Multiregion Load Forecasting for System With Large Geographical Area

Shu Fan, *Member, IEEE*, Kittipong Methaprayoon, *Member, IEEE*, and Wei-Jen Lee, *Fellow, IEEE*

***Abstract*—In a power system covering a large geographical area, a single model for load forecasting of the entire area sometimes cannot guarantee satisfactory forecasting accuracy. One of the major reasons is because of the load diversity, usually caused by weather diversity, throughout the area. Multiregion load forecasting will be a feasible and effective solution to generate more accurate forecasting results, as well as provide regional forecasts for the utilities. However, a major challenge is how to optimally partition/merge the areas according to the regional load and weather conditions. This paper investigates the electricity demand and weather data from an electric utility in Midwest, U.S. Based on the data analysis, we demonstrate the existence of weather and load diversity within its control area and then develop a short-term multiregion load forecasting system based on support vector regression for day-ahead operation and market. The proposed multiregion forecasting system can find the optimal region partition under diverse weather and load conditions and finally achieve more accurate forecasts for aggregated system load. The proposed forecasting system has been tested by using the real data from the system. The numerical results obtained for different region partition schemes validate the effectiveness of the proposed multiregion forecasting system. The detailed discussions on the forecasting results have also been given in this paper.**

***Index Terms*—Load diversity, load forecasting, multiregion, support vector regression (SVR).**

I. Introduction

LOAD forecasting is a key issue for power-system operation, planning, and marketing [1]. Many operating decisions, such as dispatch scheduling of generating capacity, reliability analysis, and maintenance plan for the generators, are based on load forecasts. Planning for future new-generation plant and transmission augmentations is also dependent on load forecasts. Therefore, load forecasting is always a popular topic in power industry.

So far, a wide variety of techniques have been proposed to forecast the electricity load [2]. In general, most of the works are focusing on forecasting models themselves, and no specific attention has been paid to the load forecasting for a system with large geographical area, although some problems are required to be resolved in real operation.

In a power system with large geographical area, the weather and electricity demand diversity across the entire area is a key issue that influences the forecasting accuracy. The weather and load diversity make it hard to predict the aggregated electricity demand by using a single forecasting model. To tackle the influence incurred by the diversity, a straightforward idea is to forecast the loads in separated subregions, i.e., multiregion load forecasting. However, a technical difficulty arises when performing the multiregion forecasting, i.e., how to find the optimal partition/combination of the areas so that more accurate forecasts can be obtained for both aggregated and regional loads.

The purpose of this paper is to develop a multiregion short-term load forecasting system for an electricity utility in Midwest U.S., the control area of which covers a large geographical area. We first investigate and quantify the weather and load diversity within the system and then develop a multiregion load forecasting system using support vector regression (SVR). The proposed multiregion forecasting system can find the optimal region partition according to the load and weather characteristics and then generate more accurate forecasts for aggregated system load. The proposed forecasting system has been tested by using the real data from the utility. For comparison, a universal forecasting model has also been developed to forecast the aggregated system load, and the numerical results validate the superiority of the proposed multiregion load forecasting system.

II. Investigation of Regional Load and Weather Characteristics

In this section, we investigate the regional weather and load characteristics as well as the dependence of electricity demand on temperature for the target system.

A. System Load

The target power system used in this paper is the control area of an electric utility in Midwest, U.S., covering a larger geographical area. The major demand of this system is residential. Initially, the control area of the utility has been divided into 24 areas mainly based on its member–owner cooperatives. The

Paper ICPSD-08-4-3, presented at the 2008 IEEE/IAS Industrial and Commercial Power Systems Technical Conference, Clearwater Beach, FL, May 4–8, and approved for publication in the IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS by the Energy Systems Committe of the IEEE Industry Applications Society. Manuscript submitted for review April 30, 2008 and released for publication December 19, 2008. First published June 2, 2009; current version published July 17, 2009.

S. Fan is with the Business and Economic Forecasting Unit, Monash University, Clayton, VIC 3800, Australia (e-mail: shu.fan@buseco.monash.edu.au).

K. Methaprayoon is with the Electric Reliability Council of Texas, Austin, TX 78744 USA (e-mail: kmethaprayoon@ercot.com).

W.-J. Lee is with the Energy Systems Research Center, University of Texas, Arlington, TX 76019 USA (e-mail: wlee@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.2009.2023569

0093-9994/$25.00 © 2009 IEEE

TABLE I
AREA NUMBER AND LOAD DATA IN 2006

Area number	Average load (MW)	Peak Load (MW)
Area001	8.111838	14.30769
Area002	25.03388	41.20340
Area003	52.70056	96.61885
Area004	61.6222	103.8052
Area005	51.39617	90.45446
Area006	27.85061	50.85374
Area007	78.43186	128.3923
Area008	10.31720	16.65236
Area009	9.618536	22.77330
Area010	23.22630	48.25764
Area011	20.19723	33.64496
Area012	19.39542	31.16093
Area013	31.17934	52.48743
Area014	42.0646	88.07723
Area015	18.36458	30.67214
Area016	1.332454	5.205772
Area017	119.0486	252.3563
Area018	31.15284	60.53848
Area019	31.23019	58.47161
Area020	34.87022	51.81867
Area021	31.82051	63.33005
Area022	16.95924	32.27306
Area023	29.7387	59.166
Area024	2.310241	6.282474
Range	1.33~119.05	5.21~252.36

yearly average and peak load of each area in 2006 are listed in Table I.

According to Table I, it can be seen that the system load in each area varies in a wide range. The ratios between peak and average load in different areas are also quite different. One of the major reasons is because of the weather and load diversity within the large area, which will further be investigated in the following section.

B. Weather Variations Throughout the Area

In recent years, demand levels have become increasingly dependent on weather conditions. This has been attributed to the augment in availability of household and commercial air conditioning units. Generally, most of the meteorological variables, including temperature, humidity, and wind speed, etc., are driving variables to the electricity consumption. On the other hand, temperature was consistently found to be the dominant factor compared with the other weather indexes. Therefore, we will focus on the temperature characteristics in this paper without losing generality.

Table II shows the mean, maximum, and minimum temperature of each area in 2006. It can be seen that the temperature values vary in a wide range across the area, particularly for the minimum temperature.

TABLE II
MEAN, MAXIMUM, AND MINIMUM TEMPERATURE OF EACH AREA IN 2006

AREA number	Mean	Maximum	Minimum
Area001	65.81763	107	11
Area002	62.13418	108	9
Area003	64.49312	107	11
Area004	64.16981	102	12
Area005	65.2785	104	15
Area006	64.25483	110	2
Area007	64.9285	108	9
Area008	65.38804	110	8
Area009	62.39155	107	2
Area010	64.86872	106	5
Area011	64.07923	106	11
Area012	63.46147	104	11
Area013	61.25713	108	4
Area014	64.88092	106	5
Area015	62.42633	107	2
Area016	69.1116	109	19
Area017	64.98563	107	13
Area018	67.35749	108	15
Area019	69.11051	109	19
Area020	65.4628	106	12
Area021	65.98925	108	4
Area022	66.14058	107	11
Area023	62.23019	104	11
Area024	66.46377	111	12
Range	62.23~69.11	102~111	2~19

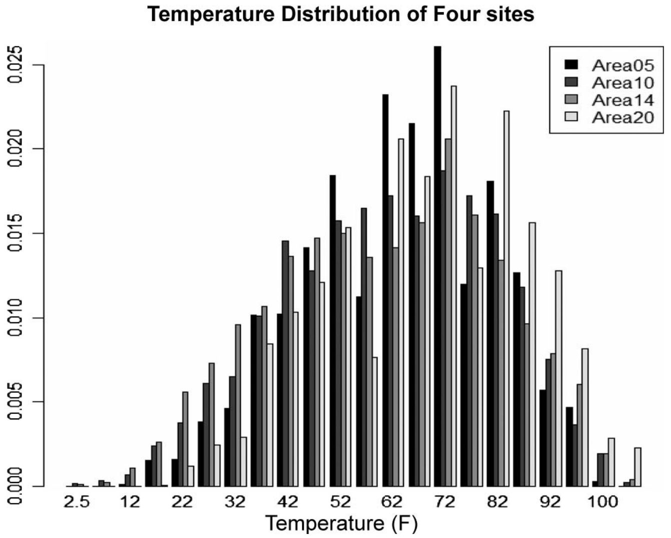

Fig. 1. Histogram plot of temperature distribution of four different sites across the control area in year 2006.

To further reveal the temperature variation in different areas, four typical sites located in different parts of the control area have been selected to illustrate the distribution of temperature.

As can be seen in Fig. 1, the temperature distributions vary significantly at different locations.

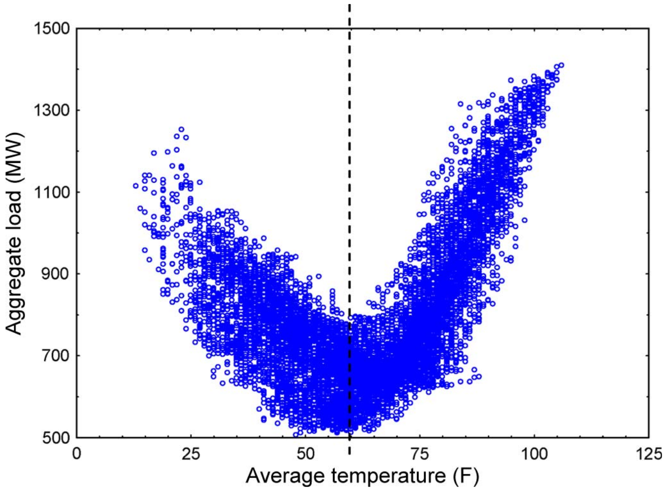

Fig. 2. Correlation between aggregated demand and temperature in 2006.

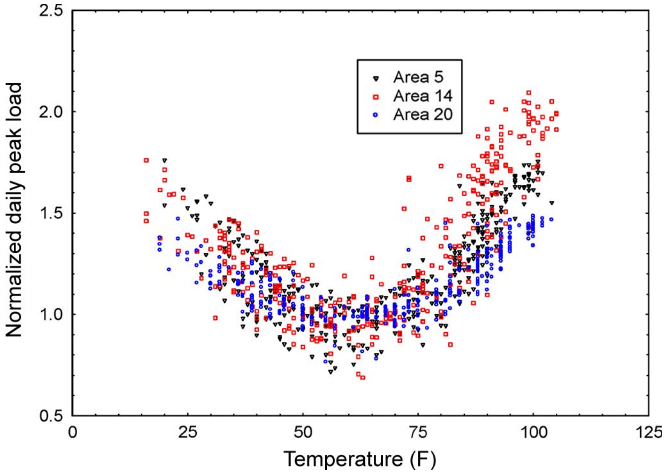

Fig. 3. Correlation between maximum daily demand and corresponding temperature for three different areas in 2006.

C. Dependence of Electricity Demand on Temperature

Since the weather diversity exists throughout the control area of the electric utility, we continue to analyze the relationship between load and temperature. The correlation between the aggregated load and average temperature for the system is shown in Fig. 2.

According to Fig. 2, there exists approximately a piecewise linear relationship of correlations between load and temperature for cold and hot days, respectively. The correlation in each segment can be computed using the following expression:

$$\rho_{t,d} = \frac{Cov(t,d)}{\sigma_t \sigma_d} \tag{1}$$

where $Cov(t,d)$ is the covariance of temperature t and load d, and σ_t and σ_d are the standard deviations for t and d. $\rho_{t,d} = 1$ corresponds to a perfect linear correlation, while an intermediate value describes partial correlations, and $\rho_{t,d} = 0$ represents no correlation at all.

The dotted line in Fig. 2 indicates the approximate separate point between the two piecewise segments which are obtained by maximizing the absolute values of the two correlation coefficients on both segments. According to our computation, the separating point is approximately 59.0°, and $\rho_{t,d}$ on the two segments are -0.69 and 0.73, which indicates a strong correlation between load and temperature for both winter and summer seasons. Note that the correlation in winter is nearly as high as that in summer; a major reason is because the heat appliances in the control area are mainly electric, instead of gas and oil heaters.

Fig. 3 shows the correlations between the maximum daily demand and the corresponding temperature for three of the aforementioned four areas in 2006. In this figure, the demands were normalized by their yearly average. As can be seen, the correlations between electricity demand and temperature are generally high, whereas the demand in different area displays different levels of dependence on temperature.

D. Load Diversity Analysis

Since the load is largely dependent on the ambient temperature and weather conditions are variable throughout the areas, it can be inferred that the electric demand would also be diverse within the large area. In this section, we quantify this load diversity within the system.

Load diversity is a reference to the level that different electricity demand patterns affect the overall system demand. Different areas can have different daily, weekly, and seasonal load profiles [3]. Load diversity can result in different areas having noncoincident load peaks. This diversity can be partly due to the existence of weather diversity throughout the wide area of a power system [4]. For a system covering a large geographic area, the load diversity will have a large influence on the aggregated load forecasting.

The level of diversity for a group of electrical loads has been defined by a coincidence factor C [5], [6]

$$C = \frac{\sum_i P_i}{P_A} \tag{2}$$

where P_i stands for the individual peak load and P_A stands for the aggregated peak load for the group of areas.

For the 24 areas, we calculate the load diversity factor by using a peak load based on increasing time interval: from daily peak to 2-daily peak until 30-daily peak. The calculation results are shown in Fig. 4. It can be seen that the factors are larger than 100%, indicating the existence of load diversity among different areas. As expected, the coincidence factor increases over longer calculation time periods.

III. Multiregion Load Forecasting

The prior analysis demonstrates the existence of diverse load patterns within the control area. For load forecasting under such a situation, it may be difficult to accurately predict the overall electricity demand by establishing a single model for the entire area. In this paper, a multiregion forecasting system is then developed to forecast the regional loads respectively based on independent models using local demand and temperature data. By using the multiregion load forecasting system,

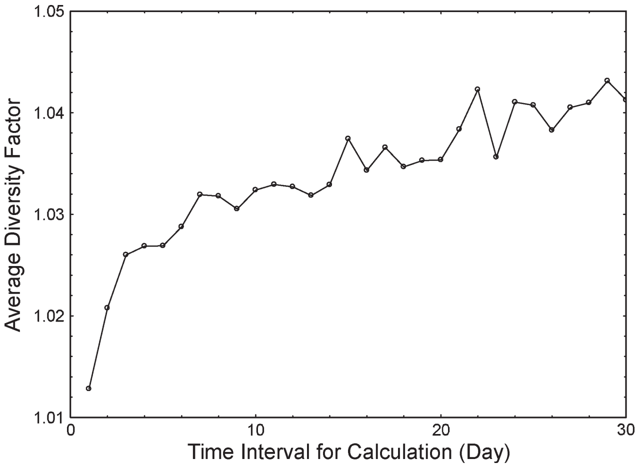

Fig. 4. Average load diversity factor for increasing calculation intervals.

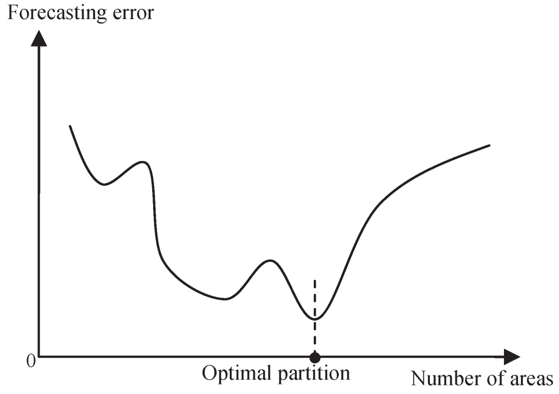

Fig. 5. Correlation between forecasting error and the number of areas.

higher accuracy for the aggregated load forecast, as well as the regional load forecasts, can be achieved, which can help system operators dispatch the generation and schedule the transmission more efficiently.

A. Influence of Region Partition to Forecasting Accuracy

When forecasting the aggregated electricity demand through the multiregion load forecasting system, the forecasting accuracy usually varies as the region partition changes. According to our calculation and past experiences in practical operation, the relationship between the aggregated load forecasting error and the number of partitioned areas can be approximately shown in Fig. 5.

In Fig. 5, the overall forecasting error decreases first as more areas have been partitioned. The forecasting error generally keeps decreasing until the number of optimal partition is reached. After that, the forecasting error keeps increasing as the number of partitioned areas increases. Here, we call this phenomenon "basin effect", which will further be explained in the numerical experiment.

B. Architecture of the Multiregion Load Forecasting System

Based on the above analysis, an adaptive multiregion load forecasting system has been developed in this paper. The structure of the proposed system is shown in Fig. 6.

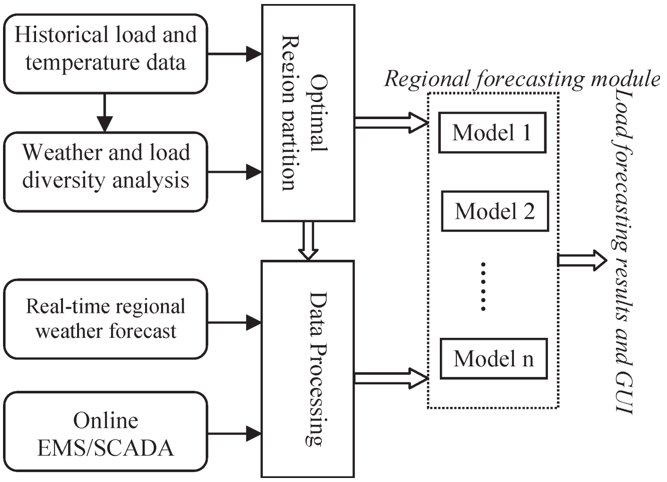
Historical load and temperature data
Weather and load diversity analysis
Optimal Region partition
Real-time regional weather forecast
Online EMS/SCADA
Data Processing
Regional forecasting module
Model 1
Model 2
......
Model n
Load forecasting results and GUI

Fig. 6. Structure of adaptive multiregion load forecasting system based on SVR.

In the proposed multiregion forecasting system, weather and load characteristics have been first analyzed by using historical data. If the system load is found to be diverse, a region partition/combination procedure is conducted to find the optimal region partition scheme, and independent regional model is then established for each partitioned area. In real-time operation, data collected from commercial weather service and EMS/SCADA system are processed and fed to corresponding regional models to generate the required forecasting results.

It is clear that the optimal region partition/combination is the key module of the entire system. Its working principle is described as follows.

First, since the entire control area has been initially divided into 24 areas based on the member cooperatives, a merging procedure is actually conducted in this paper. Theoretically, the optimal area combination scheme could be found by using exhaustive method. However, the computation burden will be intolerable if all the possible area combinations have been tried out. Since the load diversity is mainly resulting from the weather diversity and the weather diversity arises due to the spatial distance, it is then reasonable to merge the adjacent areas instead of areas isolated far from each other. Based on this assumption, a practical heuristic search procedure for the optimal region partition is proposed. The idea is to find the optimal region partition scheme by minimizing the aggregated load forecasting error for testing data set. Specifically, we first perform load forecasting individually for the 24 areas and add them together to calculate the aggregated load forecasting accuracy. Then, we choose an area located in the corner of the control area as a starting point and try to merge this area with its adjacent area one by one. A mergence will be accepted only if the aggregated load forecasting error can be lowered based on the new partition scheme, otherwise will be rejected. This procedure will continue until no more adjacent area could be merged to lower the aggregated forecasting error for the testing data set, and a new area is thus created. Then, one of the areas adjacent to the newly created area will be selected as another starting point, and the similar searching procedure will be carried out recurrently until all areas have been assigned to a unique group area, which could be regarded as the optimal partition/combination scheme. In the meanwhile, the load forecasting models for each new area are also established. Considering the weather condition in the control area may vary at different seasons; this searching procedure will be carried out for each season of the year by using the testing data set (historical regional load and local temperature) corresponding to that season. Data of two years before the testing period are used as a training sample.

C. SVR-Based Forecasting Model

In this paper, an SVR-based forecasting model has been employed for load forecasting in each area. A brief description of the SVR model is given in this section [7]–[10].

Supposing that we are given training data $(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$, where x_i's are input patterns and y_i's are the associated output value of x_i, the SVR solves an optimization problem

$$\begin{aligned} &\min_{\omega, b, \xi, \xi^*} \frac{1}{2}\omega^{\mathrm{T}}\omega + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &\text{Subject to } y_i - \left(\omega^{\mathrm{T}}\phi(x_i) + b\right) \leq \varepsilon + \xi_i^* \\ &\qquad \left(\omega^{\mathrm{T}}\phi(x_i) + b\right) - y_i \leq \varepsilon + \xi_i, \\ &\qquad \xi_i, \xi_i^* \geq 0; \qquad i = 1, \ldots, n \end{aligned} \tag{3}$$

where x_i is mapped to a higher dimensional space by the function Φ and ξ_i^* is the slack variable of the upper training error (ξ_i is the lower) subject to the ε-insensitive tube $(\omega^{\mathrm{T}}\phi(x_i) + b) - y_i \leq \varepsilon$. The constant $C > 0$ determines the trade-off between the flatness and losses. The parameters, which control regression quality, are the cost of error C, the width of the tube ε, and the mapping function Φ.

The constraints of (3) imply that we put most data x_i in the tube ε. If x_i is not in the tube, there is an error ξ_i or ξ_i^* which we tend to minimize in the objective function. SVR avoids underfitting and overfitting of the training data by minimizing the training error $C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$ as well as the regularization term $(1/2)\omega^{\mathrm{T}}\omega$. For traditional least square regression, ε is always zero and data are not mapped into higher dimensional spaces. Hence, SVR is a more general and flexible treatment on regression problems.

Since Φ might map x_i to high or infinite dimensional space, instead of solving ω for (3) in a high dimension, we deal with its dual problem, which can be derived using the Lagrange theory

$$\begin{aligned} &\max_{\alpha_i, \alpha_i^*} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)^{\mathrm{T}} Q \left(\alpha_j - \alpha_j^*\right) \\ &\qquad - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) \\ &\text{Subject to } \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \\ &\qquad 0 \leq \alpha_i, \alpha_i^* \leq C; \qquad i = 1, \ldots, n \end{aligned} \tag{4}$$

where $Q_{ij} = \phi(x_i)^{\mathrm{T}}\phi(x_j)$, and α_i and α_i^* are the Lagrange multipliers. However, this inner product may be expensive to compute because $\phi(x)$ has too many elements. Hence, we apply "kernel trick" to do the mapping implicitly. That is, to employ some special forms, inner products in a higher space yet can be calculated in the original space. Typical examples for the kernel functions are polynomial kernel $\phi(x_i)^{\mathrm{T}}\phi(x_j) = (\gamma x_1^{\mathrm{T}} x_2 + c_0)^d$ and radial basis function (RBF) kernel $\phi(x_i)^{\mathrm{T}}\phi(x_j) = e^{-\gamma(x_1 - x_2)^2}$. They are inner products in a very high dimensional space (or infinite dimensional space) but can be computed efficiently by the kernel trick even without knowing $\phi(x)$.

TABLE III

LIST OF INPUT DATA OF THE SVR NETWORK FOR REGULAR DAYS

Input	Variable name	Lagged value (hours)
1-9	Hourly load (L_1)	24,25,26,48,72,96,120, 144,168
10-19	Hourly temperature (T_1)	0,1,2,24,48,72,96,120, 144,168

Assuming that the hour of load predication is at 0, the lag 0 represents the target instant, and the 24 lagged hours mean the values that were measured 24 hours earlier than the hour of predication.

D. SVR Input Selection

Previous works have shown that the characteristics of load series between regular workdays and anomalous days, which include weekend, holidays, and days with anomalous events, are quite different [11], [12]. To achieve good forecasting results, the regular workdays and anomalous days should be treated with different schemes. Hence, we use different model feeders for regular and anomalous days. Specifically, the input data of the SVR network for regular days are shown in Table III.

As indicated, in addition to the forecasted and actual temperature, the input variables are the hourly load values of the last day available and the similar hours in the previous days or weeks. In order to capture the time series style in load, we include the electricity load and temperature of the previous seven days at the target hour. We also include the temperature information of 1 and 2 h earlier than the target hour because temperature changes normally precede load changes.

According to the historical load data, the same type of holiday showed a similar trend of load profile as in previous years, so holidays' forecasts are assessed as a function of weekend behavior in this paper. Specifically, the input data of the SVR network for anomalous days (weekend and holiday) are selected in Table IV. Here, in addition to the most recent two days' load and temperature data, eight input variables representing the data around the predicting hour in the past two weekends are used.

E. Model Training and Cross-Validation

Although SVR is shown to be resistant to the overfitting problem in solving forecasting problems of various time series [8], it is still important to apply certain techniques in the training process to achieve the highest generalization performance.

In this paper, a cross-validation procedure is conducted to optimize the tunable parameters which are the cost of error C and γ in the RBF function. Specifically, the training samples are divided into two subsets: training subset and validation subset.

TABLE IV
LIST OF INPUT DATA OF THE SVR NETWORK FOR ANOMALOUS DAYS

Input	Variable name		Lagged value (hours)
1-9	Hourly load	(L_2)	24,25,26,48,72
	Hourly load of the previous Saturday		h,h-1*
	Hourly load of the previous Sunday		h,h-1*
10-19	Hourly temperature	(T_2)	0,1,2,24,48,72
	Hourly load of the previous Saturday		h,h-1*
	Hourly load of the previous Sunday		h,h-1*

* h stands for the same clock with the target hour

The training data subset is used for updating the parameters of the network, whereas the validation data subset is adopted to monitor the training performance in the training procedure. Here, the key point is that the training process is controlled by the validation error instead of training error. In this procedure, different values of the tunable parameters of the SVR have been tried, and the optimal parameters are acquired by minimizing the prediction error for the validation set.

After the cross-validation procedure completed, the optimal parameters of the SVR model are obtained and frozen, and then, a final training process will be conducted based on the entire sample data set comprising the validation data. All the model parameters will be obtained after this learning phase and ready for the forecasting operation.

In real-time application, the cross-validation process will be conducted every week, whereas the final training process will be launched daily to update the model parameters. The validation set is selected to be located in the time period just before the target date and the same periods around the target date in the previous years, since the load patterns in these periods are usually similar to the load pattern in target period. For instance, if the cross-validation has been conducted on February 1, 2007, then data in January 2007 and February 2006 and 2005, etc., will be selected as the validation set.

For numerical experiments in this paper, we use the software LIBSVM [13], which is a library for support vector machines, including the efficient implementation of solving (4).

IV. Case Study and Numerical Results

A. Data Collection and Preprocess

The hourly electricity load data from the 24 areas and weather data observed at the corresponding local stations have been used for the study. Day-ahead load forecasting is performed in this paper. Two consecutive months of testing data, corresponding to March and April 2007, have been selected to forecast and validate the performance of the proposed model. The testing data are located in the shoulder seasons with volatile demand, which are the worst cases for load forecasting. The hourly data used as the training sample are from January 1, 2006 to December 31, 2006.

Before training, we need to preprocess the training samples. As shown in Fig. 2, there are some extremely high/low load values which deviate the mainstream of data. Basically, most of these exceptional values are caused by unusual events such as urgent maintenance of generator, contingencies and congestion events, etc. For a model based on historical data learning, these data can result in poor fit of the model to the data and should be considered as noise in the training samples. To improve the training accuracy, we applied statistics of studentized residuals via dummy regression to remove the noise data (outliers) [14]. The studentized residuals can be obtained by constructing a dummy variable representing an observation that is suspected to be an outlier and including the dummy variable in the regression model. If the dummy variable coefficient for a particular case is significant, it indicates that the observation is an outlier.

TABLE V
AGGREGATED LOAD FORECASTING RESULTS FOR DIFFERENT PARTITION SCHEME

	MAE (MW)	MAPE (%)
One area (Minimum partition)	28.14	3.36
6 areas (Optimal Partition)	23.80	2.69
24 areas (Maximum Partition)	26.68	3.16

B. Numerical Results

The criteria to compare the forecasting performance are the mean absolute error (MAE) and mean absolute percentage error (MAPE) in this paper, which indicate the accuracy of recall.

MAE is defined as

$$\text{MAE} = \sum_{i=1}^{n} \left(|d_{ai} - d_{fi}|\right)/n \tag{5}$$

where d_{ai} is the actual value, d_{fi} is the forecast value, and n is the total number of values predicted.

MAPE is given as follows:

$$\text{MAPE} = \sum_{i=1}^{n} \left(|d_{ai} - d_{fi}|^{*}100/d_{ai}\right)/n. \tag{6}$$

The optimal region partition scheme has been obtained by using the proposed searching procedure. In this numerical experiment, the 24 areas have been eventually reorganized into six groups, consisting of 4, 3, 3, 7, 3, and 4 corporative member areas, respectively. The forecasting results for minimum region partition (one area), optimal region partition, and maximum region partition (24 areas) are given in Table V. It can be seen that the optimal region partition achieved the best performances compared with the minimum and maximum region partition scheme, demonstrating the effectiveness of the proposed multiregion forecasting system.

We also investigated the forecasting results of the individual area for the three partition schemes. The ranges and algebraic average of MAPE for different areas are shown in Table VI.

Putting the results in Tables V and VI together, we can have two observations. First, the forecasting error of the individual area will generally increase when more areas have been partitioned, particularly when the maximum partition has been reached (24 areas); the MAPE for some areas can become

TABLE VI
REGIONAL FORECASTING RESULTS UNDER DIFFERENT PARTITION SCHEMES

	MAPE range (%)	Algebraic average (%)
One area (Minimum partition)	3.36	3.36
6 areas (Optimal Partition)	2.66~5.09	4.03
24 areas (Maximum Partition)	2.88~20.06	>8

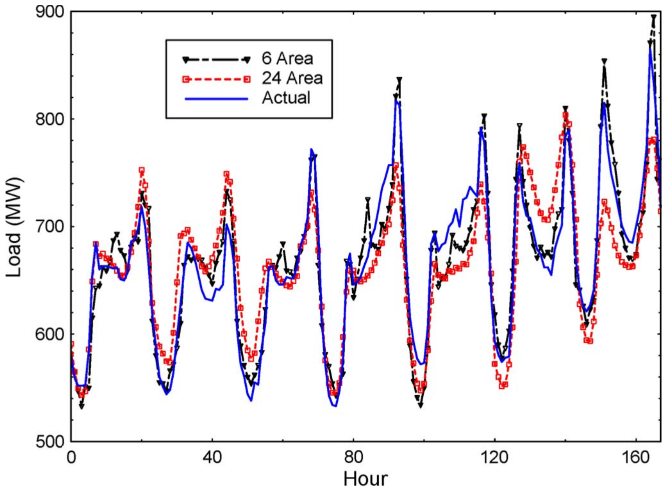

Fig. 7. Forecast results of one week (from March 30 to April 5).

considerably large, mainly because the capacities in such areas are so small that a single load may result in significant variation in the load curve, which is highly unpredictable. Another observation is that the MAPE of the aggregated load forecasting is generally lower than that of forecasting for the individual area, which becomes more significant as more areas have been partitioned.

Based on the two observations, we can further interpret the inference in Fig. 5 as follows. Initially, the number of partitioned areas increases until the optimal region partition has been reached. At this stage, although the forecasting error in the individual area may increase due to the loss of spatial smoothing effect, the overall forecasting error still goes down because of the counteracting effect of diverse error distributions in different areas. As the number of partitioned areas goes beyond the optimal number, the aggregated load forecasting error then intends to increase because the losing spatial smoothing effect overwhelms the counteracting effect of error distributions in different areas.

Finally, Fig. 7 shows the forecasting results for the testing data set. To present the load curve clearly, data of one week (from March 30 to April 5) have been selected for illustration.

V. CONCLUSION

In this paper, the load forecasting problem for a power system with large geographical area in the Midwest U.S., has been investigated. We analyze the weather and load characteristics within the control area and quantify the load diversity of the system.

Based on the fact that diverse load patterns exist in this power system, a multiregion load forecasting model has been developed. The proposed forecasting system can find the optimal region partition under diverse weather and load conditions and achieve more accurate aggregated load forecasting. The numerical results by using the real data from the utility validate the effectiveness of the proposed methodology. Discussions on influences of region partition to forecasting accuracy have also been given in this paper, which could be useful for forecasting practice.

REFERENCES

[1] G. Gross and F. Galiana, "Short term load forecasting," *Proc. IEEE*, vol. 75, no. 12, pp. 1558–1573, Dec. 1987.

[2] H. S. Hippert, C. E. Pedreira, and R. Castro, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans. Power Syst.*, vol. 16, no. 1, pp. 44–55, Feb. 2001.

[3] C. J. Ziser, Z. Y. Dong, and T. Saha, "Investigation of weather dependency and load diversity on queensland electricity demand," in *Proc. Australasian Univ. Power Eng. Conf.*, Sep. 25–28, 2005, pp. 457–462.

[4] J. D. McQuigg, S. R. Johnson, and J. R. Tudor, "Meteorological diversity-load diversity, a fresh look at an old problem," *J. Appl. Meteorol.*, vol. 11, no. 4, pp. 561–566, Jun. 1972.

[5] A. Sargent, R. P. Broadwater, J. C. Thompson, and J. Nazarko, "Estimation of diversity and kWHR-to-peak-kW factors from load research data," *IEEE Trans. Power Syst.*, vol. 9, no. 3, pp. 1450–1456, Aug. 1994.

[6] H. Lee Willis, *Spatial Electric Load Forecasting*. New York: Marcel Dekker, 1996.

[7] C. Cortes and V. Vapnik, "Support-vector network," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[8] K.-R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in *Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 243–254.

[9] B.-J. Chen, M.-W. Chang, and C.-J. Lin, "Load forecasting using support vector machines: A study on EUNITE competition 2001," *IEEE Trans. Power Syst.*, vol. 19, no. 4, pp. 1821–1830, Nov. 2004.

[10] S. Fan and L. Chen, "Short-term load forecasting based on an adaptive hybrid method," *IEEE Trans. Power Syst.*, vol. 21, no. 1, pp. 392–401, Feb. 2006.

[11] K. B. Song, Y. S. Baek, D. H. Hong, and G. Jang, "Short-term load forecasting for the holidays using fuzzy linear regression method," *IEEE Trans. Power Syst.*, vol. 20, no. 1, pp. 96–101, Feb. 2005.

[12] J. N. Fidalgo and J. A. Pecas Lopes, "Load forecasting performance enhancement when facing anomalous events," *IEEE Trans. Power Syst.*, vol. 20, no. 1, pp. 408–415, Feb. 2005.

[13] C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[14] J. Fox, *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage, 1997.

Shu Fan (M'08) received the B.S., M.S., and Ph.D. degrees from the Department of Electrical Engineering, Huazhong University of Science and Technology, Wuhan, China, in 1995, 2000, and 2004, respectively.

From 2004 to 2006, he conducted postdoctoral research, sponsored by the Japanese Government, at Osaka Sangyo University, Daito, Japan. From 2006 to 2007, he was a Visiting Scholar at the Energy Systems Research Center, University of Texas, Arlington. He is currently a Senior Research Fellow at Monash University, Clayton, Australia. His research interests include energy system forecasting, power-system control, and high-power electronics.

Kittipong Methaprayoon (S'03–M'07) received the B.S. degree in electrical engineering from Chulalongkorn University, Bangkok, Thailand, in 2000, and the M.S. and Ph.D. degrees in electrical engineering from the University of Texas, Arlington, in 2003 and 2007, respectively.

He is currently with the Electric Reliability Council of Texas, Austin. His research interests include power-system analysis, applications of artificial neural networks to power systems, and generation planning in a deregulated electricity market.

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

In 1985, he joined the University of Texas, Arlington, where he is currently a Professor in 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, and deregulation for utility companies.

Dr. Lee is a Registered Professional Engineer in the State of Texas.