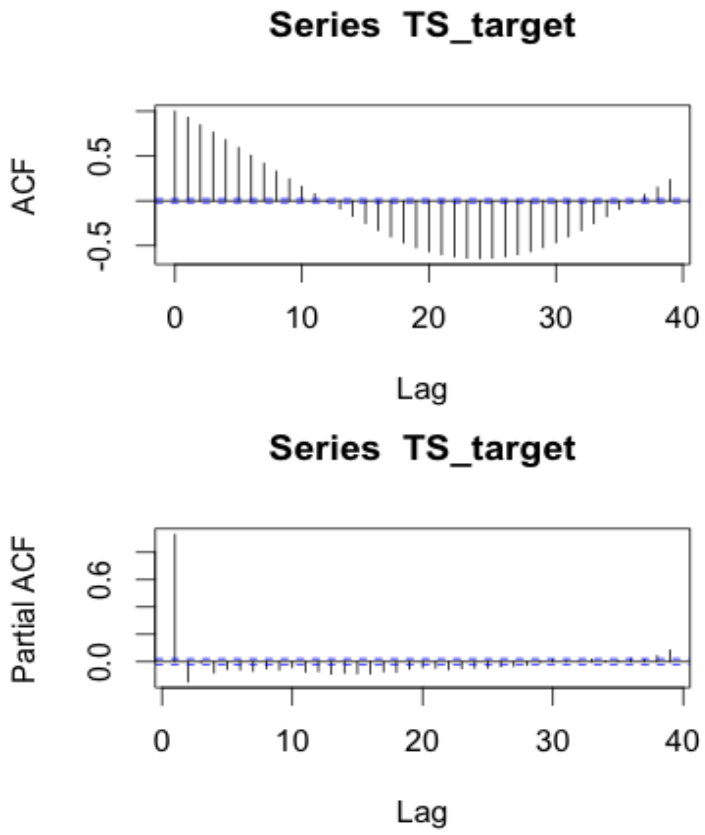


Figure 13. ACF and PACF plots of target data set

Table 3. Time Series Model Identification

Model	ACF	PACF
AR (p)	Decays	Cutoff after lag p
MA (q)	Cutoff after lag q	Decays
ARMA (p,q)	Decays	Decays

Evaluation of ST ARX Model

In this section, I introduce the *ST ARX* model for solar irradiance prediction and compare its performance to the basic *AR* model for 1-hour-ahead prediction.

A. Prediction model formulation

In order to show the contribution of neighboring data and the improvement of the prediction accuracy using *ST ARX* model, the basic *AR* model is used as the benchmark.

1) Basic AR model (Benchmark)

The basic *AR* model is a widely used prediction model with no spatial input features. Although it is easy to implement and has a good prediction performance for stationary data, it relies heavily on the latest historical data. The formulation of basic *AR* model is as follow:

$$I_{target}[t] = \sum_{n=1}^{i} \alpha_n \cdot I_{target}[t-n] + \varepsilon \quad (3)$$

Notations are shown in Table 4. A moving window process is applied, which updates the coefficients in every step after receiving the latest data point. The parameter n may vary based on the prediction time horizon and data resolution.

2) ST ARX model

The *ST ARX* model is developed based on the basic *AR* model including spatial neighboring data as an extra input, which is inspired by the strong correlations of solar irradiance data between nearby locations. The formulation of the *ST ARX* model for distribution system is as follow:

$$I_{target}[t] = \sum_{n=1}^{i} \alpha_n \cdot I_{target}[t-n] + \sum_{m=1}^{h} \beta_m \cdot I_{average}[t-m] + \varepsilon \quad (4)$$

Notations are shown in Table II. Further, I assume the time of sunrise and sunset can be known in advance, so I consider Iˆ equal to 0 W/m^2 from sunset to sunrise.

B. Forecasting metric

To compare the performance of *ST ARX* model and basic *AR* model, I use Root Mean Square Error (*RMSE*). The *RMSE* is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N}(\hat{I}[t]-I[t])^2}{N}} \tag{5}$$

where N is the total number of time steps of the time series data. The smaller values of *RMSE* indicate better prediction performance. Besides, the relative improvement is defined as:

$$Improvement = \frac{RMSE_{AR}-RMSE_{ST}}{RMSE_{AR}}\% \tag{6}$$

where $RMSE_{AR}$ is the *RMSE* value of the basic *AR* model simulation, and $RMSE_{ST}$ is the *RMSE* value of the *ST ARX* model simulation. I use the relative improvement value to show the advantages of *ST ARX* model in the following sections.

Table 4. Notations

I_{target}	Solar irradiance at target location
$I_{average}$	Averaged neighboring Solar irradiance
$\hat{I}$	Predicted solar irradiance
n	Index of historical solar irradiance at target location
m	Index of historical solar irradiance of averaged neighboring data
α	Coefficient of target variable
β	Coefficient of neighboring variable
ε	White noise

C. Summary of results

Table 5 shows the simulation results of 1-hour-ahead prediction using ST ARX model and basic AR model. I separately include the averaged neighboring inputs from six different distance ranges to show the robustness of the ST ARX model and verify the contribution of spatial neighboring inputs.

The values in the table indicate that including the spatial neighboring inputs in the model helps to improve the accuracy of solar irradiance prediction relative to the basic AR model, regardless the distance of included neighboring inputs within the 30 km range. Compared to the basic AR model, the most significant improvement of ST ARX model is about 5.3% with neighboring data at the 25 km to 30 km segment. And even for the worst case, the improvement is 3.1% at neighboring distance segment of 5 km to 10 km. The performance of the ST ARX model for solar irradiance prediction is much better than the AR model, which also verifies the significant contribution of spatial neighboring inputs. This conclusion can be extended to the 30-min-ahead and 2-hour-ahead predictions.

Table 5. RMSE Values of 1-Hour-Ahead Prediction

AR	**ST ARX Model**					
	Included Neighboring Input Distance (km)					
	<5	**<10**	**<15**	**<20**	**<25**	**<30**
20.15	19.52	19.53	19.21	19.16	19.27	19.09
Improve	3.1%	3.1%	4.7%	4.9%	4.4%	5.3%

Analysis of Spatial Neighboring Data Distance

From the results of last section we know that although including spatial neighboring data can promise improvement of prediction accuracy, the neighboring data at different distances may have different contributions to the accuracy improvement. In this section, I provide an analysis of the contribution of neighboring data from 5 km to 30 km for the 30-minute-ahead, 1-hour-ahead, and 2-hour-ahead prediction, and provide a general observation about optimal neighboring distance and prediction time scale. Besides, the difference between optimal distance and significant distance feature is discussed to stress the necessity of contribution analysis.

A. Contribution analysis for multi-time-scale prediction

In this section, the relation of prediction time-scale and distance of the neighboring inputs is investigated.

1) Optimal neighboring distance of multi-time-scale prediction

Table 6 shows the *RMSE* improvement of the *ST ARX* model for multi-time-scale prediction with six neighboring distance ranges. For 30-minute-prediction, the significant improvement is 2.1% with averaged spatial neighboring inputs at 15km to 20 km. Figure 14 (a) plots the *RMSE* value changes with the increasing distance of neighboring inputs. This result suggests that the neighboring data at 15 km to 20 km contains the most valuable spatial information of solar irradiance for 30-minute-ahead prediction in Austin area.

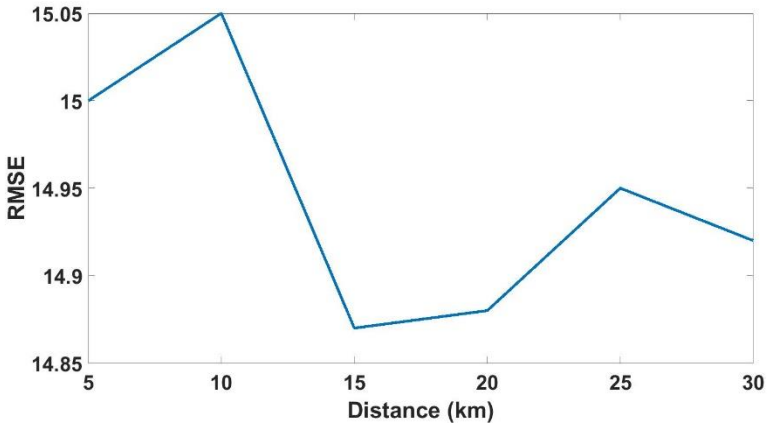
The improvement percentages in Table 6 indicate that the optimal distance of spatial neighboring data included in *ST ARX* model for 1-hour-ahead prediction is about

30 km, which is much longer compared to the 30-minute-ahead prediction. Figure14 (b) shows the *RMSE* changing by distance. The intuitive explanation of this result is based on cloud movement. The irradiance received by the solar panel is directly affected by the cloud behavior and this effect has geographical transitivity within a narrow area due to the cloud movement with wind. Therefore, for short time-scale prediction, the irradiance behavior at target location is similar to the nearby neighbors since the cloud at close distance moves to the target location after some time.

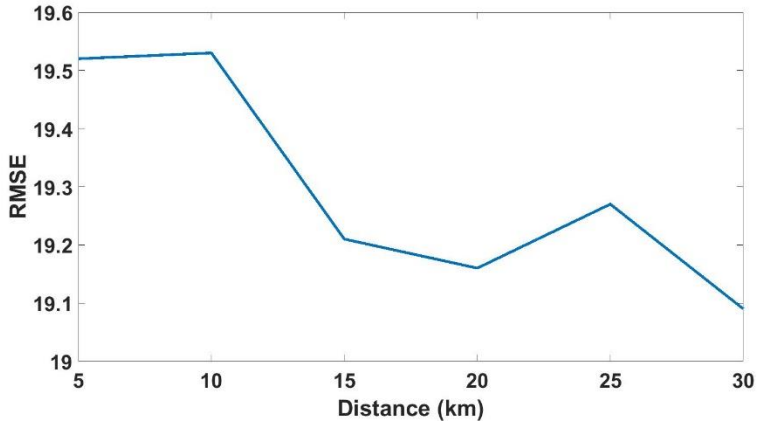
From the Table 6 and Figure 14 (c) we know that the most significant distance of spatial neighboring data in *ST ARX* model for 2-hour-ahead prediction is also around 30 km, which is the same to the 1-hour-ahead prediction. However, the relative improvement of 2-hour-ahead prediction with spatial neighboring data at significant distance is about 3.0% which is much lower than 1-hour-ahead prediction (5.3% in the Table 5). This result indicates that although the neighboring data at 30 km performs better than other close distance data, it might not be the optimal spatial inputs and the neighboring data at further distance may be more significant. Considering the service area of distribution system, I assume the longest accessible neighboring data distance is 30 km, so the neighboring data at further distance are not tested.

Table 6. Improvement Values of Multi-Time-Scale Prediction

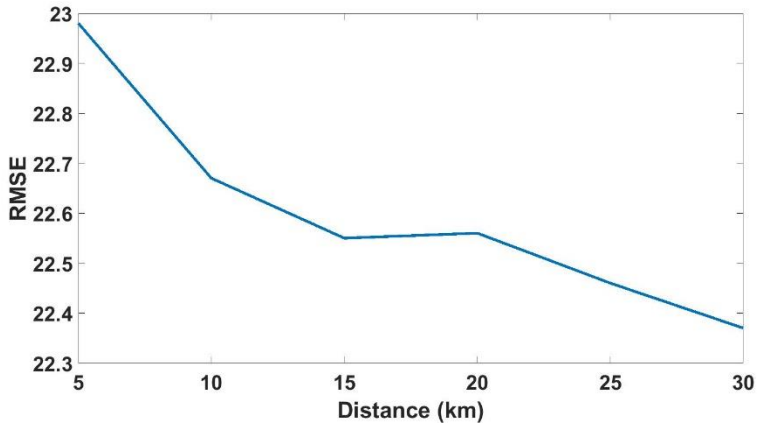
	ST ARX					
Time Scale	**Neighboring Input Distance (km)**					
	<5	**<10**	**<15**	**<20**	**<25**	**<30**
0.5h	1.3%	0.9%	2.1%	2.0%	1.6%	1.8%
1h	3.1%	3.1%	4.7%	4.9%	4.4%	5.3%
2h	0.4%	1.7%	2.3%	2.2%	2.6%	3.0%



a. 30-minute-ahead prediction



b. 1-hour-ahead prediction



c. 2-hour-ahead prediction

Figure 14. RMSE values of multi-time0scale prediction

2) Summary of results

Based on the discussion above, we have a general conclusion that the optimal distance of spatial neighboring data included in the ST ARX model is relatively long corresponding to the long time-scale prediction. The absolute significant distance for multi-time-scale prediction is not constant and may change based on the target locations, weather conditions, and local conditions at neighboring locations.

B. Discussion of optimal distance and significant feature distance

I defined the significant feature distance as the distance where the spatial neighboring inputs have the highest correlation to the target data. In other words, the significant feature distance can be identified from the results of the cross correlation check described in Section 3. The correlation values for different time lags and neighboring input distances are listed in Table 7. From the table we know that, for the 30-minutes time lag, the neighboring inputs at 5 km have the highest correlation to the target data, and for 1 hour and 2-hour time lag, the significant feature distance is 25 km. I find that the significant feature distance is not consistent with the optimal distance, and Table 8 compares these two distances for different time lags.

Table 7. Cross-correlation Values of Target Data and Neighboring Inputs

Time Lag	Neighboring Input Distance (km)					
	<5	**<10**	**<15**	**<20**	**<25**	**<30**
0.5h	0.947	0.929	0.924	0.924	0.928	0.926
1h	0.864	0.858	0.858	0.861	0.867	0.866
2h	0.816	0.813	0.815	0.817	0.824	0.822

Table 8. Comparison of Optimal Distance and Significant Feature Distance

Time	Optimal Distance	Significant Feature
0.5h	15km	5km
1h	30km	25km
2h	30km	25km

There are two possible reasons that cause this difference. Firstly, it is true that adding the most related feature in to the AR model may not result in the optimal prediction performance. To understand this, the calculation of coefficients in formula (4) should be clarified:

$$\alpha_n = corr(I_{target}[t], I_{target}[t-n]) \tag{7}$$

$$\beta_m = corr(I_{target}[t], I_{average}[t-m]) \tag{8}$$

Notations are shown in Table 4 and *corr* means correlation. From the formulas above we can know that the feature with strong correlation should have a heavy weight in the model, but cannot promise contribution to reducing the error. And the strong correlation can be caused by the collinearity of target data and neighboring inputs. The intuitive explanation is that the neighboring inputs at significant feature distance are too similar to the target data and cannot provide extra information to the prediction in order to improve the accuracy. According to this analysis, the optimal distance and significant feature distance are two different indicators. Secondly, the optimal distance is a data-driven result which is identified by prediction accuracy improvement. So the identification of optimal distance is easily affected by the quality and quantity of data set and difficult to be precise.

It is worth mentioning that both the optimal distance and significant feature distance verify the general conclusion that with increasing time scale the neighboring inputs at further distance are usually more preferable to be included in the prediction model.

Conclusions

In this chapter, I investigated the performance of the *ST ARX* model for solar irradiance prediction for distribution system operational planning. Our proposed model includes averaged spatial neighboring data within a narrowed distance range and provides a more accurate prediction when compared to the basic *AR* model. The analysis of neighboring data distance and contribution for multi-time-scale prediction concludes that longer time-scale prediction corresponds well with further optimal neighboring distance. And the optimal distance may not correspond to the distance of highest correlation.

The proposed *ST ARX* model is validated by the historical solar irradiance data for normal days which are defined as days where solar irradiance follows a periodic pattern. While for the ramping days, solely relying on the *ST ARX* model may not yield an accurate prediction and stochastic process analysis is needed to have a good understanding of irradiance spikes. Our future work will focus on the prediction model and contribution analysis of spatial neighboring data for ramping days.

CHAPTER IV

SUMMARY

In this thesis, there are two prediction problems in the distribution system are investigated. Data-driven analysis is applied to the history dataset, which contributes to parameter designs for prediction models. As for the power price forecast, series analyses are applied based on the market data to develop an efficient price prediction model for the EnergyCoupon system which is an application to incentivize the demand response in the distribution system. The designed prediction algorithm is tested in the Houston are for three months and resulted in an acceptable accuracy. The second investigation is to predict the solar power generation from individual houses in Texas area. Based on the results of data-driven analysis, the spatial relations between neighboring houses are included into the prediction model to improve accuracy. Besides, the consistency of the optimal neighboring data distance prediction time scale verifies the effects of the cloud movement to the household solar power generation.

REFERENCES

[1] A. Raruqui, S. Sergici, A. Sharif, “The impact of informational feedback on energy consumption-A survey of the experimental evidence,” Energy, vol. 35, pp. 1598-1608, April 2010.

[2] S. Darby. (2006). “The effectiveness of feedback on energy consumption,” [online]. Available: http://www.usclcorp.com/news/DEFRA-report-with-appendix.pdf.

[3] G. Wood, M. Newborough, “Dynamic energy-consumption indicators for domestic appliances: environment, behavior and design,” Energy and Buildings, vol. 35, pp. 821-841, September 2003.

[4] F. J. Nogales, J. Contreras, A. J. Conejo, and R. Espinola, “Forecasting next-day electricity prices by time series models,” IEEE Trans. Power Syst., vol. 17, no. 2, pp. 342–348, May 2002.

[5] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo, “ARIMA models to predict next-day electricity prices”, IEEE Trans. Power Syst., vol. 18, no. 3, pp. 1014-1020, Aug. 2003.

[6] A. J. Conejo, M. A. Plazas, R. Espínola, and A. B. Molina, “Day-ahead electricity price forecasting using the wavelet transform and ARIMA models,” IEEE Trans. Power Syst., vol. 20, no. 2, pp. 1035–1042, May 2005.

[7] L. Wu and M. Shahidehpour, “A hybrid model for price forecasting,” IEEE Trans. Power Syst., vol. 25, no. 3, pp. 1519–1530, Aug. 2010.

[8] T. Jonsson, P. Pinson, H. A. Nielsen, H. Madsen, and T. S. Nielsen, “Forecasting electricity spot prices accounting for wind power predictions,” IEEE Trans. Sustainable Energy, vol. 4, no. 1, pp. 210–218, Jan. 2013.

[9] K. Maciejowska and R. Weron, “Short- and mid-Term forecasting of baseload electricity prices in the U.K.: the impact of intra-day price relationships and market fundamentals,” IEEE Trans. Power Syst., vol. 31, no. 2, pp. 994–1005, March. 2016.

[10] J. H. Zhao, Z. Y. Dong, Z. Xu, and K. P. Wong, “A statistical approach for interval forecasting of the electricity price,” IEEE Trans. Power Syst., vol. 23, no. 2, pp. 267–276, May 2008

[11] L. Saini, S. Aggarwal, and A. Kumar, “Parameter optimisation using genetic algorithm for support vector machine-based price-forecasting model in national electricity market,” IET Gen., Transm., Distrib., vol. 4, no. 1, pp. 36–49, Jan. 2010.

[12] H. Zareipour, A. Janjani, H. Leung, A. Motamedi, and A. Schellenberg, “Classification of future electricity market prices,” IEEE Trans. Power Syst., vol. 26, no. 1, pp. 165–173, Feb. 2011.

[13] N. A. Shrivastava, A. Khosravi, and B. K. Panigrahi, “Prediction interval estimation of electricity prices using PSO-tuned support vector machines,” IEEE Trans. Ind. Informat., vol. 11, no. 2, pp. 322–331, Apr. 2015.

[14] P. Sarikprueck, W.-J. Lee, A. Kulvanitchaiyanunt, V. C. P. Chen, and J. Rosenberger, “Novel hybrid market price forecasting method with data clustering techniques for EV charging station application,” IEEE Trans. Ind. Appl., vol. 51, no. 3, pp. 1987–1996, May/Jun. 2015.

[15] L. Zhang, P. B. Luh, and K. Kasiviswanathan, “Energy clearing price prediction and confidence interval estimation with cascaded neural networks,” IEEE Trans. Power Syst., vol. 18, no. 1, pp. 99–105, Feb. 2003.

[16] L. Zhang and P. B. Luh, “Neural network-based market clearing price prediction and confidence interval estimation with an improved extended Kalman filter method,” IEEE Trans. Power Syst., vol. 20, no. 1, pp. 59–66, Feb. 2005.

[17] N. M. Pindoriya, S. N. Singh, and S. K. Singh, “An adaptive wavelet neural network-based energy price forecasting in electricity markets,” IEEE Trans. Power Syst., vol. 23, no. 3, pp. 1423–1432, Aug. 2008.

[18] X. Chen, Z. Y. Dong, K. Meng, Y. Xu, K. P. Wong, and H. W. Ngan, “Electricity price forecasting with extreme learning machine and Bootstrapping,” IEEE Trans. Power Syst., vol. 27, no. 4, pp. 2055–2062, Nov. 2012.

[19] S. Anbazhagan and N. Kumarappan, “Day-ahead deregulated electricity market price forecasting using recurrent neural network,” IEEE Syst. J., vol. 7, no. 4, pp. 866–872, Dec. 2013.

[20] J. C. Reston Filho, C. M. Affonso and R. C. L. Oliveira, “Pricing analysis in the Brazilian energy market: a decision tree approach,” in IEEE Bucharest PowerTech Conference, Bucharest, Romania, 2009.

[21] D. Huang, H. Zareipour, W. D. Rosehart, and N. Amjady, “Data mining for electricity price classification and the application to demand-side management,” IEEE Trans. Smart Grid, vol. 3, no. 2, pp. 808–817, Jun. 2012.

[22] C. Gonzalez, J. Mira-McWilliams, and I. Juarez, “Important variable assessment and electricity price forecasting based on regression tree models: classification and regression trees, Bagging and Random Forests,” IET Gen., Transm., Distrib., vol. 9, no. 11, pp. 1120–1128, Aug. 2015.

[23] Solar Energy Industries Association (SEIA) and GTM Research Team. “U.S. Solar Market Insight 2015 Q4,” [online]. Available: https://www.seia.org/research-resources/solar-market-insight-2015-q4

[24] National Renewable Energy Laboratry. (2013). “Metrics for evaluating the accuracy of solar power forecasting,” [online]. Avalaible: http://www.nrel.gov/docs/fy14osti/60142.pdf

[25] S. Chowdhury, G. A. Taylor, S. P. Chowdhury, A. K. Saha, Y. H. Song, "Modelling, simulation and performance analysis of a PV array in an embedded environment," IEEE 42nd International Universities Power Engineering Conference (UPEC), 2007.

[26] N. Mutoh, M. Ohno, and T. Inoue, "A method for MPPT control while searching for parameters corresponding to weather conditions for PV generation systems," IEEE Transactions on Industrial Electronics, vol. 53, pp. 1055-1065, June 2006.

[27] N. Sharma, P. Sharma, D. Irwin, P. Shenoy, “Predicting solar generation from weather forecasts using machine learning,” IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 528-533, 2011.

[28] E. Lorenz, J. Hurka, D. Heinemann, H. G. Beyer, "Irradiance forecasting for the power prediction of grid-connected photovoltaic systems," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 2, pp. 2-10, Mar. 2009.

[29] C. Yang, A. A. Thatte, and L. Xie, "Multitime-scale data-driven spatio-temporal forecast of photovoltaic generation," IEEE Transactions on Sustainable Energy, vol. 6, pp. 104-112, Jan. 2015.

[30] C. Yang, and L. Xie, "A novel ARX-based multi-scale spatio-temporal solar power forecast model," 2012 North American Power Symposium (NAPS), Champaign, IL, 2012, pp. 1-6.

[31] M. Diagne, M. David, P. Lauret, J. Boland, N. Schmutz, "Review of solar irradiance forecasting methods and a proposition for small-scale insular grids," Renewable and Sustainable Energy Reviews, vol. 27, pp. 65-76, Nov. 2013.

[32] C. Voyant, M. Muselli, C. Paoli, M. Nivet, "Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation," Energy, vol. 39, pp. 341-355, Mar. 2012.

[33] H. Y. Cheng, C. C. Yu, and S. J. Lin, "Bi-model short-term solar irradiance prediction using support vector regressors," Energy, vol. 70, pp. 121-127, Jun. 2014.

[34] S. Cao, "Total daily solar irradiance prediction using recurrent neural networks with determinants," Asia-Pacific Power and Energy Engineering Conference, Chengdu, China, 2010.

[35] A. Mellit, 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 Energy, 84(5), pp. 807-821, 2010.

[36] Z. Wang, F. Wang, and S Su, "Solar irradiance short-term prediction model based on BP neural network," Energy Procedia, vol. 12, pp. 488-494,2011.

[37] 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, 85(11), pp. 2856-2870, 2011.

[38] G. Dedieu, PY. Deschamps, and YH. Kerr, "Satellite estimation of solar irradiance at the surface of the earth and of surface albedo using a physical model applied to Metcosat data," Journal of Climate and Applied Meteorology, vol. 26, pp. 79-87, 1987.

[39] J. D.Tarpley, “Estimating incident solar radiation at the surface from geostationary satellite data,” Journal of Applied Meteorology, 18(9), pp. 1172-1181, 1979.

[40] A. Hammer, D. Heinemann, E. Lorenz, and B. Luckehe, “Short-term forecasting of solar radiation: a statistical approach using satellite data,” Solar Energy, 67(1), pp. 139-150,1999.

[41] O. Şenkal, and T. Kuleli, "Estimation of solar radiation over Turkey using artificial neural network and satellite data," Applied Energy, vol. 86, pp. 1222-1228, 2009.

[42] P. Mathiesen, J. Kleissl, “Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States,” Solar Energy, 85(5), pp. 967-977, 2011.

[43] S. Pelland, G. Galanis, G. Kallos, “Solar and photovoltaic forecasting through post-processing of the global environmental multiscale numerical weather prediction model,” Progress in Photovoltaics: Research and Applications, 21(3), pp. 284-296,2013.

[44] National Renewable Energy Laboratory, “National solar radiation data base,” [online]. Available: http://rredc.nrel.gov/solar/old_data/nsrdb

APPENDIX

Table 9. Physical Location of Neighboring Data

Customer	Latitude	Longitude	Distance (km)
Target	30.21	-97.74	0
	Distance < 5 km		
Customer 1	30.21	-97.70	4
Customer 2	30.21	-97.78	4
Customer 3	30.25	-97.74	3
	Distance < 10 km		
Customer 1	30.21	-97.66	8
Customer 2	30.13	-97.74	9
Customer 3	30.13	-97.78	10
Customer 4	30.25	-97.82	9
Customer 5	30.29	-97.74	10
	Distance < 15 km		
Customer 1	30.21	-97.58	15
Customer 2	30.13	-97.62	15
Customer 3	30.13	-97.82	12
Customer 4	30.25	-97.86	12
Customer 5	30.33	-97.70	14
	Distance < 20 km		
Customer 1	30.21	-97.54	19
Customer 2	30.90	-97.58	20
Customer 3	30.05	-97.82	19
Customer 4	30.25	-97.90	16
Customer 5	30.37	-97.70	18
	Distance < 25 km		
Customer 1	30.21	-97.50	23
Customer 2	30.01	-97.86	25
Customer 3	30.21	-97.98	23
Customer 4	30.41	-97.66	24
Customer 5	30.41	-97.78	23
	Distance < 30 km		
Customer 1	30.21	-97.42	30
Customer 2	30.05	-97.54	26
Customer 3	29.97	-97.86	29
Customer 4	30.21	-98.02	27
Customer 5	30.41	-97.54	29