Novel Hybrid Market Price Forecasting Method with Data Clustering Techniques for EV Charging Station Application

Piampoom Sarikprueck
Student Member, IEEE
Energy Systems Research Center
University of Texas at Arlington
Arlington, TX 76019, USA
piampoom.sarikprueck@mavs.uta.edu

Wei-Jen Lee
Fellow, IEEE
Energy Systems Research Center
University of Texas at Arlington
Arlington, TX 76019, USA
wlee@uta.edu

Asama Kulvanitchaiyanunt
Student Member, IEEE
Center on Stochastic Modeling, Optimization and Statistics
University of Texas at Arlington
Arlington, TX 76019, USA
asama.kulvanitchaiyanunt@mavs.uta.edu

Victoria C.P. Chen
Center on Stochastic Modeling, Optimization and Statistics
University of Texas at Arlington
Arlington, TX 76019, USA
vchen@uta.edu

Jay Rosenberger
Center on Stochastic Modeling, Optimization and Statistics
University of Texas at Arlington
Arlington, TX 76019, USA
jrosenberg@uta.edu

***Abstract*—In addition to provide charging service, Electric Vehicle (EV) charging station equipped with distributed energy storage system can also participate in the deregulate market to optimize the cost of operation. To support this function, it is necessary to achieve sufficient accuracy on the forecasting of energy resources and market prices. The deregulated market price prediction presents challenges since the occurrence and magnitude of the price spikes are difficult to estimate. This paper proposes a hybrid method for very-short term market price forecasting to improve prediction accuracy on both non-spike and spike wholesale market prices. First, Support Vector Classification is carried out to predict spike price occurrence and Support Vector Regression is used to forecast magnitude for both non-spike and spike market prices. Additionally, three clustering techniques including Classification and Regression Trees, K-means, and Stratification methods are introduced to mitigate high error spike magnitude estimation. The performance of the proposed hybrid method is validated with the ERCOT wholesale market price. The results from proposed method show significant improvement over typical approaches.**

***Index Terms*—EV Charging infrastructure, deregulated market, market price forecasting, support vector machine, data clustering.**

I. Introduction

Electric vehicles (EV) are currently promoted in the US and other countries for electrification of the transportation to improve the energy efficiency of the transportation sector and reduce the greenhouse gas emission. To promote the deployment and public acceptance of EV, it is necessary to reduce/eliminate the range anxiety of EV users. A well-planned fast (Level 3) charging infrastructure plays an important role for EV penetration. Therefore, one should consider the EV charging infrastructure from the regional point of view. In addition, it is desired to integrate renewable energy sources including wind and solar energy with electricity from power grid into EV charging station for sustainable future development [1,2].

The EV charging station with distributed energy storage system can also participate in deregulated market. Since the wholesale price of the electricity shows considerably volatility in the deregulated market, accuracy of market price prediction is one of the most important tasks to maximize the profit of the charging station.

Typically, the electric price forecasting method in the deregulated market can be separated into simulation and statistical approaches [3]. Though the simulation method can estimate market price accurately, it needs a lot of data from actual electrical models for simulation [4]. Therefore, the statistical approaches with artificial intelligence (AI) algorithms such as Neural Networks (NN) combined with Fuzzy c-mean [5,6], NN based on similar day method [7], and Autoregressive moving average [8, 9] have been commonly applied. All of them show sufficient forecasting accuracy but they normally can only predict non-spike electric price. A few hybrid models with classification algorithms such as Radial Basic Function NN and Support Vector Machines (SVM) [10, 11] have been conducted to estimate electric price both non-spike and spike prices conditions in deregulated market. However, the forecasting timeframes and training input parameters have not been described clearly in previous studies. Also, the spike price forecasting in these hybrid models is performed by only typical AI methods. These three important issues can significantly influence the electric price prediction performance.

This paper proposes a hybrid market price forecasting method (HMPFM) with data clustering techniques. The goal of clustering technique is to dissect spike prices in several ranges before performing the spike price magnitude

This paper is partially supported by NSF Grant ECCS-1128871

forecasting. This novel technique can improves the accuracy of spike price magnitude forecasting to enhance overall market price prediction. Since SVM has been proficiently conducted for predicting both classification and regression in various applications [12-15], Support Vector Classification (SVC) is adopted to predict spike price occurrence and Support Vector Regression (SVR) is used for market price magnitude prediction on both non-spike and spike prices. This paper implements three clustering algorithms including Classification and Regression Trees (CART), K-means, and Stratification methods because the Stratification method is the simplest clustering technique and CART and K-means approaches have been successfully applied for several research topics [16-19].

In this paper, the regional EV charging stations are considered to locate in Dallas/Fort Worth (DFW) metroplex. Electric Reliability Commission of Texas (ERCOT) takes the responsibility to serve electricity in this area and deploys 15-minutes time interval of market price. Therefore, the 15-minute ahead HMPFM with data clustering techniques is performed and validated with 2011 ERCOT wholesale market price data.

The rest of paper is organized following by modeling of the proposed regional EV charging stations in section II. Then, the framework of HMPFM with data clustering techniques is proposed in section III. Next, all of implemented algorithms including SVM, CART, K-means and Stratification methods are briefly described in Section IV. Finally, section V and VI are a case study to illustrate the proposed approach and conclusion, respectively.

II. Regional EV Charging Station System in ERCOT Deregulated Market

The goal of the proposed EV charging station design is to build a fast charging station equipped with distributed energy storage system that uses solar, wind energy, and electricity from power grid to simultaneously charge multiple EVs. The participation of this EV charging station system in the deregulated market highlights the benefit of wind and solar energy as well as distributed energy storage system in [1] with the optimal operational strategies. However, the operation charging station should be determined from the regional point of view to achieve global optimization. Hence, the proposed regional EV charging station system with n stations is shown in Fig. 1.

In this study, the regional EV charging station system are designed to build nearby the power nodes in DFW area represented by red circles in Fig. 2. These power nodes can have different nodal market prices at different locations and can serve as Point of Interconnection (POI) of DC fast (Level 3) charging between each charging station and power grid. ERCOT wholesale market prices [20] of these power nodes in July 2011 are depicted in Fig. 3. The normal market prices are less than 50 $/MWh; however, the spike prices are able to suddenly occur and their magnitudes can change suddenly from normal prices up to 2000 $/MWh. Moreover, the spike prices can happen either only one or several time durations. Because of these volatile scenarios in ERCOT nodal deregulated market, it is important for the regional charging station system to improve price forecasting accuracy to maximize its profit.

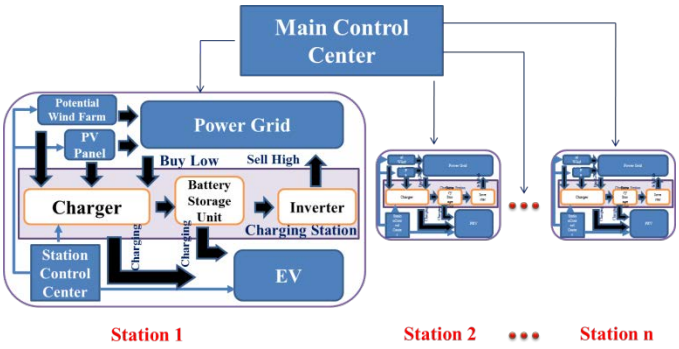


Fig 1 Configuration of EV Charging Infrastructure

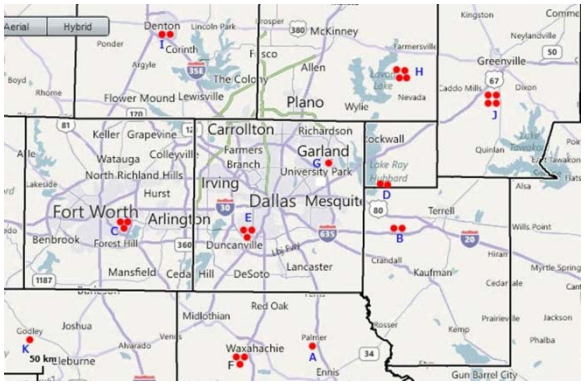


Fig. 2 Power nodes in DFW area

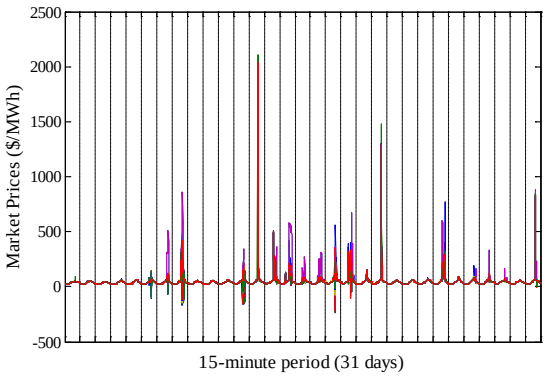


Fig. 3 DFW market prices in July 2011

III. Hybrid Method for Market Prices Forecasting

The framework of HMPFM with data clustering techniques is depicted in Fig. 4. There are two main stages of the proposed method including spike price occurrence and price magnitude predictions. First, the spike occurrence forecasting

is performed. If the result of this prediction is yes, the spike price magnitude prediction will be performed; otherwise, the non-spike price magnitude prediction is processed. Details in each process are described below.

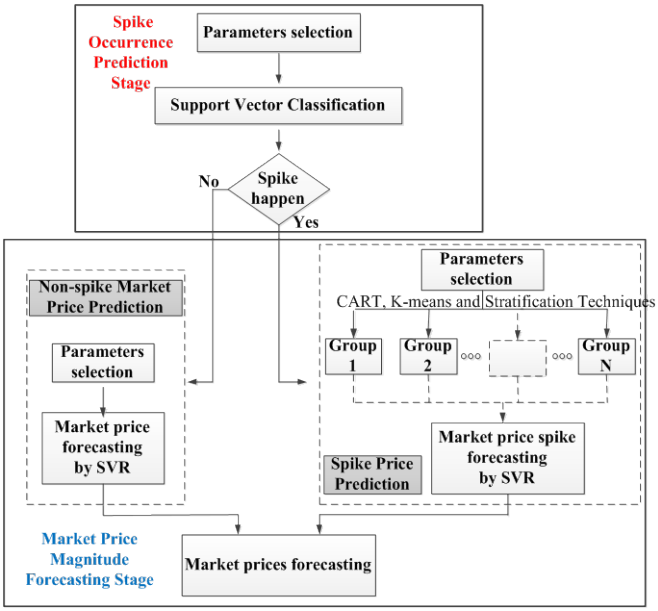


Fig. 4 The hybrid market price prediction framework

A. Spike Market Price Occurrence Prediction

According to several previous research works [10,11], there are three spike price definitions: 1) *An abnormal high price* is a price that is substantially higher than normal 2) *An abnormal jump price* is a difference between two adjacent prices that is greater than a given threshold 3) *A negative price* is where the price falls below zero.

An abnormal high price is the main focus in this paper. Levels of this type of spike price can be defined by statistical methods. [10, 11] show that it can either be calculated by either one standard deviation threshold or two standard deviation threshold. In order to escalate spike event number for improving the forecasting accuracy, the spike price is defined by a one standard deviation threshold and is calculated by (1) in this study.

$$spike = \mu \pm \sigma \tag{1}$$

where μ and σ are a mean and a standard deviation of market price, respectively (43.59 and 162.32 $/MWh for DFW market price in 2011).

The SVC is a selected algorithm to predict the spike price occurrence considering several impact parameters such as historical market prices, load profiles, etc. The spike price occurrence forecasting is performed for several models in this paper to identify model with the best performance.

B. Non-Spike Market Price Prediction

Due to the inconsiderable magnitude of non-spike price in 15-minute period, typical AI forecasting method can be adequately conducted to predict non-spike price condition. SVR is selected to estimate the magnitude of non-spike price considering the similar impact parameters as the spike price occurrence prediction. All spike prices are removed prior to perform the forecasting in several models in this process to identify model with the best performance.

C. Spike Market Price Prediction

Spike prices in DFW market fluctuate between less than -120 $/MWh and more than 3000 $/MWh in 2011[20]. Since this widespread distribution of spike prices can affect their magnitude estimation inaccuracy by typical AI forecasting approaches, clustering methods are introduced to divide spike prices into appropriate clusters before SVR performs their magnitude prediction. This paper implements three clustering algorithms including CART, K-means, and Stratification methods. The model with the best performance of various models considering impact parameters is obtained by performing the comprehensive HMPFM with this three proposed data clustering techniques.

IV. Support Vector Machines and Data Clustering Techniques

A. Support Vector Machines (SVM)

SVM is a machine learning method that conducts the learning procedure by statistical theory. It can be separated into two groups consisting of the classification and regression methods. The basic concept of these two approaches [21] is briefly described as follows.

1) Support Vector Classification (SVC)

Fig. 5 (a) illustrates linearly separable of SVC along with hyperplane $w \cdot x + b = 0$. The definition of $x=(x_1, x_2, \ldots, x_l)$ is the total number of market price events, w is the vector and b is the scalar that define the characteristics of the hyperplane. Moreover, $y_i = +1$ and $y_i = -1$ represent non-spike and spike classes, respectively. Thus, two constraints regarding this two classes separable hyperplane are shown in (2) and (3).

$$w^T \cdot x + b \geq +1 \tag{2}$$

$$w^T \cdot x + b \leq -1 \tag{3}$$

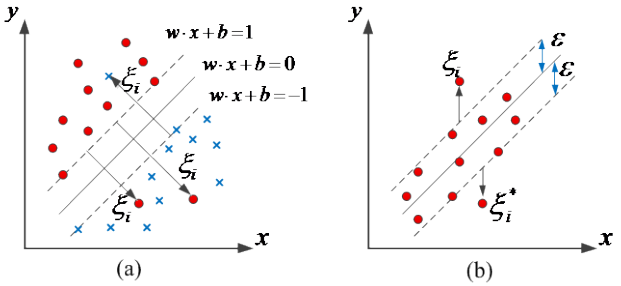


Fig. 5 Support Vector Machine (a) Classification and (b) Regression

The target of optimal separable hyperplane is to maximize the margin so the objective function and constraint of this problem become (4) and (5)

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \tag{4}$$

Subject to

$$y_i(w^T \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,...,l \text{ and } C > 0 \tag{5}$$

where C is a regularization parameter defined by the error penalty and ξ_i is a slack variable determined by the distance between incorrectly classified x_i and margin.

Lagrange multiplier is applied to solve (4) and (5). By solving the minimization problem, x_i becomes a dot product function. For nonlinear separable in high dimensional feature space, x_i can be mapped into $\phi(x_i)$ leading to a linearly separable problem. Kernel function is an efficient technique which is applied for solving this problem. In this paper, the radial basis function (RBF) kernel given the satisfactory SVM prediction performance [14, 15] is used to perform all of forecasting and can be described as (6).

$$K(x, x_i) = \langle \phi(x) \cdot \phi(x_i) \rangle = \exp(-\frac{\|x - x_i\|}{2\sigma^2})^2 \tag{6}$$

2) Support Vector Regression (SVR)

The concept of SVR is slightly different from SVC as shown in Fig. 5 (b). The loss function insensitive band (ε) and slack variable (ξ_i) are introduced and defined as cost of errors. To maximize margin, equation (7) and (8) describe objective function and problem constraints regarding ε and ξ_i . Techniques to remedy this regression problem are similar to classification solution by applying Lagrange multiplier and Kernel function as explained in the previous section.

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i + \xi_i^* \tag{7}$$

Subject to

$$\begin{aligned} &y_i - (w^T \cdot x_i + b) \leq \varepsilon + \xi_i, \\ &(w^T \cdot x_i + b) - y_i \leq \varepsilon + \xi_i^*, \\ &\xi_i, \xi_i^* \geq 0, i = 1,...,l \text{ and } C > 0 \end{aligned} \tag{8}$$

B. Data Clustering Techniques

1) Classification and Regression Trees (CART)

CART is a binary recursive partitioning clustering technique [22,23]. Target variables can be either categorical or continuous values in classification or regression scenarios, respectively. The clustering method in this paper focuses on regression technique since the magnitude of the spike price is considered continuous. Regarding the regression algorithm itself, two main important stages are carried out to determine optimal clusters including growing and pruning processes. In the former stage, CART ultimately enforces maximum possible terminal nodes from their parents by splitting rule as $x_i \leq d$. Thus, if predictor value (x_i) is less than or equal to a setting value (d), this variable will be a left children node member. Conversely, it will be assigned to right children node group. This rule is implemented with least square function and goodness of split as (9, 10) for growing optimal terminal nodes. In the latter stage, minimal cost tree by lowest mean square error is employed for pruning the generated tree from the first stage.

$$SS(t) = \sum (y_{i(t)} - \bar{y}_{(t)})^2 \tag{9}$$

$$\phi(t) = SS(t) - SS(t_R) - SS(t_L) \tag{10}$$

where $y_{i(t)}$ is the target of x_i in node t, $\bar{y}_{(t)}$ is the mean of target values in node t, $SS(t)$, $SS(t_R)$ and $SS(t_L)$ are sum square errors of the parent node, right children node, and left children node, serially, $\phi(t)$ is a goodness of split which shows the highest value for the best split.

2) K-means Clustering[24]

This algorithm separates d-dimensional vector space of data point (x_i), $D = \{x_i | i = 1,..,N\}$ into k partitions by minimize cost function as (11).

$$Cost = \sum_{i=1}^{n}(\arg\min_j \|x_i - c_j\|^2) \tag{11}$$

where c_j are k-centroid clusters in set $C = \{c_j | j = 1,...,k\}$

To reach the aim of cost minimization, this algorithm performs iteratively two-step procedures. First, c_j are initialized randomly and data points are assigned to the closest centroid by implementing a Euclidean distance function. Second, new c_j are computed by assigned data from the first step. This iteration is repeated until c_j are stabilized.

3) Stratification method

Employing this clustering technique is a simple process based on statistical data. To have sufficient data in each group, this technique divides d-dimensional vector space $D = \{x_i | i = 1,..,N\}$ equally into k clusters considering different target ranges that are different spike price ranges in this paper.

V. CASE STUDY

The regional EV charging station system is determined to be built near the power nodes in DFW area for level 3 DC fast charging. Since the ERCOT's wholesale market prices in each cluster in Fig. 2 are similar, only one set of market price is used at each cluster. Cluster E which us near Dallas is used to illustrate proposed market prices predicting method. First, correlation analysis is carried out to select input parameters for SVM process. Then, the HMPFM with data clustering techniques is implemented following the framework in section III. Finally, the comprehensive results are presented/discussed to verify prediction performance. The proposed approach is then applied to other power nodes to improve the forecasting accuracy for other EV charging station locations in DFW area.

A. Parameters Selection

Typically, one can obtain historical market prices, temperatures, and load profiles before performing 15-minute ahead market price forecasting while several factors such as generator contingencies and transmission constraints remain unknown prior to predict the market price. Other factors, such as fuel prices and day ahead load forecast, are less influence for very-short term market price forecasting. Therefore, correlation analyses of historical market prices, temperatures, and load profiles are studied. Temperature, load profile and electric price data are extracted from National Climatic Data Center (NCDC) [25] and ERCOT websites [20]. Fig. 6 depicts the correlation results between the market price and 15-minute until 12-hour time lags of three impact parameters.

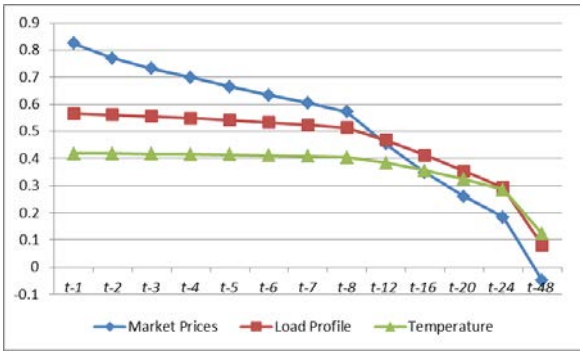


Fig. 6 Correlation analyses between market price and three impact parameters [(*t*-1),(*t*-2),…, (*t*-n) are 1,2,…,n prior times in 15-minute period.]

According to Fig.6, all correlations decrease significantly when prior times increase. Historical market prices show strong auto correlation with coefficients of greater than 0.7 till 1 hour before, so this parameter is decided as one important predictor. Moreover, both historical load profiles and temperatures give moderate correlations to market price with coefficient exceeding 0.4. Although these two parameters present less correlations than historical market prices, they are included as input parameters for further improving the forecasting accuracy.

B. Spike Market Price Occurrence Prediction

This paper introduces $P(in)$ and $P(out)$ given by (12) and (13) in order to specify spike occurrence prediction accuracy. These two indices provide classification precision of predicted spikes and incorrect classification of predicted non-spikes. The effective classification forecasting is determined by high $P(in)$ and low $P(out)$.

$$P(in) = P(\textit{correctly predicted spike} \mid \textit{predicted spike}) \quad (12)$$

$$P(out) = P(\textit{incorrectly predicted nonspike} \mid \textit{predicted nonspike}) \quad (13)$$

SVC is used to perform the spike price occurrence estimation in several models following these steps. First, due to the most significant impact of historical market prices corresponding strong auto correlation, they are selected to run spike price occurrence prediction for four time lag models. Second, the classification executes the forecasting separately for three time lags of temperature and load profile combined with the model with the best prediction performance from the first step. Lastly, the combination of the best prediction performance model of temperature and load profile obtained from the second step is evaluated in order to examine the possible classified performance improvement. Two-thirds of the year 2011 data in each month are employed for training while the remaining one third is used for testing. The spike price occurrence forecasting results are shown in table I.

TABLE I
SPIKE MARKET PRICE OCCURRENCE PREDICTION RESULTS

Models	*P(in)*	*P(out)*
mp(*t*-1)	0.73	0.0046
mp(*t*-1&*t*-2)	0.78	0.0046
mp(*t*-1,…,*t*-3)	0.77	0.0049
mp(*t*-1,…,*t*-4)	0.75	0.0049
mp(*t*-1&*t*-2)&*T*(*t*-1)	0.78	0.0046
mp(*t*-1&*t*-2)&*L*(*t*-1)	0.78	0.0052
mp(*t*-1&*t*-2)&*T*(*t*-1&*t*-2)	0.78	0.0046
mp(*t*-1&*t*-2)&*L*(*t*-1&*t*-2)	0.80	0.0048
mp(*t*-1&*t*-2)&*T*(*t*-1,…,*t*-3)	0.78	0.0046
mp(*t*-1&*t*-2)&*L*(*t*-1,…,*t*-3)	0.78	0.0049
mp(*t*-1&*t*-2)&*T*(*t*-1)& *L*(*t*-1&*t*-2)	0.80	0.0048

mp is a market price, *T* is a temperature and *L* is a load profile

A Significant low *P(out)* in table I is a result of high non-spike and low spike price number compared to total number of testing data. Following the procedure above, the model of historical market prices are simulated from *mp*(*t*-1) until *mp*(*t*-1,..,*t*-4). Spike price occurrence prediction by model of *mp*(*t*-1&*t*-2) yields the best result compared to other models with highest *P(in)* and lowest *P(out)* of 0.78 and 0.0046, respectively. Then, this model combined with *L*(*t*-1&*t*-2) enhances classification performance and provides the most accurate model compared to other combination models. This model is selected for spike price occurrence prediction in the HMPFM. In addition to improve classification performance, two adjustable parameters in SVC including Regularization (*C*) and Bandwidth (*B*) are tuned. Initial setting for *C* and *B* are 10 and 2, serially. Finally, the best parameter setting by *C* =5000 and *B* =20 elevates *P(in)* to 0.85 and stabilizes *P(out)* at 0.0046.

C. Non-spike Market Price Prediction

SVR is carried out to estimate magnitude of non-spike prices in the same way as the spike occurrence prediction. The forecasting performance is evaluated by Mean Absolute Percentage Error (*MAPE*) calculated by (14). The forecasting results are shown in table II.

$$MAPE = \frac{1}{N}\sum_{j=1}^{N}\frac{\left|P_j^{true} - P_j^{fst}\right|}{P_j^{true,N}} \quad (14)$$

where P_j^{true} is an actual market price at time *j*, P_j^{fst} is a forecasting market price at time *j* and $P_j^{true,N}$ is an average of recorded market prices over *N* period.

TABLE II
NON-SPIKE MARKET PRICE PREDICTION RESULTS

Models	*MAPE (%)*	Models	*MAPE (%)*
mp(*t*-1)	6.02	*mp*(*t*-1&*t*-2)&*T*(*t*-1&*t*-2)	5.93
mp(*t*-1&*t*-2)	5.94	*mp*(*t*-1&*t*-2)&*L*(*t*-1&*t*-2)	5.94
mp(*t*-1,…,*t*-3)	5.95	*mp*(*t*-1&*t*-2)&*T*(*t*-1,…,*t*-3)	5.93
mp(*t*-1,…,*t*-4)	6.02	*mp*(*t*-1&*t*-2)&*L*(*t*-1,…,*t*-3)	5.96
mp(*t*-1&*t*-2)&*T*(*t*-1)	5.94	*mp*(*t*-1&*t*-2)& *T*(*t*-1&*t*-2)&*L*(*t*-1&*t*-2)	5.92
mp(*t*-1&*t*-2)&*L*(*t*-1)	6.02		

The results in table II show the prediction performance of SVR. Temperature and load profile can enhance forecasting precision slightly. The model of *mp*(*t*-1&*t*-2) including *T*(*t*-1&*t*-2) and *L*(*t*-1&*t*-2) offers the best result with 5.92 % MAPE compared to the results of other models. This model is selected in the HMPFM for non-spike price estimation.

D. Spike Market Price Prediction

Three clustering techniques consisting of CART, K-means, and Stratification methods are utilized to enhance market price prediction in the deregulated market. This section presents clustering selection results of three proposed approaches prior to perform comprehensive HMPFM in the next stage.

1) Classification and Regression Trees (CART)

CART employs ten-fold cross validation considering historical market prices, temperatures, and load profiles as predictors and determining market price as target. Minimum numbers of target data in parent nodes are assigned from 10 to 70 and suitable numbers of data in each terminal node are one-third of the assigned number in parent nodes recommended by software inventor [25]. The optimal results appropriately specify different terminal nodes that are suitable number of clusters for each model. CART provides regression tree rules for each terminal node to settle proper clusters prior to performing spike prediction. Example regression tree rules of the model including *mp*(*t*-1) and *T*(*t*-1,…,*t*-3) are shown in table III. For instance, the rule for the 6th cluster is *mp*(*t*-1) fallen between 2086.89 and 3000.6 $/MWh.

TABLE III
EXAMPLE REGRESSION TREE RULES OBTAINED BY CART

Terminal Nodes	Rules
1	*mp*(*t*-1)<=816.95 and *T*(*t*-3)<=3.3
2	*mp*(*t*-1)<=816.95 and *T*(*t*-3)>3.3 and *T*(*t*-3)<=28.05
3	*mp*(*t*-1)<=275.22 and *T*(*t*-3)>28.05
4	*mp*(*t*-1)> 275.22 and *mp*(*t*-1)<=816.95 and *T*(*t*-3)>28.05
5	*mp*(*t*-1)> 816.95 and *mp*(*t*-1)<=2086.89
6	*mp*(*t*-1)>2086.89and *mp*(*t*-1)<=3000.66
7	*mp*(*t*-1)>3000.66

2) K-means

K-means clustering is performed to obtain proper clusters and is yielded separated input parameters for each group. Then, input parameters in each group are averaged to be the decision values. The lowest Euclidean distance calculated by (15) is carried out for selecting appropriate groups prior to predict magnitude of spike price. An example result from K-means of the model including *mp*(*t*-1,…,*t*-3) is shown in table IV.

$$d_n = \sqrt{\sum_{t=-1}^{-T} (X_{predicting(t)} - Y_{n(t)})^2} \qquad (15)$$

where d_n is a Euclidean distance for n^{th} cluster, X is an input parameter value, Y is an average decision value and T is a parameter at each several t prior times.

TABLE IV
4 CLUSTERS BY K-MEANS

Group	Group 1			Group 2		
Average decision values	*mp*(*t*-3)	*mp*(*t*-2)	*mp*(*t*-1)	*mp*(*t*-3)	*mp*(*t*-2)	*mp*(*t*-1)
	169.90	199.05	282.13	520.67	1042.90	2144.89
Group	Group 3			Group 4		
Average decision values	*mp*(*t*-3)	*mp*(*t*-2)	*mp*(*t*-1)	*mp*(*t*-3)	*mp*(*t*-2)	*mp*(*t*-1)
	2762.85	2929.57	2977.58	2655.70	2043.38	1059.57

As the results, one can see that K-means clustering is able to separate input parameters for each group effectively. All average decision values of input parameters are less than 282.13 and more than 2762.85 $/MWh in cluster 1 and 3, respectively. In addition, the average decision values of input parameters in cluster 2 give an increasing trend while they show a decreasing trend in cluster 4. The suitable number of clusters is discussed in the comprehensive results.

3) Stratification

The Stratification method equally dissects number of cluster members based on total spike price number. According to different levels of spike prices specified by dissection, input parameters are separated in the same category and time such as *mp*(*t*-1), *T*(*t*-1), etc. As with the K-means method, input parameters in each group are averaged to be the decision values. The lowest Euclidean distance defined by (15) is employed to select appropriate clusters before performing the prediction. Example result by four groups of the model including *mp*(*t*-1,…,*t*-3) is shown in table V. The proper number of clusters is discussed in the next section.

TABLE V
4 CLUSTERS BY STRATIFICATION METHOD

Group (no.of spike price)	Group 1 (66)			Group 2 (65)		
Range ($/MWh)	[-250,300)			[300-550)		
Average decision values	*mp*(*t*-3)	*mp*(*t*-2)	*mp*(*t*-1)	*mp*(*t*-3)	*mp*(*t*-2)	*mp*(*t*-1)
	141.66	184.34	250.40	435.48	379.74	371.56
Group (no.of spike)	Group 3 (69)			Group 4 (72)		
Range ($/MWh)	[550,2000)			[2000,3500)		
Average decision values	*mp*(*t*-3)	*mp*(*t*-2)	*mp*(*t*-1)	*mp*(*t*-3)	*mp*(*t*-2)	*mp*(*t*-1)
	634.93	661.46	739.53	2045.60	2270.62	2534.08

E. Comprehensive Results

The selected models from spike occurrence prediction and non-spike market price prediction are inducted to perform HMPFM combined with three proposed clustering techniques. CART computationally assigns the optimal number of clusters by software itself while the preliminary clusters for K-means and Stratification methods are set at four. Following the similar procedure for spike occurrence and non-spike price magnitude prediction, the results of spike price magnitude prediction are obtained from the comprehensive HMPFM tested in several models as shown in table VI. The prediction performance is evaluated by MAPE.

TABLE VI
COMPREHENSIVE MARKET PRICE FORECASTING RESULTS

Models	CART [MAPE (%)]	Models	K-means [MAPE (%)]	Models	Stratification [MAPE (%)]
mp(*t*-1)	15.65	*mp*(*t*-1)	15.86	*mp*(*t*-1)	16.00
mp(*t*-1&*t*-2)	15.76	*mp*(*t*-1&*t*-2)	16.69	*mp*(*t*-1&*t*-2)	15.68
mp(*t*-1,…,*t*-3)	15.87	*mp*(*t*-1,…,*t*-3)	15.75	*mp*(*t*-1,…,*t*-3)	16.30
mp(*t*-1,…,*t*-4)	15.87	*mp*(*t*-1,…,*t*-4)	15.83	*mp*(*t*-1,…,*t*-4)	16.17
mp(*t*-1) &*T*(*t*-1)	16.37	*mp*(*t*-1,…,*t*-3) &*T*(*t*-1)	15.32	*mp*(*t*-1&*t*-2) &*T*(*t*-1)	16.55
mp(*t*-1) &*L*(*t*-1)	16.63	*mp*(*t*-1,…,*t*-3) &*L*(*t*-1)	16.28	*mp*(*t*-1&*t*-2) &*L*(*t*-1)	16.56
mp(*t*-1) &*T*(*t* -1&*t*-2)	16.09	*mp*(*t*-1,…,*t*-3) &*T*(*t*-1&*t*-2)	15.40	*mp*(*t*-1&*t*-2) &*T*(*t*-1&*t*-2)	16.45
mp(*t*-1) &*L*(*t*-1&*t*-2)	16.50	*mp*(*t*-1,…,*t*-3) &*L*(*t*-1&*t*-2)	15.17	*mp*(*t*-1&*t*-2) &*L*(*t*-1&*t*-2)	16.48
mp(*t*-1) &*T*(*t*-1,…,*t*-3)	15.86	*mp*(*t*-1,…,*t*-3) &*T*(*t*-1,…,*t*-3)	15.32	*mp*(*t*-1&*t*-2) &*T*(*t*-1,…,*t*-3)	16.41
mp(*t*-1) &*L*(*t*-1,…,*t*-3)	15.30	*mp*(*t*-1,…,*t*-3) &*L*(*t*-1,…,*t*-3)	15.19	*mp*(*t*-1&*t*-2) &*L*(*t*-1,…,*t*-3)	16.43
mp(*t*-1) &*T*(*t*-1,…,*t*-3) &*L*(*t*-1,…,*t*-3)	15.29	*mp*(*t*-1,…,*t*-3) &*T*(*t*-1) &*L*(*t*-1&*t*-2)	15.50	*mp*(*t*-1&*t*-2) &*T*(*t*-1,..,*t*-3) &*L*(*t*-1,..,*t*-3)	16.33

According to table VI, each model provides similar market price forecasting accuracy by the three proposed clustering approaches. The models with best performance of spike price forecasting from the comprehensive HMPFM are *mp*(*t*-1)&*L*(*t*-1,…,*t*-3), *mp*(*t*-1,…,*t*-3)&*L*(*t*-1&*t*-2) and *mp*(*t*-1&*t*-2) for CART, K-means and Stratification methods, respectively. These predictions give the lowest MAPE with 15.29 %, 15.17 % and 15.68 %, serially. In addition, the number of clusters is adjusted from two to six to compare the optimum results in K-means and Stratification methods are shown in table VI. The maximum of six clusters is chosen for ensuring sufficient data in each group. The best prediction performance for K-means is the same as above while three clusters of Stratification method yields the lowest MAPE with 15.30 %.

To illustrate apparently improvement of the HMPFM combined with three clustering techniques compared to other prediction methods, Fig.7 depicts comparison results between the best cases of three proposed approaches and other general prediction methods including normal SVM (NSVM) and typical hybrid SVM (THSVM). MAPE reduce remarkably from 20.59 % and 16.95 % by NSVM and THSVM to about 15 % by three proposed methods for entire year results.

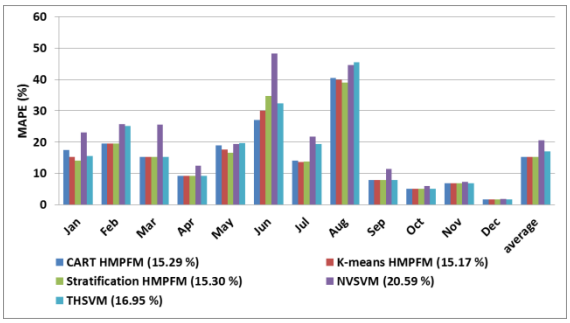
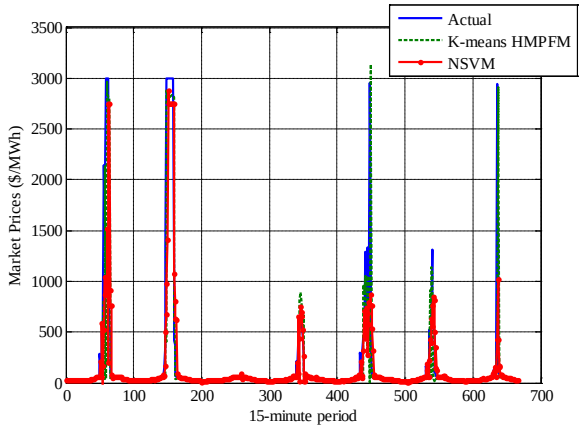
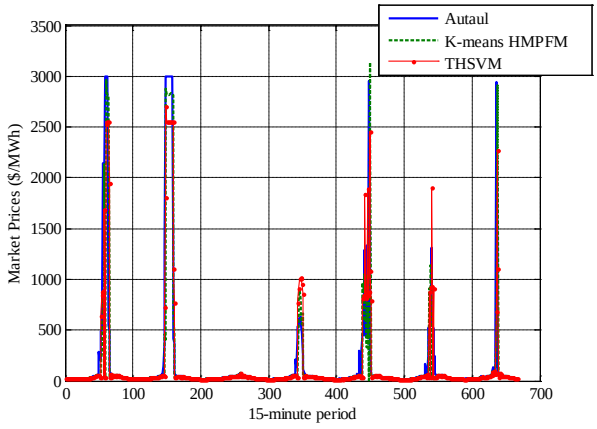


Fig. 7 The market prices forecasting comparison results of various approaches

As also shown in Fig. 8, since K-means HMPFM gives the most accurate result compared to the other two proposed data clustering techniques, it is applied to the proposed method, NSVM and THSVM for comparison. Three prediction methods yield comparable and satisfactory results of non-spike price estimation. Fig. 8 (a) shows that while the NSVM is not able to attain the spike price forecasting, the proposed approach can efficiently predict spike price occurrence and its magnitude. In addition, spike price magnitude prediction by THSVM provides more error than the forecasting by K-means HMPFM as depicted in Fig. 8 (b).



(a)



(b)

Fig. 8 The comparison of market price prediction results from proposed K-means hybrid SVM (a) with normal SVM (b) with typical hybrid SVM

Since the ERCOT wholesale market prices among different clusters are different, it is necessary to verify the performance of K-means HMPFM for all power nodes in DFW area to cover all locations of EV charging station system. Market prices for all power nodes are predicted by the proposed method and the prediction results are shown in Fig. 9. One can see that the proposed method yield similar results for all power nodes with average MAPE of 16.11 %.

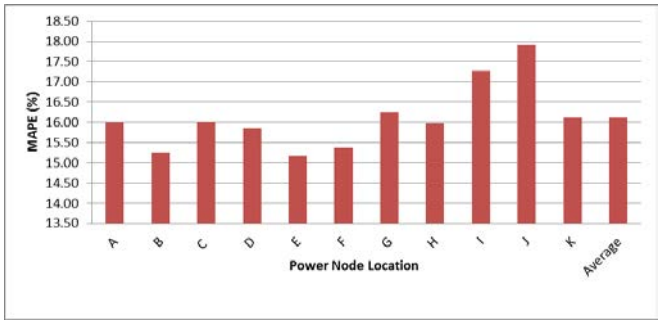


Fig. 9 Market price prediction results from K-means HMPFM for all power nodes in DFW area.

Although the K-means HMPFM provides acceptable results, there are still errors from the prediction. One approach to analyze the forecast uncertainty is the Martingale Model Forecast Evolution (MMFE) [26]. In the multiplicative model, MMFE determines the forecast change error as the log normal function by (16). An example forecast change error distribution of Dallas power node is depicted in Fig. 10. The uncertainty in stochastic cost minimizing problem for EV charging station system can be generated by this probability density function. The further study for optimal operation of regional EV charging station system applied the uncertainty function by K-means HMPFM will be focused in future work.

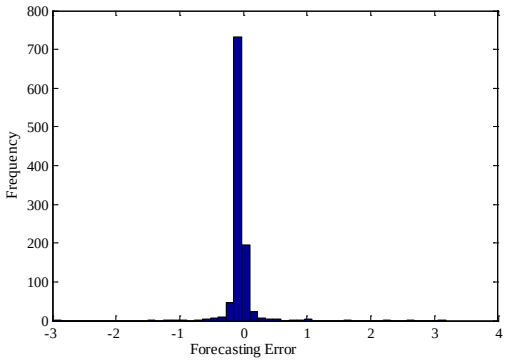


Fig. 10 Forecast change error distribution by MMFE

$$\varepsilon = \ln\left(\frac{MP_A}{MP_F}\right) \tag{16}$$

where ε is a forecast change error, MP_A and MP_F are actual market price and forecasting market price, respectively.

VI. Conclusion

This paper presents a novel HMPFM with data clustering techniques including CART, K-means, and Stratification methods to improve the accuracy of the wholesale electric price prediction in the deregulated market. The selected input models for SVM in spike price occurrence, non-spike and spike price magnitude estimations consider three historical impact parameters consisting of market price, temperature, and load profile. The proposed K-means HMPFM shows the effective prediction performance validated by ERCOT wholesale market price in DFW area. This proposed approach improves the prediction accuracy significantly compared to general market price prediction approaches. One can apply MMFE to evaluate the uncertainty by using probability density function of market price forecasting errors. This uncertainty can lead to stochastic optimization problem of the regional EV charging stations with distributed energy storage systems participated in the deregulated market in the future.

References

[1] F. Huang, P. Sarikprueck, Y. Cheng, and Wei-Jen Lee, "Design Optimization of PHEV Charging Station," *IEEE-IAS ICPS Annual Conference*, Louisville, KY, USA. 2012

[2] A. Kulvanitchaiyanunt, V. Chen, J. Rosenberger, Wei-Jen Lee, and P. Sarikprueck, "Control for a System of PHEV Charging Stations," *The Industrial and Systems Engineering Research Conference*, San Juan, Puerto Rico, 2013

[3] J. Bastian, J. Zhu, V. Banunarayanan, and R. Mukerji, "Forecasting energy prices in a competitive market," *IEEE computer application in Power*, Vol.12, issue 3, pp. 40-45, Jul. 1999

[4] C. P. Rodriguez, and G. J. Anders, "Energy Price Forecasting in the Ontario Competitive Power System Market," *IEEE Trans. on Power Systems*, Vol. 19, issue. 1, pp. 366-374, Feb. 2004

[5] Y.Y Hong, and C.Y. Hsiao, "Locational marginal price forecasting in deregulated electricity markets using artificial intelligence," *IEE Proc. Generation, Transmission and Distribution*, Vol. 149, issue 5, pp. 621-626, Dec. 2002

[6] K.Meng, Z.Y. Dong, and K.P.Wong, "Self-adaptive radial basis function neural network for short-term electricity price forecasting," *IET. Generataion, Transmission and Distribution*, Vol. 3, issue 4, pp. 325-335, Mar. 2009

[7] P. Mandal, T. Senjyu, N. Urasaki, T. Funabashi, and A. K. Srivastava, "A Novel Approach to Forecast Electricity Price for PJM Using Neural Network and Similar Days Method," *IEEE Trans. on Power Systems*, Vol. 22, issue. 4, pp. 2058-2065, Nov. 2007

[8] J. Contreras, R. Espínola, F. J. Nogales, and A. J. Conejo, "ARIMA Models to Predict Next-Day Electricity Prices," *IEEE Trans. on Power Systems*, Vol. 18, issue. 3, pp. 1014-1020, Aug. 2003

[9] 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. on Power Systems*, Vol. 20, issue. 2, pp. 1035-1042, May. 2005

[10] Y. Baez-Rivera, B. Rodriguez-Medina, and A. K. Srivastava, "An Attempt to Forecast Price Spikes in Electric Power Markets," *NAPS*, Carbondale, IL, USA, 2006, pp. 143-148

[11] J. H. Zhao, Z. Y. Dong, X. Li, and K. P. Wong, "A Framework for Electricity Price Spike Analysis With Advanced Data Mining Methods," *IEEE Trans. on Power Systems*, Vol. 22, issue. 1, pp. 376-385, Feb. 2007

[12] B. Ernst, B. Oakleaf, M. L. Ahlstrom, M. Lange, C. Moehrlen, B. Lange, U. Focken, and K. Rohrig, "Predicting the Wind." *IEEE Power & Energy Magzine*, Vol. 5, issue. 6, pp. 78-89, Nov. 2007

[13] R. Xu, H. Chen and X. Sun, "Short-term Photovoltaic Power Forecasting with Weighted Support Vector Machine," IEEE International Conference on ICAL, Zhengzhou, China, 2012, pp. 248-253
[14] Y.Liu, S.Jie, Y.Yang, and Wei-Jen Lee, "Short-Term Wind-Power Prediction Based on Wavelet Transform–Support Vector Machine and Statistic-Characteristics Analysis," IEEE Trans. on Industrial Applications, Vol. 48, issue. 4, pp. 1136-1141, May. 2012
[15] S.Jie, Wei-Jen Lee, Y.Liu, Y.Yang, and P.wang, "Forecasting Power Output of Photovoltaic Systems Based on Weather Classification and Support Vector Machines," IEEE Trans. on Industrial Applications, Vol. 48, issue. 3, pp. 1064-1069, Mar. 2012
[16] K. R. Skinner, D. C. Montgomery, G. C. Runger, J. W. Fowler, D. R. McCarville, T. R. Rhoads, and J. D. Stanley, "Multivariate Statistical Methods for Modeling and Analysis of Wafer Probe Test Data," IEEE Trans. on Semiconductor Manufacturing, Vol. 15, issue. 4, pp. 523-530, Nov. 2002
[17] M. Seera, C. P. Lim, D. Ishak, and H. Singh, "Fault Detection and Diagnosis of Induction Motor Using Motor Current Signature Analysis and a Hybrid FMM–CART Model," IEEE Trans. on Neural Networks and Learning Systems, Vol.23, issue.1, pp. 97-108 Jan. 2012
[18] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers," IEEE Trans. on Power Systems, Vol. 22, issue. 3, pp. 1120-1128, Aug. 2007
[19] P. K. Dash, S. R. Samantaray, and G. Panda, "Fault Classification and Section Identification of an Advanced Series-Compensated Transmission Line Using Support Vector Machine," IEEE Trans. on Power Delivery, Vol. 22, issue. 1, pp. 67-73, Jan. 2007
[20] ERCOT Market Information, [Online] Available: http://www.ercot.com/mktinfo/
[21] H. Xue, Q.Yang, and S. Chen, "SVM: Support Vector Machine," in *The Top Ten Algorithm in Data Mining*, Boca Raton, FL: A Chapman and Hall Book, 2009, ch.3, pp.37-59
[22] D. Steinberg, "CART: Classification and Regression Trees," in *The Top Ten Algorithm in Data Mining*, Boca Raton, FL: A Chapman and Hall Book, 2009, ch.10, pp.179-201
[23] Y.Yohannes, and P.Webb, *Classification and Regression Trees, CART: A User Manual for Identifying Indicators of Vulnerability to Famine and Chronic Food Insecurity*, International Food Policy Research Institute
[24] J. Ghosh, and A. Liu, "K-means," in *The Top Ten Algorithm in Data Mining*, Boca Raton, FL: A Chapman and Hall Book, 2009, ch.2, pp.21-35 National Climate Data Center, [Online] Available: http://www.ncdc.noaa.gov/
[25] D. Steinberg, and M. Golovnya, "CART 6.0:User's Guide," Salford Systems, San Diego, CA, 2007
[26] Heath, D.C., Jackson, "Modelling the evolution of demand forecasts with application to safetystock analysis in production distribution systems." Technical Report No. 989. School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY., 1991

BIOGRAPHY

Piampoom Sarikprueck (S'11) received the B.Eng and M.Eng. degrees from King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok, Thailand in 2002 and 2005, respectively. He is a Ph.D candidate in Electrical Engineering in the Energy Systems Research Center at the University of Texas at Arlington in Arlington, Texas USA. His research areas of interest include energy conversion, rotating electric machines, power system, renewable energy application and statistical forecasting.

Wei-Jen Lee (S'85-M'85-SM'97-F'07) received the B.S. and M.S. degrees from National Taiwan University, Taipei, Taiwan, R.O.C., and the Ph.D. degree from the University of Texas, Arlington, in 1978, 1980, and 1985, respectively, all in Electrical Engineering. In 1985, he joined the University of Texas, Arlington, where he is currently a professor of the Electrical Engineering Department and the director of the Energy Systems Research Center. He has been involved in research on power flow, transient and dynamic stability, voltage stability, short circuits, relay coordination, power quality analysis, renewable energy, and deregulation for utility companies. Prof. Lee is a Fellow of IEEE and registered Professional Engineer in the State of Texas.

Asama Kulvanitchaiyanunt (S'13) is a postdoctoral researcher at the University of Texas at Arlington. She holds a B.S. in Chemical Engineering from Chulalongkorn University, Bangkok, Thailand, a MS in Industrial Engineering from Lehigh University, Bethlehem, PA, and Ph.D. in Industrial Engineering from the University of Texas, Arlington. Her research interests include mathematical programming, dynamic programming, statistical modeling, machine learning, and optimization in energy application. She is a member of the Center on Stochastic Modeling Optimization and Statistics at UTA.

Victoria C. P. Chen is Professor and Interim Chair of Industrial and Manufacturing Systems Engineering at The University of Texas at Arlington. She holds a B.S. in Mathematical Sciences from The Johns Hopkins University, and M.S. and Ph.D. in Operations Research and Industrial Engineering from Cornell University. Dr. Chen is actively involved with the Institute for Operations Research and Management Science (INFORMS), having co-founded the INFORMS Section on Data Mining, and served as officers for the Section on Data Mining and Forum for Women in OR/MS. She is currently serving on the INFORMS Subdivision Council. Dr. Chen is a guest editor for the Annals of Operations Research and an associate editor for the journal Elemental Science of the Anthropocene. Dr. Chen's research utilizes statistical perspectives to create new decision-making methodologies that accommodate the uncertainty and complexity of real world problems. She has expertise in the design of experiments, statistical modeling, and data mining, particularly for computer experiments and stochastic optimization.

Jay Rosenberger is an Associate Professor of Industrial & Manufacturing Systems Engineering at the University of Texas at Arlington (UTA). He holds a B.S. in Mathematics from Harvey Mudd College, an M.S. in Industrial Engineering and Operation Research from the University of California at Berkeley, and a Ph.D. in Industrial Engineering from the Georgia Institute of Technology. His research interests include mathematical programming, applied simulation and optimization of statistical metamodel of complex systems. He has applied his methodological research to solve numerous real-world problems including those in transportation, healthcare, defense and energy. He is the Director of the Center on Stochastic Modeling Optimization and Statistics at UTA. He was the Vice Chair and Chair of the INFORMS Health Applications Section in 2007 and 2008, respectively. His graduate research on airline planning won the 2003 Pritsker Doctoral Dissertation award. Prior to joining the faculty at UTA, he worked in the Operations Research and Decision Support Department at American Airlines.