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

Piampoom Sarikprueck, *Student Member, IEEE*, Wei-Jen Lee, *Fellow, IEEE*, Asama Kulvanitchaiyanunt, *Student Member, IEEE*, Victoria C. P. Chen, and Jay Rosenberger

***Abstract*—In addition to providing charging service, an electric vehicle charging station equipped with a distributed energy storage system can also participate in the deregulated 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 nonspike 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 the magnitude for both nonspike 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 Electric Reliability Commission of Texas wholesale market price. The results from the proposed method show a significant improvement over typical approaches.**

***Index Terms*—Data clustering, deregulated market, electric vehicle (EV) charging infrastructure, market price forecasting, support vector machine (SVM).**

I. Introduction

ELECTRIC vehicles (EVs) are currently promoted in the U.S. 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 the EV charging station for sustainable future development [1], [2].

The EV charging station with a distributed energy storage system can also participate in the deregulated market. Since the wholesale price of electricity shows considerable 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]. Although 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 (NNs) 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 nonspike electric prices. A few hybrid models with classification algorithms such as radial basic function NN and support vector machines (SVMs) [10], [11] have been conducted to estimate the electric price (both nonspike and spike price conditions) in the 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 the clustering technique is to dissect spike prices in several ranges before performing the spike price magnitude forecasting. This novel technique can improve 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 nonspike and spike prices. This paper implements three clustering algorithms including classification and regression trees (CART), K-means, and stratification methods because the

Manuscript received June 25, 2014; revised September 19, 2014 and October 29, 2014; accepted November 25, 2014. Date of publication December 18, 2014; date of current version May 15, 2015. Paper 2014-ESC-0246.R2, presented at the 2014 IEEE Industry Applications Society Annual Meeting, Vancouver, BC, Canada, October 5–9, and approved for publication in the IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS by the Energy Systems Committee of the IEEE Industry Applications Society.

P. Sarikprueck and W.-J. Lee are with the Energy Systems Research Center, The University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: piampoom.sarikprueck@mavs.uta.edu; wlee@uta.edu).

A. Kulvanitchaiyanunt, V. C. P. Chen, and J. Rosenberger are with the Center on Stochastic Modeling, Optimization and Statistics, The University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: asama.kulvanitchaiyanunt@mavs.uta.edu; vchen@uta.edu; jrosenberg@uta.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIA.2014.2379936

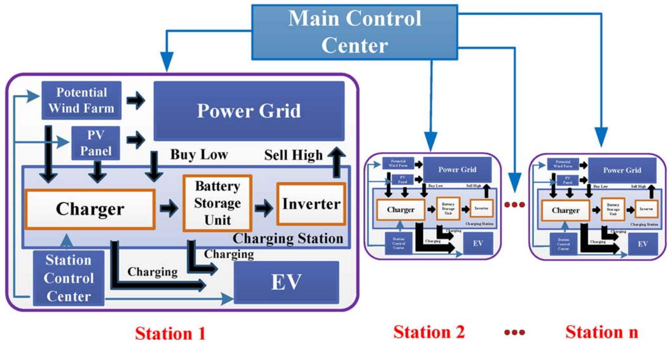

Fig. 1. Configuration of the EV charging infrastructure.

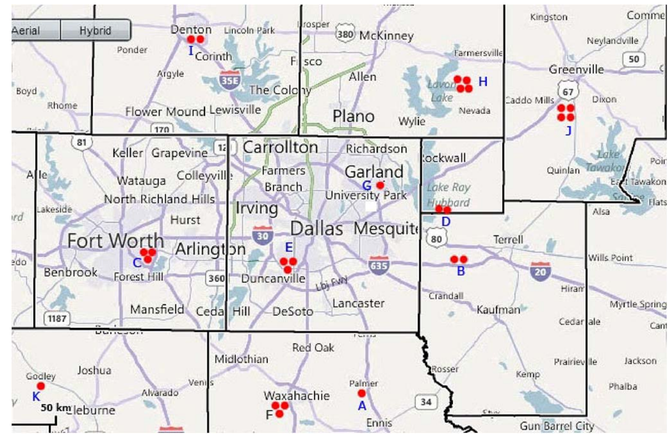

Fig. 2. Power nodes in the DFW area.

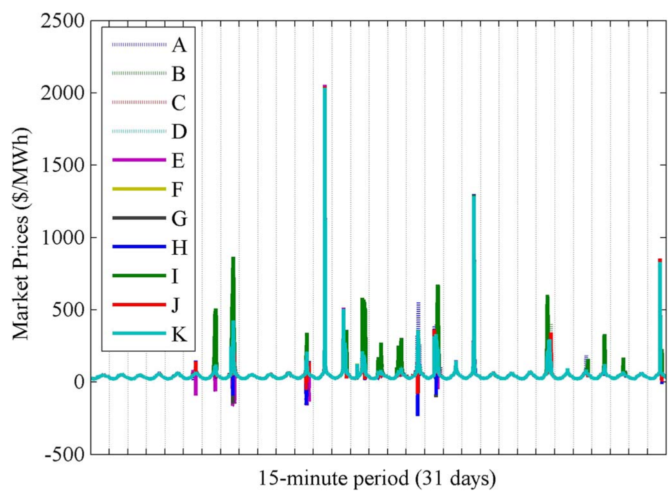

Fig. 3. DFW market prices in July 2011.

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-min time interval of market price. Therefore, the 15-min ahead HMPFM with data clustering techniques is performed and validated with 2011 ERCOT wholesale market price data.

The rest of this paper is organized as follows. Modeling of the proposed regional EV charging stations is presented in Section II. Then, the framework of HMPFM with data clustering techniques is proposed in Section III. Next, all of the implemented algorithms including SVM, CART, K-means, and stratification methods are briefly described in Section IV. Finally, Sections V and VI present a case study to illustrate the proposed approach and the 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 is designed to build nearby the power nodes in the DFW area represented by red circles in Fig. 2. There are 26 power nodes in 11 clusters. These power nodes can have different nodal market prices at different clusters and have similar market prices within the same clusters. They can serve as a point of interconnection of dc fast (level 3) charging to the power grid. The ERCOT wholesale market prices [20] of these 11 clusters 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.

III. Hybrid Method for Market Price 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 nonspike price magnitude prediction is processed. The details in each process are described in the following discussion.

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; and 3) *a negative price* is where the price falls below zero.

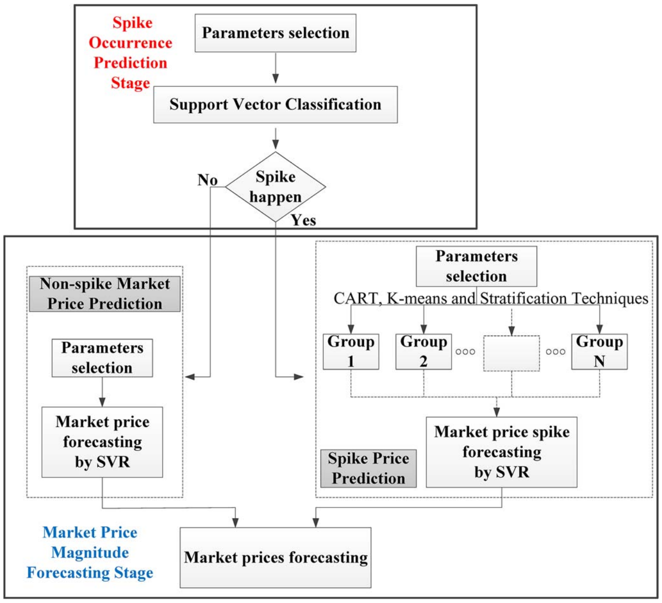

Fig. 4. Hybrid market price prediction framework.

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

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

where μ and σ are the mean and standard deviation of the market price, respectively (43.59 and 162.32 \$/MWh for the 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 the model with the best performance.

B. Nonspike Market Price Prediction

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

C. Spike Market Price Prediction

Spike prices in the 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.

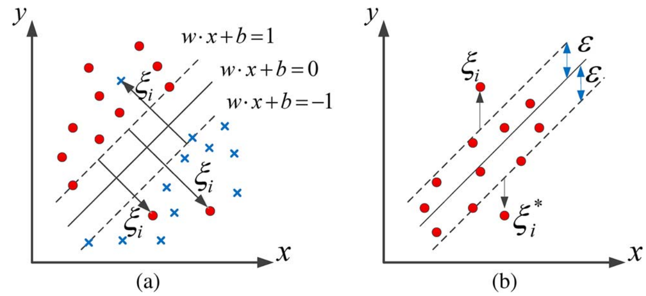

Fig. 5. SVM. (a) Classification. (b) Regression.

IV. SVMs and Data Clustering Techniques

A. SVMs

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) SVC: Fig. 5(a) illustrates the linear separability 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 defines the characteristics of the hyperplane. Moreover, $y_i = +1$ and $y_i = -1$ represent nonspike and spike classes, respectively. Thus, two constraints regarding this two-class separable hyperplane are shown in the following:

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

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

The target of the 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

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

where C is a regularization parameter defined by the error penalty and ξ_i is a slack variable determined by the distance between the incorrectly classified x_i and the 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 kernel given the satisfactory SVM prediction performance [14], [15] is used to perform all of the forecasting and can be described as (6)

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

2) 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 the margin, (7) and (8) describe the objective function and problem constraints regarding ε and ξ_i. The techniques used to remedy this regression problem are similar to the classification solution by applying the Lagrange multiplier and kernel function as explained in the previous section

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

Subject to

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

B. Data Clustering Techniques

1) 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 the 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 the 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 $x_i \le d$. 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) and (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

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

$$\phi(t) = SS(t) - SS(t_R) - SS(t_L) \quad (10)$$

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

2) K-Means Clustering [24]: This algorithm separates the d-dimensional vector space of data point (x_i), $D = \{x_i | i = 1, \ldots, N\}$ into k partitions by minimizing the cost function as

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

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

To reach the aim of cost minimization, this algorithm performs iteratively two-step procedures. First, c_j is initialized randomly, and data points are assigned to the closest centroid by implementing a Euclidean distance function. Second, a new c_j is computed by assigned data from the first step. This iteration is repeated until c_j is 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 the 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 is near Dallas is used to illustrate the proposed market price predicting method. First, correlation analysis is carried out to select the input parameters for the 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 the prediction performance. The proposed approach is then applied to other power nodes to improve the forecasting accuracy for other EV charging station locations in the DFW area.

A. Parameter Selection

Typically, one can obtain historical market prices, temperatures, and load profiles before performing a 15-min ahead market price forecasting, while several factors such as generator contingencies and transmission constraints remain unknown prior to predicting the market price. Other factors, such as fuel prices and day-ahead load forecast, have less influence on 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 the National Climatic Data Center [25] and ERCOT websites [20]. Fig. 6 depicts the correlation results between the market price and 15 min until 12-h time lags of three impact parameters.

According to Fig. 6, all correlations decrease significantly when prior times increase. Historical market prices show a strong autocorrelation with coefficients of greater than 0.7 until 1 h before, so this parameter is decided as one important

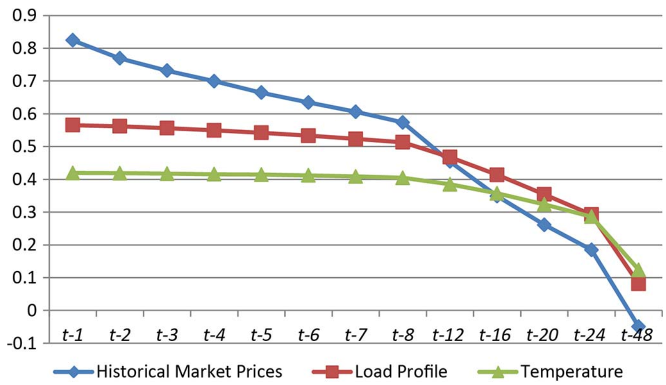

Fig. 6. Correlation analyses between market price and three impact parameters [$(t-1), (t-2), \ldots, (t-n)$ are the $1, 2, \ldots, n$ prior times in a 15-min period].

predictor. Moreover, both historical load profiles and temperatures give moderate correlations to market price with a 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 $P(\text{in})$ and $P(\text{out})$ given by (12) and (13) in order to specify the spike occurrence prediction accuracy. These two indices provide classification precision of predicted spikes and incorrect classification of predicted nonspikes. The effective classification forecasting is determined by high $P(\text{in})$ and low $P(\text{out})$

$$P(\text{in}) = P(\text{correctly predicted spike} | \text{ predicted spike}) \tag{12}$$

$$P(\text{out}) = P(\text{incorrectly predicted nonspike} | \text{ predicted nonspike}). \tag{13}$$

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 to a strong autocorrelation, they are selected to run spike price occurrence prediction for four time lag models. Second, the classification performs 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. Previous study shows that the dependence of temperature and load profile can be either strong or weak. In addition, it can have a positive or negative correlation [26]. Different dependences of these two parameters motivate the final evaluation step considering all combinations to examine the possible further prediction improvement. Two-thirds of the 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.

A significantly low $P(\text{out})$ in Table I is a result of high nonspike and low spike price number compared to the total number of testing data. Following the aforementioned procedure, the model of historical market prices is simulated from $mp(t-1)$ until $mp(t-1,\ldots,t-4)$. Spike price occurrence prediction by the model of mp $(t-1$ and$t-2)$ yields the best result compared to other models with the highest $P(\text{in})$ and the lowest $P(\text{out})$ of 0.78 and 0.0046, respectively. Then, this model combined with L $(t-1$ and$t-2)$ enhances classification performance and provides the most accurate model compared to the other combination models. This model is selected for spike price occurrence prediction in the HMPFM.

In order to improve classification performance, two adjustable parameters in SVM, including regularization (C) and bandwidth (B), are tuned. The model from the previous step is conducted to modify both two parameters. The initial settings for C and B are 10 and 2, respectively. The modification results are shown in Tables II and III. The modification of regularization and bandwidth notably improves the classification accuracy. The most efficient parameter setting ($C = 5000$ and $B = 20$) elevates $P(\text{in})$ to 0.85 and stabilizes $P(\text{out})$ at 0.0046.

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)	0.78	0.0052
mp(*t*-1&*t*-2)&*T*(*t*-1)&*L*(*t*-1&*t*-2)	0.80	0.0048
mp(*t*-1&*t*-2)&*T*(*t*-1)&*L*(*t*-1,…,*t*-3)	0.78	0.0049
mp(*t*-1&*t*-2)&*T*(*t*-1&*t*-2)&*L*(*t*-1)	0.78	0.0052
mp(*t*-1&*t*-2)&*T*(*t*-1&*t*-2)&*L*(*t*-1&*t*-2)	0.80	0.0048
mp(*t*-1&*t*-2)&*T*(*t*-1&*t*-2)&*L*(*t*-1,…,*t*-3)	0.78	0.0049
mp(*t*-1&*t*-2)&*T*(*t*-1,…,*t*-3)&*L*(*t*-1)	0.78	0.0052
mp(*t*-1&*t*-2)&*T*(*t*-1,…,*t*-3)&*L*(*t*-1&*t*-2)	0.80	0.0056
mp(*t*-1&*t*-2)&*T*(*t*-1,…,*t*-3)&*L*(*t*-1,…,*t*-3)	0.78	0.0049

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

TABLE II
SPIKE OCCURRENCE PREDICTION RESULTS BY SVM PARAMETER MODIFICATION ($C = 10$)

B	0.4	2	10	20	30	100
P(in)	0.71	0.80	0.80	0.81	0.79	0.78
P(out)	0.0061	0.0048	0.0049	0.0048	0.0049	0.0052

TABLE III
SPIKE OCCURRENCE PREDICTION RESULTS BY SVM PARAMETER MODIFICATION ($B = 20$)

C	0.1	10	100	1000	5000	10000
P(in)	0.78	0.81	0.83	0.84	0.85	0.84
P(out)	0.0053	0.0048	0.0048	0.0049	0.0046	0.0045

C. Nonspike Market Price Prediction

SVR is carried out to estimate the magnitude of nonspike prices in the same way as the spike occurrence prediction. The forecasting performance is evaluated by mean absolute

TABLE IV
NONSPIKE MARKET PRICE PREDICTION RESULTS

Models	MAPE (%)
mp(*t*-1)	6.02
mp(*t*-1&*t*-2)	5.94
mp(*t*-1,…,*t*-3)	5.95
mp(*t*-1,…,*t*-4)	6.02
mp(*t*-1&*t*-2)&*T*(*t*-1)	5.94
mp(*t*-1&*t*-2)&*L*(*t*-1)	6.02
mp(*t*-1&*t*-2)&*T*(*t*-1&*t*-2)	5.93
mp(*t*-1&*t*-2)&*L*(*t*-1&*t*-2)	5.94
mp(*t*-1&*t*-2)&*T*(*t*-1,…,*t*-3)	5.93
mp(*t*-1&*t*-2)&*L*(*t*-1,…,*t*-3)	5.96
mp(*t*-1&*t*-2)&*T*(*t*-1)&*L*(*t*-1)	6.00
mp(*t*-1&*t*-2)&*T*(*t*-1)&*L*(*t*-1&*t*-2)	5.93
mp(*t*-1&*t*-2)&*T*(*t*-1)&*L*(*t*-1,…,*t*-3)	5.94
mp(*t*-1&*t*-2)&*T*(*t*-1&*t*-2)&*L*(*t*-1)	6.00
mp(*t*-1&*t*-2)&*T*(*t*-1&*t*-2)&*L*(*t*-1&*t*-2)	5.92
mp(*t*-1&*t*-2)&*T*(*t*-1&*t*-2)&*L*(*t*-1,…,*t*-3)	5.93
mp(*t*-1&*t*-2)&*T*(*t*-1,…,*t*-3)&*L*(*t*-1)	5.98
mp(*t*-1&*t*-2)&*T*(*t*-1,…,*t*-3)&*L*(*t*-1&*t*-2)	5.92
mp(*t*-1&*t*-2)&*T*(*t*-1,…,*t*-3)&*L*(*t*-1,…,*t*-3)	5.93

percentage error (MAPE) calculated by (14). The forecasting results are shown in Table IV

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

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

The results in Table IV show the prediction performance of SVR. Temperature and load profile can enhance forecasting precision slightly. The model of mp $(t-1$ and$t-2)$ including T $(t-1$ and$t-2)$ and L $(t-1$ and$t-2)$ offers the best result with 5.92% MAPE compared to the results of the other models. This model is selected in the HMPFM for nonspike 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 the clustering selection results of the three proposed approaches prior to performing comprehensive HMPFM in the next stage.

1) 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 the parent nodes recommended by the software inventor [27]. The optimal results appropriately specify different terminal nodes that are a 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 $T(t-1,\ldots,t-3)$ are shown in Table V. For instance, the rule for the sixth cluster is $mp(t-1)$ fallen between 2086.89 and 3000.6 \$/MWh.

TABLE V
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

TABLE VI
FOUR CLUSTERS BY K-MEANS

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

2) K-Means: K-means clustering is performed to obtain proper clusters and has yielded separate input parameters for each group. Then, the input parameters in each group are averaged to be the decision values. The lowest Euclidean distance calculated by (15) is carried out for selecting the appropriate groups prior to predicting the magnitude of the spike price. An example result from K-means of the model including $mp(t-1,\ldots,t-3)$ is shown in Table VI

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

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

As the results, one can see that K-means clustering is able to separate the input parameters for each group effectively. All average decision values of the input parameters are less than 282.13 and more than 2762.85 \$/MWh in clusters 1 and 3, respectively. In addition, the average decision values of the 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 the number of cluster members based on the total spike price number. According to different levels of spike prices specified by dissection, the 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, the input parameters in each group are averaged to be the decision values. The lowest Euclidean distance defined by (15) is employed to select the appropriate clusters before performing the prediction. Example result by four groups of the model including $mp(t-1,\ldots,t-3)$ is shown in Table VII. The proper number of clusters is discussed in the next section.

TABLE VII
FOUR CLUSTERS BY THE STRATIFICATION METHOD

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

TABLE VIII
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) &L(t-1)	15.28	mp(t-1,…,t-3) &T(t-1) &L(t-1)	15.21	mp(t-1&t-2) &T(t-1) &L(t-1)	16.41
mp(t-1) &T(t-1) &L(t-1&t-2)	15.25	mp(t-1,…,t-3) &T(t-1) &L(t-1&t-2)	15.50	mp(t-1&t-2) &T(t-1) &L(t-1&t-2)	16.37
mp(t-1) &T(t-1) &L(t-1,…,t-3)	15.28	mp(t-1,…,t-3) &T(t-1) &L(t-1,…,t-3)	15.19	mp(t-1&t-2) &T(t-1) &L(t-1,…,t-3)	16.36
mp(t-1) &T(t-1&t-2) &L(t-1)	15.31	mp(t-1,…,t-3) &T(t-1&t-2) &L(t-1)	15.26	mp(t-1&t-2) &T(t-1&t-2) &L(t-1)	16.36
mp(t-1) &T(t-1&t-2) &L(t-1&t-2)	15.42	mp(t-1,…,t-3) &T(t-1&t-2) &L(t-1&t-2)	15.24	mp(t-1&t-2) &T(t-1&t-2) &L(t-1&t-2)	16.34
mp(t-1) &T(t-1&t-2) &L(t-1,…,t-3)	15.35	mp(t-1,…,t-3) &T(t-1&t-2) &L(t-1,…,t-3)	15.86	mp(t-1&t-2) &T(t-1&t-2) &L(t-1,…,t-3)	16.33
mp(t-1) &T(t-1,…,t-3) &L(t-1)	15.32	mp(t-1,…,t-3) &T(t-1,…,t-3) &L(t-1)	15.51	mp(t-1&t-2) &T(t-1,…,t-3) &L(t-1)	16.35
mp(t-1) &T(t-1,…,t-3) & L(t-1&t-2)	15.34	mp(t-1,…,t-3) &T(t-1,…,t-3) &L(t-1&t-2)	15.88	mp(t-1&t-2) &T(t-1,…,t-3) &L(t-1&t-2)	16.34
mp(t-1) &T(t-1,..,t-3) &L(t-1,..,t-3)	15.29	mp(t-1,…,t-3) &T(t-1,..,t-3) &L(t-1,..,t-3)	15.32	mp(t-1&t-2) &T(t-1,..,t-3) &L(t-1,..,t-3)	16.33

E. Comprehensive Results

The selected models from spike occurrence prediction and nonspike 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 the K-means and stratification methods are set at four. Following the similar procedure for spike occurrence and nonspike price magnitude prediction, the results of spike price magnitude prediction are obtained from the comprehensive HMPFM tested in several models as shown in Table VIII. The prediction performance is evaluated by MAPE.

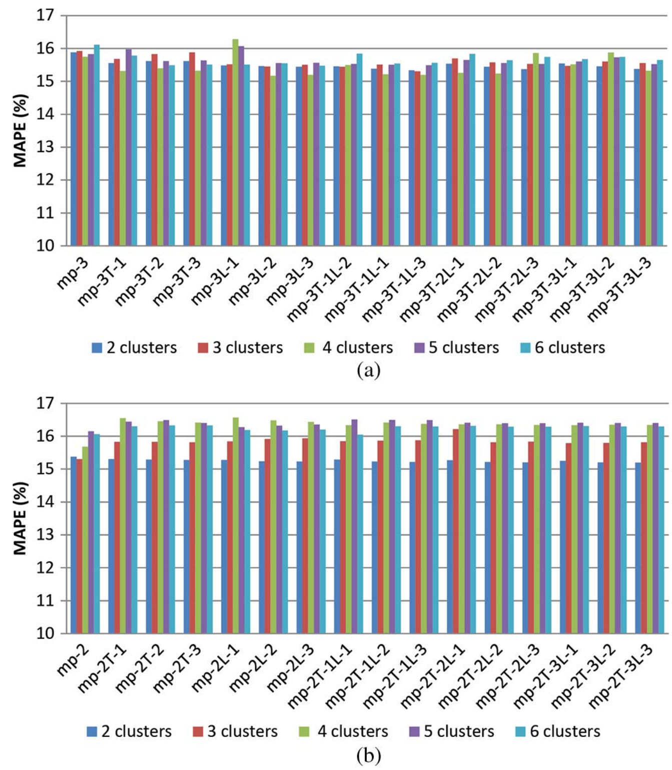

Fig. 7. Comparison of market price prediction results with different clusters. (a) K-means method. (b) Stratification method.

According to Table VIII, each model provides similar market price forecasting accuracy by the three proposed clustering approaches. The models with the best performance of spike price forecasting from the comprehensive HMPFM are $mp(t-1)$ and$T(t-1)$ and$L$ $(t-1$ and$t-2)$, $mp(t-1,\ldots,t-3)$ and$L$ $(t-1$ and$t-2)$, and mp $(t-1$ and$t-2)$ for the CART, K-means, and stratification methods, respectively. These predictions give the lowest MAPE with 15.25%, 15.17%, and 15.68%, respectively.

In addition, the number of clusters is adjusted from two to six with different temperature and load profile combinations to compare and select the optimal results from the K-means and stratification methods. The results are shown in Fig. 7. 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 previously discussed, while two clusters with the model of mp $(t-1$ and$t-2)$ and$T(t-1,\ldots,t-3)$ and$L(t-1,\ldots,t-3)$ for the stratification method give the best prediction performance and improve the MAPE from 15.68% to 15.19%.

To illustrate apparently the improvement of the HMPFM combined with three clustering techniques compared to other prediction methods, Fig. 8 depicts the comparison results between the best cases of the three proposed approaches and

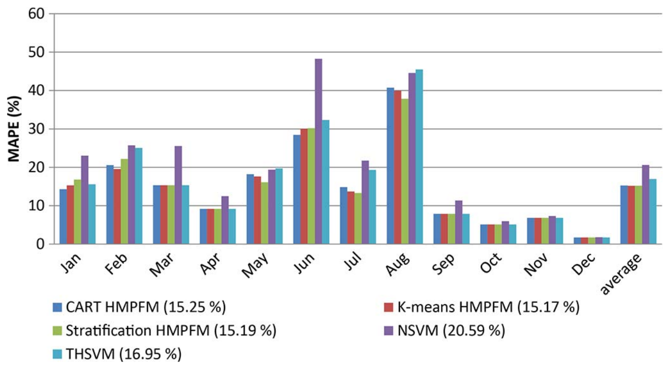

Fig. 8. Market price forecasting comparison results of various approaches.

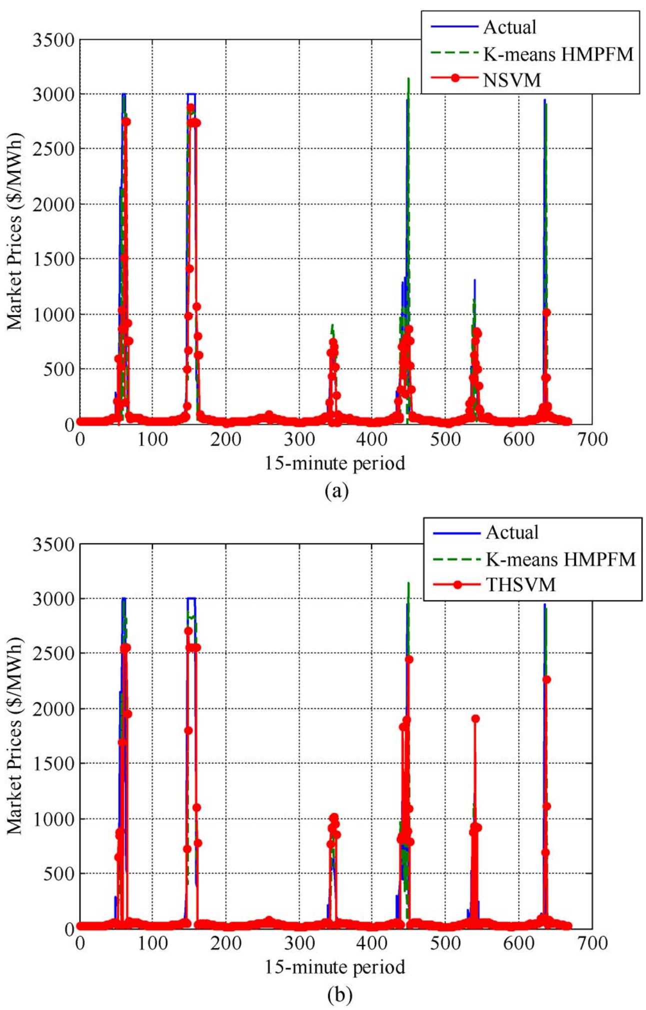

Fig. 9. Comparison of market price prediction results from the proposed K-means hybrid SVM. (a) With NSVM. (b) With THSVM.

the other general prediction methods including normal SVM (NSVM) and typical hybrid SVM (THSVM). Considering the same parameters and data with the proposed approaches, NSVM is carried out to forecast the market price only by the traditional SVM algorithm, and THSVM is executed by the hybrid of SVC and SVR without the proposed data clustering techniques. These two methods are programmed with MATLAB. MAPEs reduce remarkably from 20.59% and 16.95% by NSVM and THSVM to 15.25%, 15.17%, and 15.19% from the

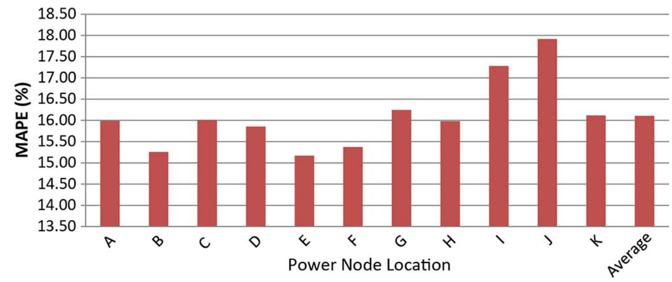

Fig. 10. Market price prediction results from K-means HMPFM for all power nodes in the DFW area.

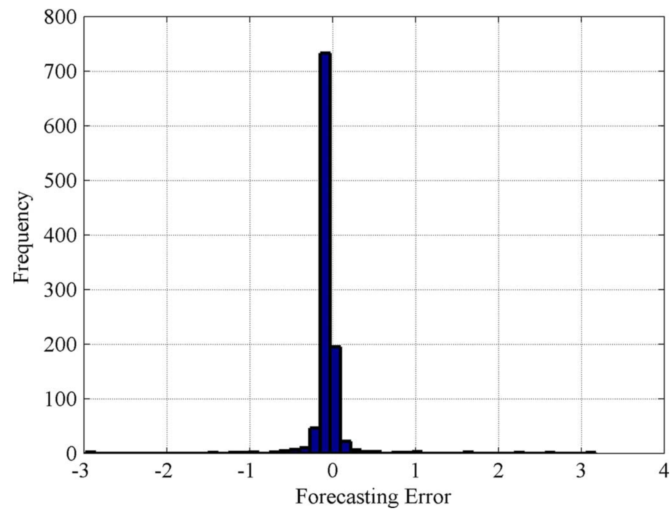

Fig. 11. Forecast change error distribution by MMFE.

best prediction performance models by HMPFM with th CART, K-means, and stratification methods, respectively.

As also shown in Fig. 9, since K-means HMPFM with four clusters of $mp(t-1,\ldots,t-3)$ and$L$ $(t-1$ and$t-2)$ 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 nonspike price estimation. Fig. 9(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. 9(b).

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 the DFW area to cover all locations of the EV charging station system. The market prices for all power nodes are predicted by the proposed method, and the prediction results are shown in Fig. 10. One can see that the proposed method yields similar results for all power nodes with an average MAPE of 16.11%.

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) [28]. In the multiplicative model, MMFE determines the forecast change error as the log normal function by (16). An example forecast change error distribution of the Dallas power node is depicted in Fig. 11. The uncertainty in the stochastic cost minimizing problem for the EV charging station system can be generated by this probability density function. The further study for optimal operation of the regional EV charging station system applying the uncertainty function by K-means HMPFM will be the focus in future work.

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

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

VI. Conclusion

This paper has presented 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 and nonspike 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 the ERCOT wholesale market price in the 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 W.-J. Lee, "Design optimization of PHEV charging station," in *Conf. Rec. IEEE IAS Annu. Meeting*, Louisville, KY, USA, 2012, pp. 1–7.

[2] A. Kulvanitchaiyanunt, V. Chen, J. Rosenberger, W.-J. Lee, and P. Sarikprueck, "Control for a system of PHEV charging stations," in *Proc. Ind. Syst. Eng. Res. Conf.*, San Juan, Puerto Rico, 2013, pp. 356–366.

[3] J. Bastian, J. Zhu, V. Banunarayanan, and R. Mukerji, "Forecasting energy prices in a competitive market," *IEEE Comput. Appl. Power*, vol. 12, no. 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. Power Syst.*, vol. 19, no. 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. Gener., Transmiss. Distrib.*, vol. 149, no. 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. Gener., Transmiss. Distrib.*, vol. 3, no. 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. Power Syst.*, vol. 22, no. 4, pp. 2058–2065, Nov. 2007.

[8] J. Contreras, R. Espnola, 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.

[9] A. J. Conejo, M. A. Plazas, R. Espnola, 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.

[10] Y. Baez-Rivera, B. Rodriguez-Medina, and A. K. Srivastava, "An attempt to forecast price spikes in electric power markets," in *Proc. 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. Power Syst.*, vol. 22, no. 1, pp. 376–385, Feb. 2007.

[12] B. Ernst *et al.*, "Predicting the wind," *IEEE Power Energy Mag.*, vol. 5, no. 6, pp. 78–89, Nov./Dec. 2007.

[13] R. Xu, H. Chen, and X. Sun, "Short-term photovoltaic power forecasting with weighted support vector machine," in *Proc. IEEE Int. Conf. ICAL*, Zhengzhou, China, 2012, pp. 248–253.

[14] Y. Liu, S. Jie, Y. Yang, and W.-J. Lee, "Short-term wind-power prediction based on wavelet transform-support vector machine and statistic-characteristics analysis," *IEEE Trans. Ind. Appl.*, vol. 48, no. 4, pp. 1136–1141, May 2012.

[15] S. Jie, W.-J. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting power output of photovoltaic systems based on weather classification and support vector machines," *IEEE Trans. Ind. Appl.*, vol. 48, no. 3, pp. 1064–1069, Mar. 2012.

[16] K. R. Skinner *et al.*, "Multivariate statistical methods for modeling and analysis of wafer probe test data," *IEEE Trans. Semicond. Manuf.*, vol. 15, no. 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. Neural Netw. Learn. Syst.*, vol. 23, no. 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. Power Syst.*, vol. 22, no. 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. Power Del.*, vol. 22, no. 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, USA: Chapman & Hall, 2009, ch. 3, pp. 37–59.

[22] D. Steinberg, "CART: classification and regression trees," in *The Top Ten Algorithms in Data Mining*. Boca Raton, FL, USA: Chapman & Hall, 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*. Washington DC, USA: International Food Policy Research Institute, 1999.

[24] J. Ghosh and A. Liu, "K-means," in *The Top Ten Algorithms in Data Mining*. Boca Raton, FL, USA: Chapman & Hall, 2009, ch. 2, pp. 21–35.

[25] National Climate Data Center. [Online]. Available: http://www.ncdc.noaa.gov/

[26] S. Fan, L. Chen, and W.-J. Lee, "Machine learning based switching model for electricity load forecasting," *Energy Convers. Manag.*, vol. 49, no. 6, pp. 1331–1344, Jun. 2008.

[27] D. Steinberg and M. Golovnya, *CART 6.0: User's Guide*. San Diego, CA, USA: Salford Systems, 2007.

[28] Heath and D. C. Jackson, "Modelling the evolution of demand forecasts with application to safety stock analysis in production distribution systems," School Oper. Res. Ind. Eng., Cornell Univ., Ithaca, NY, USA, Tech. Rep. 989, 1991.

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 currently working toward the Ph.D. degree in the Energy Systems Research Center, The University of Texas at Arlington, Arlington, TX, USA.

From 2002 to 2005, he was a Researcher with the Engineering Service and Developing Center, KMITL, where he has been a Lecturer with the Department of Electrical Engineering since 2005 (currently on sabbatical leave). His research areas of interest include energy conversion, rotating electric machines, power systems, renewable-energy applications, and statistical forecasting.

Wei-Jen Lee (S'85–M'85–SM'97–F'07) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1978 and 1980, respectively, and the Ph.D. degree in electrical engineering from The University of Texas at Arlington (UTA), Arlington, TX, USA, in 1985.

In 1985, he joined the UTA, where he is currently a Professor with the Department of Electrical Engineering 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 Registered Professional Engineer in the State of Texas.

Asama Kulvanitchaiyanunt (S'13) received the B.S. degree in chemical engineering from Chulalongkorn University, Bangkok, Thailand, the M.S. degree in industrial engineering from Lehigh University, Bethlehem, PA, USA, and the Ph.D. degree in industrial engineering from The University of Texas at Arlington (UTA), Arlington, TX, USA.

She is a Postdoctoral Researcher with the UTA, where she is a member of the Center on Stochastic Modeling Optimization and Statistics. Her research interests include mathematical programming, dynamic programming, statistical modeling, machine learning, and optimization in energy application.

Victoria C. P. Chen received the B.S. degree in mathematical sciences from The Johns Hopkins University, Baltimore, MD, USA, and the M.S. and Ph.D. degrees in operations research and industrial engineering from Cornell University, Ithaca, NY, USA.

She is a Professor and the Interim Chair of Industrial and Manufacturing Systems Engineering with The University of Texas at Arlington, Arlington, TX, USA. She is actively involved with the Institute for Operations Research and Management Science (INFORMS), having cofounded the INFORMS Section on Data Mining and served as Officer for the Section on Data Mining and Forum for Women in OR/MS. She is currently serving on the INFORMS Subdivision Council. She is a Guest Editor of the *Annals of Operations Research* and an Associate Editor of the journal *Elemental Science of the Anthropocene*. Her 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 received the B.S. degree in mathematics from Harvey Mudd College, Claremont, CA, USA, the M.S. degree in industrial engineering and operation research from the University of California at Berkeley, Berkeley, CA, USA, and the Ph.D. degree in industrial engineering from Georgia Institute of Technology, Atlanta, GA, USA.

He is an Associate Professor of industrial and manufacturing systems engineering with The University of Texas at Arlington (UTA), Arlington, TX, USA. He is the Director of the Center on Stochastic Modeling Optimization and Statistics, UTA. Prior to joining the faculty at UTA, he was with the Operations Research and Decision Support Department, American Airlines. His research interests include mathematical programming, applied simulation, and optimization of statistical metamodels of complex systems. He has applied his methodological research to solve numerous real-world problems including those in transportation, healthcare, defense, and energy.

Dr. Rosenberger 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.