Short-Term Prediction of Wind Farm Power: A Data Mining Approach

Andrew Kusiak, *Member, IEEE*, Haiyang Zheng, and Zhe Song, *Student Member, IEEE*

***Abstract*—This paper examines time series models for predicting the power of a wind farm at different time scales, i.e., 10-min and hour-long intervals. The time series models are built with data mining algorithms. Five different data mining algorithms have been tested on various wind farm datasets. Two of the five algorithms performed particularly well. The support vector machine regression algorithm provides accurate predictions of wind power and wind speed at 10-min intervals up to 1 h into the future, while the multilayer perceptron algorithm is accurate in predicting power over hour-long intervals up to 4 h ahead. Wind speed can be predicted fairly accurately based on its historical values; however, the power cannot be accurately determined given a power curve model and the predicted wind speed. Test computational results of all time series models and data mining algorithms are discussed. The tests were performed on data generated at a wind farm of 100 turbines. Suggestions for future research are provided.**

***Index Terms*—Data mining algorithms, multiperiod prediction, multiscale prediction, time series model, wind farm power prediction.**

I. Introduction

WIND POWER generation is rapidly expanding into a large-scale industry. As most wind farms are relatively new, it is natural that their performance has not been adequately studied. Prediction of the power produced by a wind farm at different time scales is of interest to the electricity grid.

A number of different approaches have been applied to forecast wind speed and the power produced by wind farms. Potter and Negnevitsky [6] applied the adaptive neurons fuzzy inference approach to forecast short-term wind speed and direction. Barbounis *et al.* [20] used the nonlinear recursive least-squares method to train a recurrent neural network (NN) based on the meteorological data. Their model has improved the accuracy of long-term wind speed and power forecasting. Damousis *et al.* [8] developed a fuzzy logic model and trained it with a genetic algorithm. The model was then used to forecast wind speed over horizons ranging from 0.5 to 2 h. Li *et al.* [19] compared regression and NN models for wind turbine power estimation, and reported that the NN model outperformed the regression model. Sfetsos [3] presented a novel method to forecast the mean hourly wind speed using a time series analysis, and showed that the developed model outperformed the conventional forecasting models. Torres *et al.* [23] built the autoregressive moving average (ARMA) model based on time series data after transformation and standardization, and predicted mean hourly wind speed for up to 10 h ahead.

Developing prediction models for wind farms is a challenge, as the power is mainly determined by the wind speed that is difficult to forecast accurately. The wind speed depends on parameters such as air pressure, temperature, terrain topography, etc. The stochastic nature of a wind farm environment calls for new modeling approaches to accurately predict the power to be produced in the future time periods.

Data mining is a promising approach to model wind farm performance. Numerous successful applications of data mining in manufacturing, marketing, medical informatics, and the energy industry have been reported in the literature [1], [10], [15], [16].

In this paper, a data mining approach has been applied to build time series models for the prediction of wind farm power over short horizons, e.g., 10–70 min as well as longer horizons, e.g., 1–4 h. Two different methodologies for power prediction have been employed. The models are built using historical data collected by supervisory control and data acquisition (SCADA) systems installed at a wind farm. A short-term power prediction is important in dispatching power to meet customer needs. For long horizon predictions, meteorological data are usually used.

II. Basic Methodologies for Wind Power Prediction

A. Time Series Prediction Modeling

Time series prediction [5], [24] focuses on determining future events based on known events, measured typically at successive times and spaced at (often uniform) time intervals. The basic time series prediction model is as follows [11]:

$$\hat{y}(t+T) = f(y(t), y(t-T), \ldots, y(t-mT)) \qquad (1)$$

where T is the sampling time (time interval), $\hat{y}(t+T)$ is the predicted parameter, $y(t), y(t-T), \ldots, y(t-mT)$ are the current and past observed parameters, and $m+1$ is the number of inputs (predictors) to the model.

To obtain an accurate prediction model with a data mining approach, appropriate predictors need to be selected. Data mining offers different algorithms to perform this task. For example, the boosting tree algorithm [13], [14] can be used to select the best predictors, as well as the wrapper approach [26] using the genetic or the best-first search algorithms [9], [12].

To maximize the performance of the prediction model, a boosting tree algorithm was employed to select a set of the most important predictors among $\{y(t), y(t-T), \ldots, y(t-mT)\}$. Two metrics, i.e., the absolute error (2) and the relative error (3)

Manuscript received March 5, 2008; revised May 27, 2008; accepted September 4, 2008. First published January 13, 2009; current version published February 19, 2009. This work was supported by Iowa Energy Center under Grant IEC 07-01. Paper no. TEC-00080-2008.

The authors are with the Intelligent Systems Laboratory, Department of Mechanical and Industrial Engineering, The University of Iowa, Iowa City, IA 52242-1527 USA (e-mail: andrew-kusiak@uiowa.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/TEC.2008.2006552

0885-8969/$25.00 © 2009 IEEE

Authorized licensed use limited to: The University of Iowa. Downloaded on February 19, 2009 at 20:53 from IEEE Xplore. Restrictions apply.

TABLE I
DATASET DESCRIPTION

Data set	Start Time Stamp	End Time Stamp	Description
1	1/1/06 1:40 AM	1/31/06 11:50 PM	Total data set; 4455 observations
2	1/1/06 1:40 AM	1/25/06 8:00 PM	Training data set; 3568 observations
3	1/25/06 8:10 PM	1/31/06 11:50 PM	Test data set; 887 observations

were used to select the accurate model (1) extracted with data mining algorithms.

$$\text{Absolute error} = |\hat{y}(t+T) - y(t+T)| \tag{2}$$

$$\text{Relative error} = \left| \frac{\hat{y}(t+T) - y(t+T)}{y(t+T)} \right| \times 100\% \tag{3}$$

where $\hat{y}(t+T)$ is the predicted parameter and $y(t+T)$ is the observed (measured) parameter.

B. Data Description

The data used in this research were generated at a wind farm of 100 turbines. The data were collected by a SCADA system installed at each wind turbine. The SCADA system of every wind turbine collects data for more than 120 parameters. Though the data are sampled at high frequency, e.g., 2 s, it is averaged and stored at 10-min intervals (referred to as the 10-min average data). The data used in this research were collected over a period of 1 month for all turbines of the wind farm. The wind speed was shown to be an important predictor of wind farm power in the previous research [25]. The time series prediction models for wind speed and wind farm power are discussed in Sections III and IV.

The wind speed and power recorded at 10-min intervals resulted in 4455 instances (dataset 1 in Table I), starting from "January 1, 2006 1:40 A.M." and continuing up to "January 31, 2006 11:50 P.M." During this time period, the overall wind farm performance was normal. Dataset 1 was divided into two subsets, dataset 2 and dataset 3. Dataset 2 contains 3568 data points and was used to develop a prediction model with data mining algorithms. Dataset 3 includes 887 data points and was used to test the prediction performance of the model learned from dataset 2. For the testing data, the mean and standard deviation of the two statistical measures (2) and (3) are the most important indicators for selecting the data mining algorithms to learn model (1) of Section II.A.

C. Feature Selection

Important predictors are determined by the importance index generated by the boosting tree algorithm [13], [14]. The basic idea of the boosting tree algorithm is to build a number of trees (e.g., binary trees) splitting the dataset and approximating the underlying function. The importance of each predictor is measured by its contribution to the prediction accuracy on the training dataset.

It is not surprising to observe that the importance of the predictors (past and present values of the model (1)) is ranked in the order of time sequence. The ranking order is as follows: $I[y(t)] > I[y(t-T)] > I[y(t-2T)], \ldots, I[y(t-(m-1)T] > I[y(t-mT)]$, where $I[\cdot]$ is the importance (rank) of predictors.

TABLE II
RANK ORDER OF PREDICTORS

Predictor	Variable Rank	Importance
y(t)	100	1.000
y(t - T)	91	0.907
y(t - 2T)	88	0.877
y(t - 3T)	85	0.845
y(t - 4T)	80	0.796
y(t - 5T)	76	0.755
y(t - 6T)	73	0.731
y(t - 7T)	72	0.719
y(t - 8T)	69	0.688
y(t - 9T)	66	0.665

In this paper, three different models are considered: a 10-min time series model of wind speed, a 10-min time series model of wind farm power, and a 1-h time series model of mean hourly wind farm power.

The performance of the hourly time series model applied to wind speed is rather poor; however, the wind farm power time series model performs well. Therefore, the wind speed results produced by the hourly time series model are not discussed. Note that here the mean hourly power is the average of the wind farm power produced over an hour.

Table II shows the importance of ten predictors computed by the boosting tree algorithm based on the 10-min power data.

It is important to select predictors with the highest information content among $\{y(t), y(t-T), \ldots, y(t-mT)\}$ to maximize prediction accuracy. A threshold value of 0.75 has been established heuristically to select the predictors for the three models. For the 10-min time series models of wind speed and power, six predictors $\{y(t), y(t-T), y(t-2T), y(t-3T), y(t-4T), y(t-5T)\}$ have been selected. For the hourly time series model of power, the selected predictors are $\{y(t), y(t-T), y(t-2T), y(t-3T)\}$.

The threshold value of 0.75 used in computation has produced good quality results. A lower threshold value leads to more predictors. A large number of predictors could result in inferior performance of extracted models due to "the course of dimensionality" principle.

D. Multiperiod Predictions

Using the selected predictors, model (1) predicts the values of wind speed and power at future time periods.

Fig. 1(a)–(c) illustrates the concept of a multiperiod prediction with the 10-min time series model. In this model, the sampling time T is 10 min. In Fig. 1(a), using the average measured values at the intervals $[t=-60,\ t=-50)$, $[t=-50,\ t=-40), \ldots,\ [t=-10,\ t=0-)$, the average value at the subsequent interval $[t=0,\ t=10)$ is predicted. In Fig. 1(b), based on the average measured value at the intervals $[t=-50, t=-40), [t=-40, t=-30), \ldots, [t=-10, t=0)$ and

Authorized licensed use limited to: The University of Iowa. Downloaded on February 19, 2009 at 20:53 from IEEE Xplore. Restrictions apply.

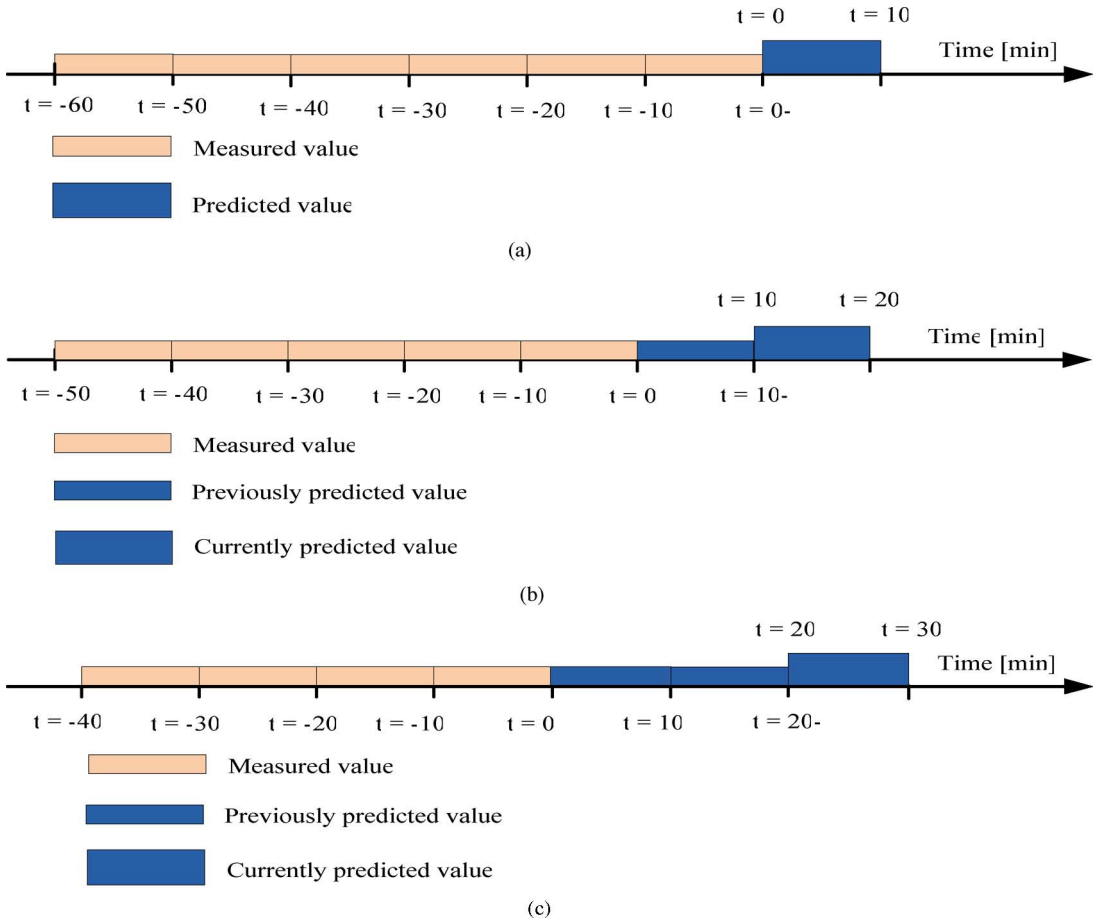

Fig. 1. Description of the 10-min time series prediction model. (a) 10-min ahead prediction. (b) 20-min ahead prediction. (c) 30-min ahead prediction.

previously predicted value at $[t = 0,\ t = 10-)$ as predictors, the average value at the subsequent interval $[t = 10, t = 20)$ is predicted. In Fig. 1(c), using the average measured values at the intervals $[t = -40,\ t = -30) \ldots, [t = -10,\ t = 0)$ and previously predicted values at $[t = 0,\ t = 10)$ and $[t = 10,\ t = 20-)$ as predictors, the average value at the subsequent interval $[t = 20, t = 30)$ is predicted. Similarly, the average values at intervals $[t = 30,\ t = 40)$, $[t = 50,\ t = 60)$ are predicted.

Fig. 2(a)–(c) illustrate the concept of a multiperiod prediction with the hourly time series model. In this model, the sampling time T is an hour. In Fig. 2(a), using the mean measured hourly power at the intervals $[t = -4,\ t = -3)$, $[t = -3,\ t = -2)$, $[t = -2, t = -1)$, $[t = -1, t = 0-)$, the mean hourly power at the subsequent interval $[t = 0,\ t = 1)$ is predicted. In Fig. 2(b), based on the mean measured hourly power at the intervals $[t = -3,\ t = -2)$, $[t = -2,\ t = -1)$, $[t = -1,\ t = 0)$ and the previously predicted value at $[t = 0,\ t = 1-)$ as predictors, the mean hourly power at the subsequent interval $[t = 1,\ t = 2)$ is predicted. In Fig. 2(c), using the mean measured hourly power at the intervals $[t = -2,\ t = -1)$, $[t = -1,\ t = 0)$ and the previously predicted values at $[t = 0,\ t = 1)$ and $[t = 1,\ t = 2-)$ as predictors, the mean hourly power at the subsequent interval $[t = 2,\ t = 3)$ is predicted. Similarly, the mean hourly power at $[t = 3,\ t = 4)$ is predicted.

E. Integrated k Nearest Neighbor (kNN) and Time Series Prediction Model

Based on the time series prediction model, two ways to predict the short-term wind farm power are proposed. One is to directly use the power values measured in the past to predict the future power. The other is to use the wind speed measured in the past to predict the future wind speed first, and then use the predicted wind speed to predict the wind farm power.

The basic equation of wind power density [2] is shown in

$$P_w = 0.5\rho v^3 \tag{4}$$

where P_w is the power density (watts per square meter), ρ is the air density (in kilogram per cubic centimeter), and v is the horizontal component of the mean freestream wind velocity (meter per second).

As the hub of the turbine is usually located 60–80 m above the ground, the air density ρ is frequently considered constant at that height. Though wind direction changes, the yaw position is controlled to face the wind to capture the maximum energy from the wind. Therefore, the wind speed is a significant predictor of the wind farm power.

Knowing the wind speed, power produced by a wind farm is usually computed using the power curve function provided

Authorized licensed use limited to: The University of Iowa. Downloaded on February 19, 2009 at 20:53 from IEEE Xplore. Restrictions apply.

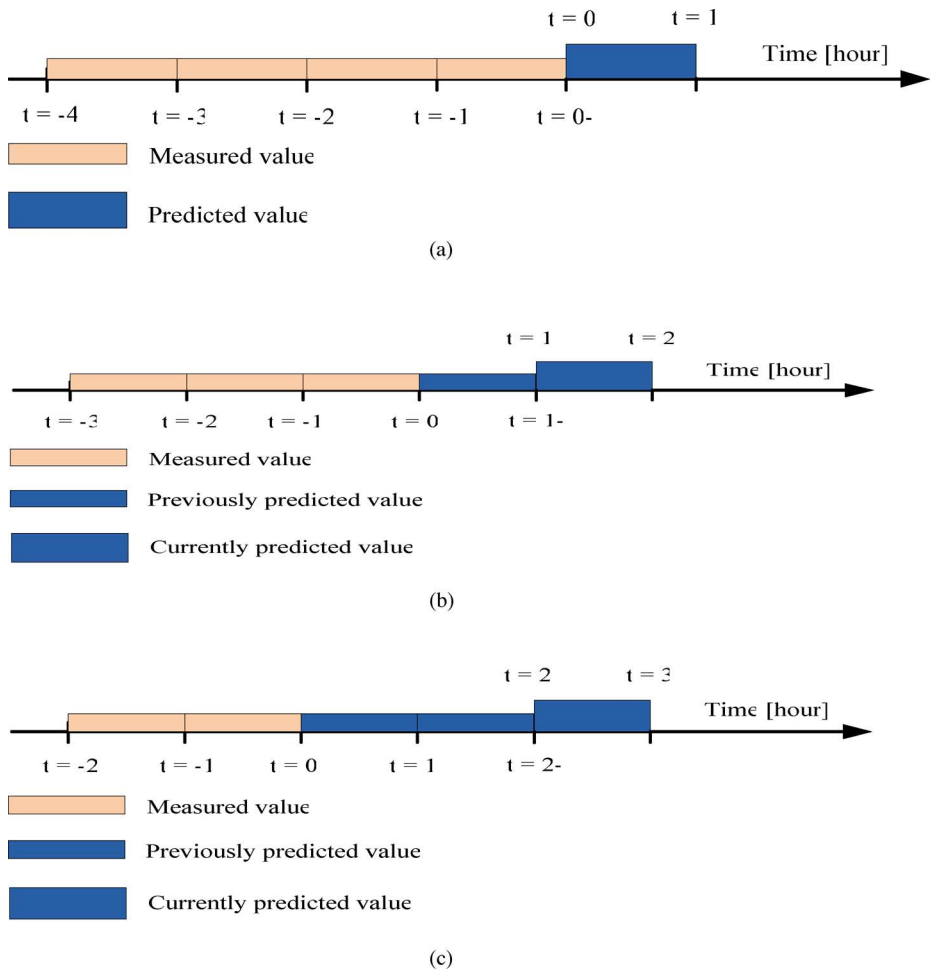

Fig. 2. Description of the hourly time series prediction model. (a) 1-h ahead prediction. (b) 2-h ahead prediction. (c) 3-h ahead prediction.

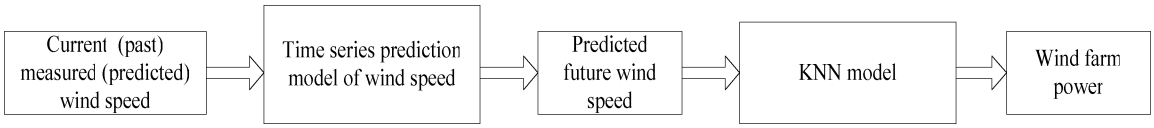

Fig. 3. Structure of the integrated prediction model.

by the turbine manufacturer. For wind farms containing diverse type turbines, different power curves are used. Factors such as turbine location, wind conditions, operations, and control make the actual power curve different from the one offered by the turbine manufacturer. Using the manufacturer's power curve function to compute the wind farm power output for known wind speed leads to significant errors due to the difference between the actual power curve and the one provided by the manufacturer.

Computing wind farm power, based on the current wind speed, has been discussed in the literature [25]. It has been shown that the kNN model [9] accurately determines wind farm power, given the wind speed collected at the corresponding 10-min time interval. In this paper, the time series model predicting the wind speed and the kNN model are combined to predict wind farm power as shown in Fig. 3.

In Fig. 3, the future wind speed is predicted with the time series model. Then, the predicted wind speed is used as an input of the kNN model to compute the wind farm power.

III. Wind Speed Time Series Prediction

A. Algorithm Selection for the 10-min Time Series Prediction Model

The important predictors $\{y(t), y(t-T), y(t-2T), y(t-3T), y(t-4T), y(t-5T)\}$ of the time series model for wind speed have been selected in Section II-C. The relative error (2) and absolute error (3) have been used to select the most suitable algorithm for building the time series model (1).

Five data mining algorithms that appeared to be the most promising have been used to construct the 10-min time series

Authorized licensed use limited to: The University of Iowa. Downloaded on February 19, 2009 at 20:53 from IEEE Xplore. Restrictions apply.

TABLE III
ERROR STATISTICS OF DIFFERENT MODELS BASED ON DATASET 3 OF TABLE I

Algorithm	Mean Absolute Error (m/s)	Absolute Error Std (m/s)	Mean Relative Error (%)	Relative Error Std (%)
SVMreg	0.198	0.181	3.514	5.730
MLP	0.266	0.388	4.661	6.739
M5P tree	0.206	0.267	3.953	6.035
REP tree	0.249	0.256	4.576	6.564
Bagging tree	0.202	0.198	3.665	5.723

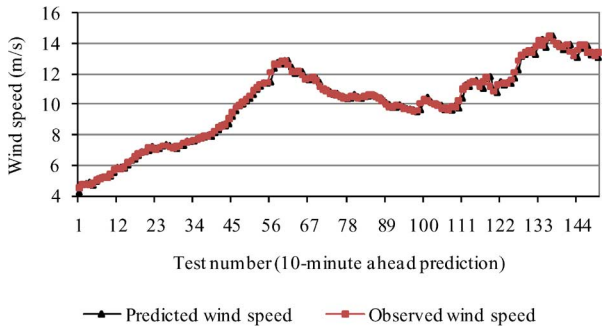

Fig. 4. Predicted and observed wind speed for the first 150 data points from dataset 3 of Table I.

prediction model (1) for wind speed. They include: the support vector machine regression (SVMreg) algorithm [4], [18], multilayer perceptron (MLP) algorithm [9], [17], M5P tree algorithm [7], [9], [27], Reduced Error Pruning (REP) tree (decision or regression tree) [9], [28], and the bagging tree [9], [21], [22]. Table III summarizes the model's prediction accuracy based on the dataset 3 of Table I. The SVMreg algorithm outperformed the other four algorithms. The MLP and REP tree algorithms performed the worst. The SVMreg algorithm was finally selected for building the 10-min time series wind speed prediction model.

Fig. 4 shows the first 150 observed and predicted wind speeds from dataset 3 of Table I (the 10-min ahead prediction). It is obvious that the observed and predicted wind speeds are almost identical.

B. Multiperiod Prediction With 10-min Time Series Model

Based on the approach described in Section II-D, the time series model built by the SVMreg algorithm is used for multiperiod predictions. The test dataset for the 10-min ahead predictions containing 887 data points is reduced by one, when the prediction period moves forward by one step. Figs. 5(a)–(e) illustrates the first 150 observed and predicted wind speeds at 20-, 30-, 40-, 50-, and 60-min ahead periods, respectively.

Tables IV and V show the absolute and relative errors for each of the one- to six-period ahead predictions. The standard deviation, mean, and maximum of the absolute and relative errors all increase as the prediction horizon increases. However, the minimum error does not substantially increase, which could be due to the relative stability of the wind.

Persistent forecasting with traditional methods has been studied in the literature [5], [29]. The time series model built by data mining algorithms has enhanced the prediction accuracy by at least 20%. Further improvements can likely be made as data mining offers a variety of algorithms.

IV. WIND FARM POWER TIME SERIES PREDICTION

A. Algorithm Selection for the 10-min Time Series Prediction Model

The total power of the wind farm analyzed in this section was scaled to the interval [0, 100 MW].

Five of the most promising data mining algorithms (the same as in Section III-A) were applied to extract the 10-min time series prediction model of wind power. Table VI summarizes the prediction accuracy based on the dataset 3 of Table I. The SVMreg algorithm outperformed the other four algorithms. The MLP and REP tree algorithms performed the worst. Therefore, the SVMreg algorithm was finally selected to build the 10-min time series prediction model (1) of wind farm power.

Fig. 6 shows the first 200 observed and predicted (10-min ahead) wind power values from dataset 3 of Table I. It is easy to see that both the observed and predicted wind farm power values are almost identical.

B. Multiperiod Power Prediction With the 10-min Time Series Model

In this section, the time series model learned by the SVMreg algorithm is used to predict the wind farm power over 10-min intervals, 10–60 min ahead into the future. The test data for the 10-min prediction used in Section IV-A containing 887 points is reduced by one for each of the next 10-min period predictions.

Fig. 7(a)–(e) shows the first 200 observed and predicted wind farm power values over 20-, 30-, 40-, 50-, and 60-min future time intervals, respectively.

Tables VII and VIII summarize the statistics of absolute and relative errors of the future power prediction over six different 10-min intervals. The mean, the standard deviation, and the maximum error all increase when the prediction horizon increases. However, the minimum error remains relatively stable. The relative errors reported in Table VIII are for power when the wind farm power is greater than 7 MW (7% of the maximum power), as the relative error at the low power output is usually large, while the absolute error remains small. Note that the prediction accuracy with smaller relative error is more meaningful when the power generated is greater than 7 MW.

C. Algorithm Selection for the Hourly Time Series Prediction Model

The important predictors of the hourly time series model of wind farm power that were selected by the boosting tree algorithm are $\{y(t), y(t-T), y(t-2T), y(t-3T)\}$. The sampling time for this model is an hour, and thus all variables need to be at the same time scale. As the original dataset 1 in Table I contains 10-min average data, every six consecutive data points are aggregated into a mean hourly power value (the average of six measured power values). Therefore, the original 4455 dataset has been aggregated into a dataset with 742 data points

Authorized licensed use limited to: The University of Iowa. Downloaded on February 19, 2009 at 20:53 from IEEE Xplore. Restrictions apply.

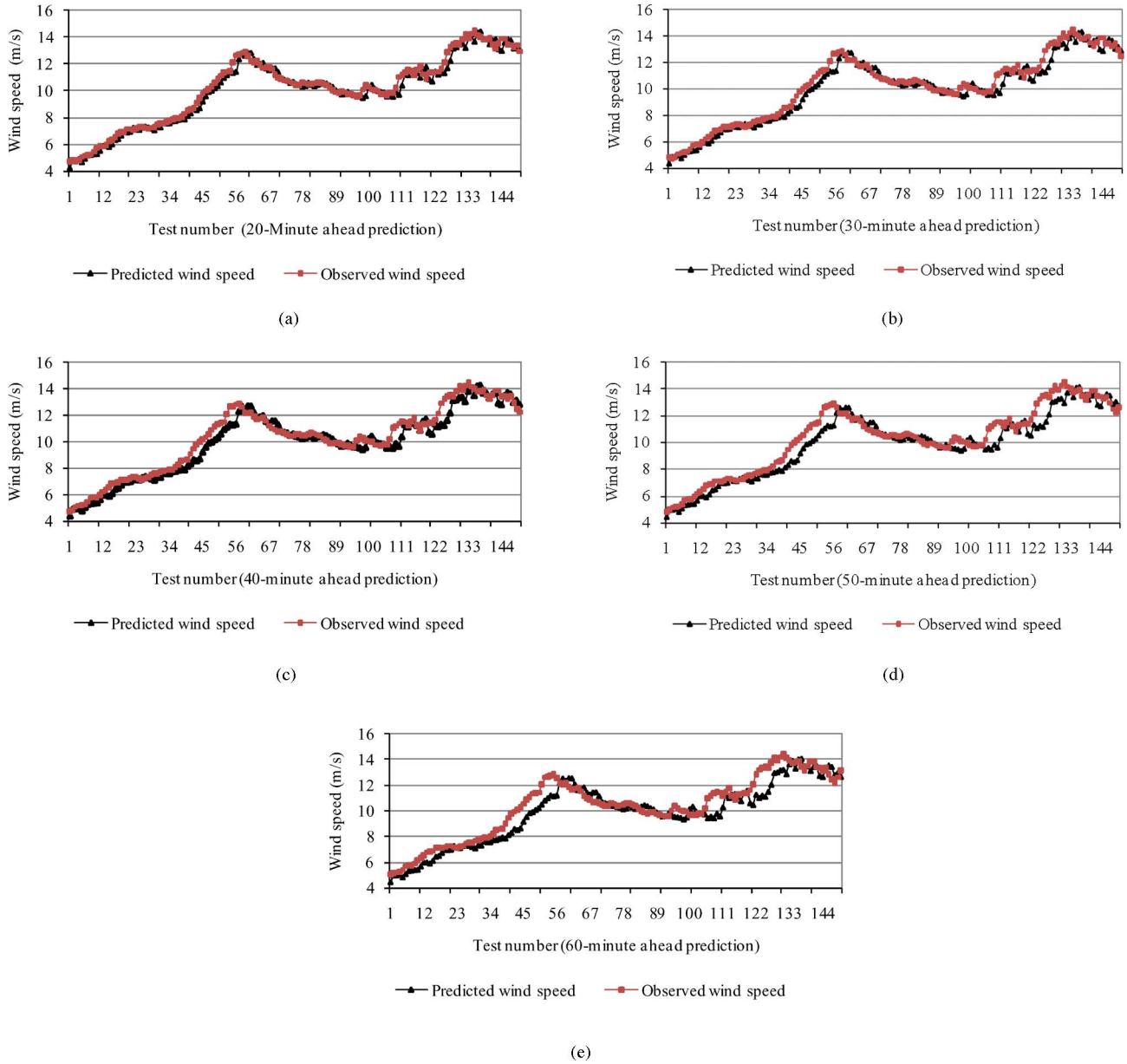

Fig. 5. Observed and predicted wind speeds at different periods for the first 150 data points from dataset 3 of Table I. (a) 20-min ahead prediction. (b) 30-min ahead prediction. (c) 40-min ahead prediction. (d) 50-min ahead prediction. (e) 60-min ahead prediction.

TABLE IV
ABSOLUTE ERROR STATISTICS FOR 10–60-min AHEAD PREDICTIONS

Absolute Error (m/s)	Mean	Standard Deviation	Maximum	Minimum
10-minute ahead prediction	0.198	0.181	1.568	0.001
20-minute ahead prediction	0.361	0.311	2.300	0.001
30-minute ahead prediction	0.485	0.403	2.475	0.001
40-minute ahead prediction	0.580	0.474	2.923	0.004
50-minute ahead prediction	0.657	0.542	3.050	0.001
60-minute ahead prediction	0.733	0.593	3.119	0.001

TABLE V
RELATIVE ERROR STATISTICS FOR 10–60-min AHEAD PREDICTIONS

Relative Error (%)	Mean	Standard Deviation	Maximum	Minimum
10-minute ahead prediction	3.514	5.730	80.239	0.008
20-minute ahead prediction	6.659	10.859	129.826	0.006
30-minute ahead prediction	9.294	15.613	180.392	0.029
40-minute ahead prediction	11.322	20.141	260.219	0.041
50-minute ahead prediction	12.976	24.763	353.865	0.009
60-minute ahead prediction	14.571	28.139	425.971	0.028

described in Table IX. This dataset begins from "January 1, 2006 1:30 A.M." and continues to "January 31, 2006 11:30 P.M." All the time stamps in Table IX denote the hourly intervals, e.g., "January 1, 2006 5:30 A.M." represents the mean hourly power during the hourly interval from "January 1, 2006 4:30 A.M." to "January 1, 2006 5:30 A.M." Dataset 1 of Table I was divided

Authorized licensed use limited to: The University of Iowa. Downloaded on February 19, 2009 at 20:53 from IEEE Xplore. Restrictions apply.

TABLE VI
ERROR STATISTICS OF FIVE DIFFERENT ALGORITHMS BASED ON DATASET 3 OF TABLE I

Algorithm	Mean Absolute Error (MW)	Absolute Error Std (MW)	Mean Relative Error (%)	Relative Error Std (%)
SVMreg	2.147	2.499	6.613	7.785
MLP	2.981	3.456	8.973	9.873
M5P tree	2.354	2.578	6.892	7.983
REP tree	4.893	5.673	9.123	10.124
Bagging tree	2.335	3.135	8.876	9.754

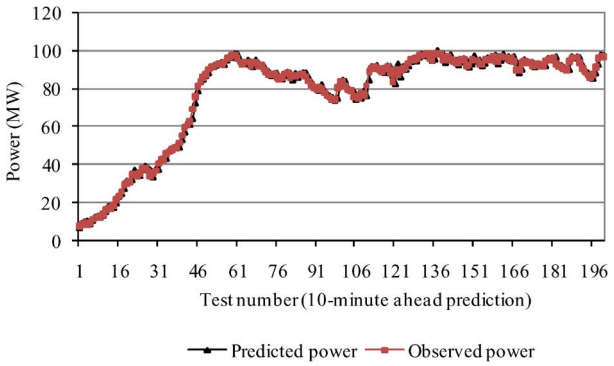

Fig. 6. Observed and predicted wind power for the first 200 data points from dataset 3 of Table I.

into two datasets, dataset 2 and dataset 3. Dataset 2 contains 593 data points and is used to develop a prediction model with data mining algorithms. Dataset 3 includes 149 data points and it tests the prediction performance of the model learned from dataset 2.

Five of the most promising data mining algorithms (the same used previously) were selected to construct the hourly time series model. Table X summarizes the prediction accuracy of different algorithms based on dataset 3 of Table IX. Here, the MLP algorithm outperformed the other four algorithms. The REP algorithm performed the worst. Based on its performance, the MLP algorithm was selected to build the hourly time series model (1) for the mean hourly wind farm power prediction.

Fig. 8 shows the observed and predicted wind power for dataset 3 of Table IX. This is an hour ahead prediction, and the predicted power closely trails the observed power.

D. Four-Period Forward Prediction of the Hourly Time Series Model

Based on the approach discussed in Section II-D, multiperiod predictions were tested. The accuracy of the time series model decreases as the time horizon extends. The test data of Table IX for hourly predictions contains 149 data points, and it will be reduced by one data point for each subsequent prediction interval. Fig. 9(a)–(c) illustrates the observed and predicted mean hourly wind farm power over 2-, 3-, and 4-h ahead time intervals, respectively.

Tables XI and XII show the absolute and relative error statistics for the mean hourly power prediction over four different hourly intervals. The mean, the standard deviation, and the maximum error all increase as the prediction horizon lengthens. The relative error here is computed when the mean hourly power is greater than 7 MW (7% of the maximum power).

V. INTEGRATED MODEL FOR WIND POWER PREDICTION

A. kNN Model for Wind Power Prediction

Integrating the kNN model with the wind speed time series model for power prediction has been inspired by the wind industry practice. The prevailing approach to the wind farm power prediction is to forecast the wind speed first and use it to compute power based on a predefined power curve function.

The kNN is a machine learning algorithm predicting unknown value for an instance (here power) using the data supporting that instance. The predicted value is associated with the majority votes of these k neighbors. Euclidean distance is often used to measure the closeness of the data points.

In this paper, the kNN algorithm predicts wind farm power based on the wind speed. The basic steps of the kNN algorithm are as follows.

1) Represent each instance in a multidimensional space.
2) Divide the entire dataset into training and test datasets.
3) Given a test instance, a distance metric is computed between the test instance and all training instances, then the k nearest neighbor instances are selected from the training data.
4) Compute the distance-weighted (or nonweighted) average of output of the k nearest neighbor instances selected from the training data. This average becomes the predicted value for the test (unknown) instance.

Previous research [9], [25] has shown that the kNN model is quite accurate for computing wind farm power given the wind speed as input. Using the measured wind speed, the wind farm power is computed fairly accurately when the wind farm is operating under normal conditions. The normal conditions exclude states where the wind speed is too low or too high, turbines undergo maintenance, and the turbine power output is low due to other issues. When the wind speed is below the cut-in speed, the power output of a wind turbine is about zero. When the wind speed is above the rated speed, the power output of a wind turbine is almost constant. Removing the corresponding data points allows the kNN model to identify the relationship between wind speed and power output.

To build a kNN model, the original dataset of Table I has been preprocessed into the format shown in Table XIII. Dataset 1 was divided into two data subsets, dataset 2 and dataset 3. Dataset 2 contains 3476 data points and is used to develop a prediction model with the kNN algorithm. Dataset 3 of 871 data points was used to test the prediction performance of the model learned from dataset 2. Table XIV shows the error statistics of the kNN model for dataset 3 of Table XIII.

B. Comparison of the Integrated Model and the Time Series Models

The computational experience reported in the previous sections showed that the kNN algorithm provided accurate power

Authorized licensed use limited to: The University of Iowa. Downloaded on February 19, 2009 at 20:53 from IEEE Xplore. Restrictions apply.

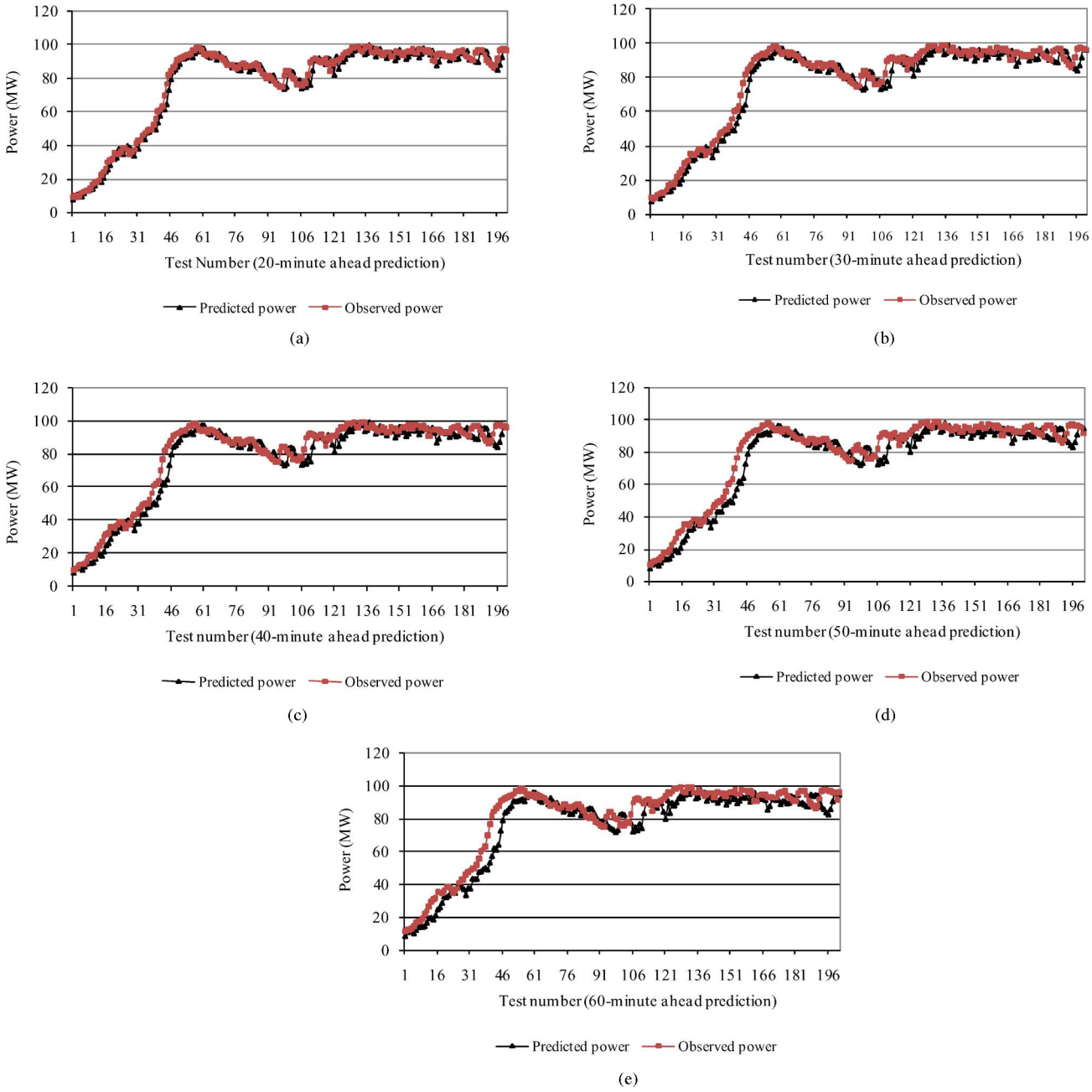

Fig. 7. Observed and predicted wind farm power at different periods for the first 200 data points from dataset 3 of Table I. (a) 20-min ahead prediction. (b) 30-min ahead prediction. (c) 40-min ahead prediction. (d) 50-min ahead prediction. (e) 60-min ahead prediction.

TABLE VII
ABSOLUTE ERROR STATISTICS FOR 10–60-min AHEAD PREDICTIONS

Absolute Error (MW)	Mean	Standard Deviation	Maximum	Minimum
10-minute ahead prediction	2.214	2.503	26.357	0.001
20-minute ahead prediction	3.902	4.085	38.254	0.001
30-minute ahead prediction	5.146	5.152	42.032	0.006
40-minute ahead prediction	6.063	5.92	50.19	0.008
50-minute ahead prediction	6.724	6.57	51.306	0.019
60-minute ahead prediction	7.388	6.99	54.784	0.003

TABLE VIII
RELATIVE ERROR STATISTICS FOR 10–60-min AHEAD PREDICTIONS

Relative Error (%)	Mean	Standard Deviation	Maximum	Minimum
10-minute ahead prediction	6.613	7.785	66.780	0.001
20-minute ahead prediction	11.894	13.696	99.114	0.004
30-minute ahead prediction	15.928	18.7161	150.671	0.011
40-minute ahead prediction	18.741	22.375	198.538	0.018
50-minute ahead prediction	20.873	9.862	189.414	0.022
60-minute ahead prediction	22.997	28.877	250.681	0.005

Authorized licensed use limited to: The University of Iowa. Downloaded on February 19, 2009 at 20:53 from IEEE Xplore. Restrictions apply.

TABLE IX
DESCRIPTION OF THE HOURLY DATASET

Data set	Start Time Stamp	End Time Stamp	Description
1	1/1/06 5:30 AM	1/31/06 11:30 PM	Total data set; 742 observations
2	1/1/06 5:30 AM	1/25/06 6:30 PM	Training data set; 593 observations
3	1/25/06 7:30 PM	1/31/06 11:30 PM	Test data set; 149 observations

TABLE X
PREDICTION ACCURACY OF FIVE DIFFERENT ALGORITHMS BASED ON DATASET 3 OF TABLE IX

Algorithm	Mean Absolute Error (MW)	Absolute Error Std (MW)	Mean Relative Error (%)	Relative Error Std (%)
SVMreg	6.128	9.164	20.591	26.347
MLP	5.937	8.243	19.317	27.217
M5P tree	6.164	9.264	20.691	29.564
REP tree	6.559	10.654	21.397	29.684
Bagging tree	6.367	8.956	21.654	29.354

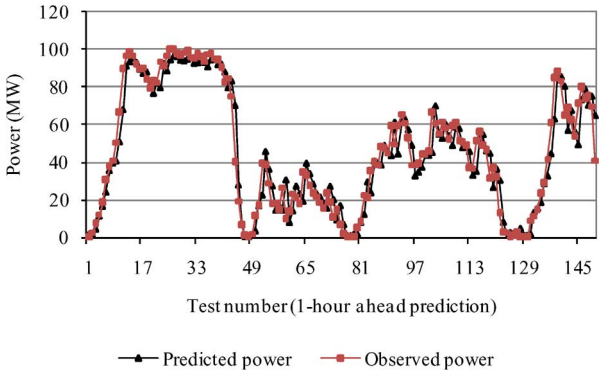

Fig. 8. Observed and predicted hourly power of dataset 3 from Table IX.

predictions. It is desirable to make sure that the wind speed predictions are dependable. As presented in Section II-E, the kNN model and the 10-min time series wind speed prediction model have been integrated to predict future power. The 10-min time series model for the wind speed prediction was discussed in Section III.

The test data set for the 10-min ahead prediction of 871 data points will be reduced by one for each future prediction interval. Fig. 10(a)–(f) shows the first 200 observed and predicted wind farm power over future 10-, 20-, 30-, 40-, 50-, and 60-min time intervals, respectively.

Tables XV and XVI show the absolute and relative error statistics of the integrated prediction model. Like the time series prediction model, the performance of the integrated model decreases as the prediction horizon increases. The mean, the standard deviation, and the maximum error all increase in time. However, the minimum error remains relatively stable.

The computational results reported in this paper have shown that the 10-min time series model for wind power prediction outperforms the integration model. Though the kNN model and the 10-min wind speed time series prediction model perform well individually, the integrated model produces a larger error

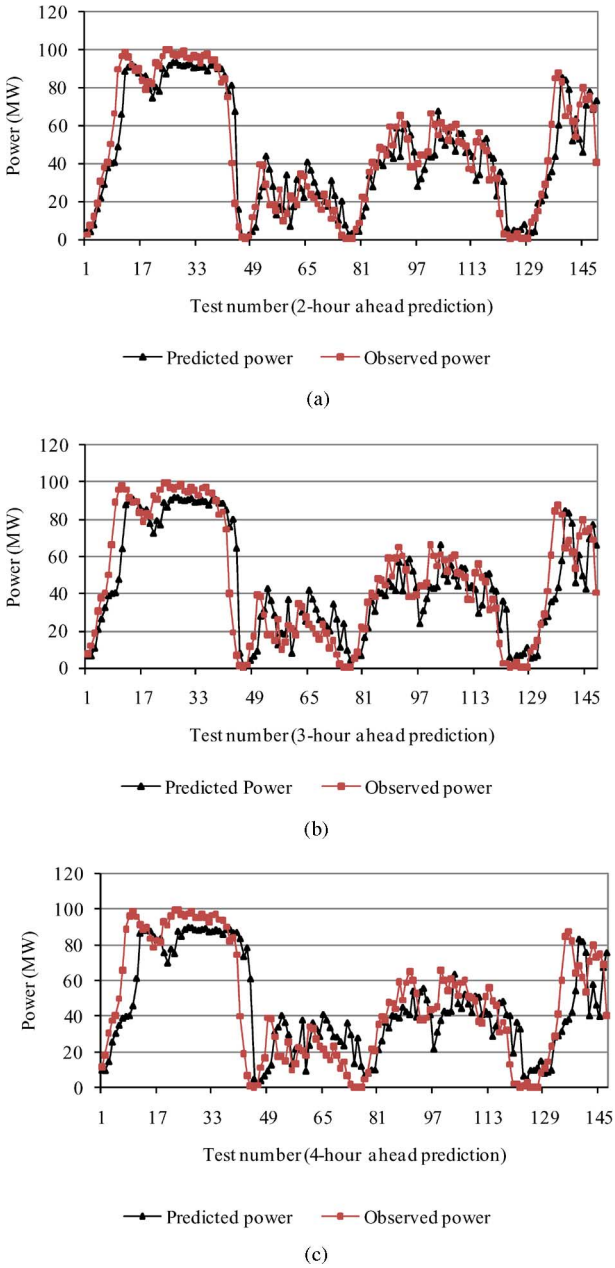

Fig. 9. Observed and predicted hourly power at different future intervals. (a) 2-h ahead prediction. (b) 3-h ahead prediction. (c) 4-h ahead prediction.

TABLE XI
ABSOLUTE ERROR STATISTICS FOR DIFFERENT HOURLY INTERVALS

Absolute Error (MW)	Mean	Standard Deviation	Maximum	Minimum
1-h ahead prediction	5.938	5.738	30.379	0.001
2-h ahead prediction	9.475	9.048	48.828	0.142
3-h ahead prediction	12.004	11.421	61.014	0.016
4-h ahead prediction	15.214	13.399	71.855	0.105

when predicting future power. This could be due to the fact that the power is a cubic function of the wind speed, as indicated by the wind power density function (4) of Section II-E.

Authorized licensed use limited to: The University of Iowa. Downloaded on February 19, 2009 at 20:53 from IEEE Xplore. Restrictions apply.

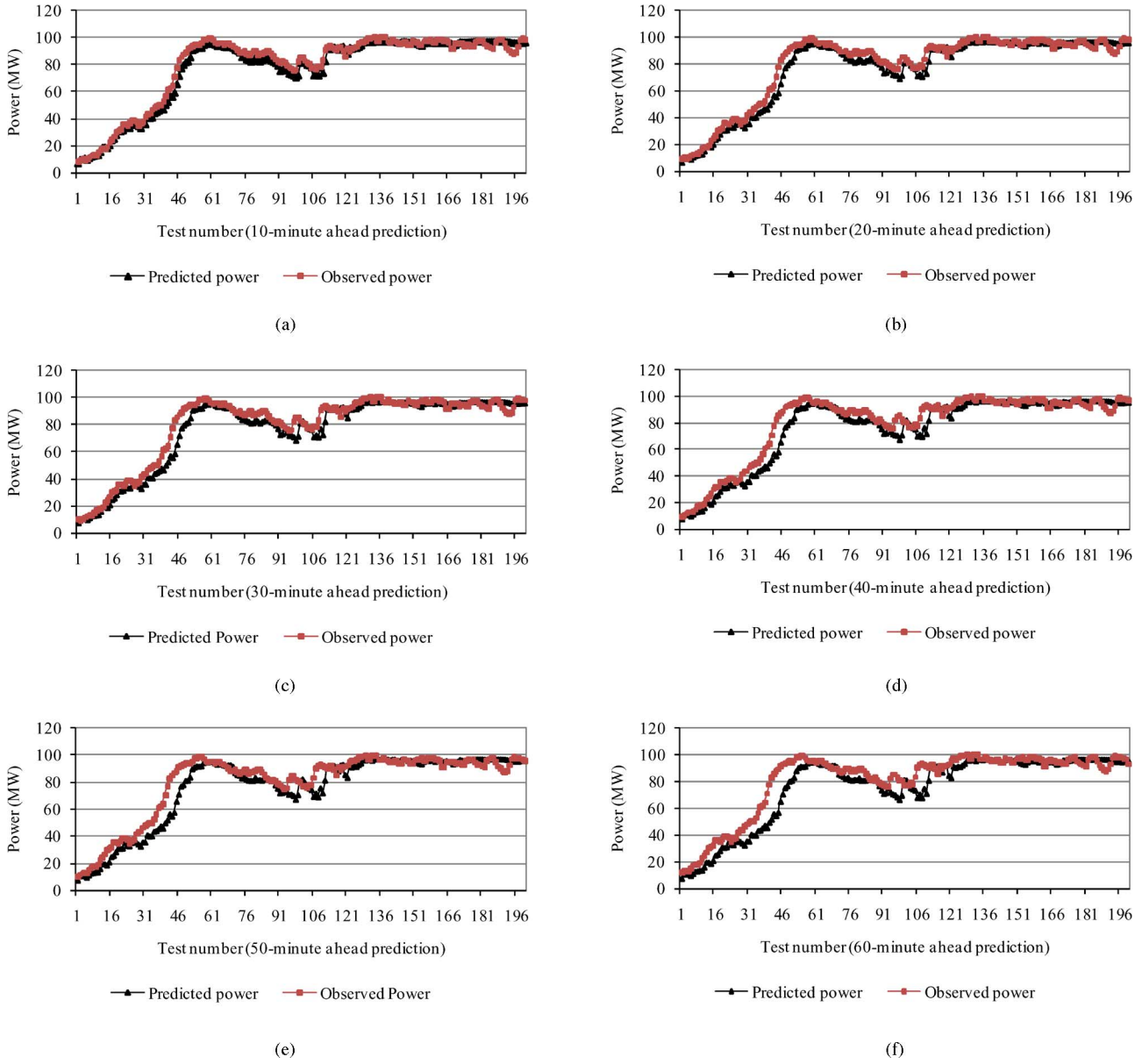

Fig. 10. Observed and predicted power at different future intervals. (a) 10-min ahead prediction. (b) 20-min ahead prediction. (c) 30-min ahead prediction. (d) 40-min ahead prediction. (e) 50-min ahead prediction. (f) 60-min ahead prediction.

TABLE XII
RELATIVE ERROR STATISTICS FOR DIFFERENT HOURLY INTERVALS

Relative Error (%)	Mean	Standard Deviation	Maximum	Minimum
1-h ahead prediction	19.317	27.217	219.845	0.011
2-h ahead prediction	29.210	37.293	258.665	0.158
3-h ahead prediction	31.783	37.887	323.215	0.018
4-h ahead prediction	36.865	40.469	289.159	0.365

Additionally, the wind speed in the kNN model is too sensitive as a predictor for wind farm power, and thus it might lead to a worse prediction for the integration model. The integration of the two models did not improve prediction accuracy.

TABLE XIII
DATASET DESCRIPTION

Data set	Start Time Stamp	End Time Stamp	Description
1	1/1/06 12:00 AM	1/31/06 11:50 PM	Total data set; 4347 observations
2	1/1/06 12:00 AM	1/25/06 6:20 PM	Training data set; 3476 observations
3	1/25/06 6:30 PM	1/31/06 11:50 PM	Test data set; 871 observations

TABLE XIV
ERROR STATISTICS OF THE kNN MODEL

Algorithm	Mean Absolute Error (MW)	Absolute Error Std (MW)	Mean Relative Error (%)	Relative Error Std (%)
kNN (k = 250)	1.689	1.628	4.231	3.167

Authorized licensed use limited to: The University of Iowa. Downloaded on February 19, 2009 at 20:53 from IEEE Xplore. Restrictions apply.

TABLE XV
ABSOLUTE ERROR STATISTICS OVER DIFFERENT 10-min TIME INTERVALS

Absolute Error (MW)	Mean	Standard Deviation	Maximum	Minimum
10-minute ahead prediction	2.915	2.814	23.249	0.001
20-minute ahead prediction	4.360	4.294	36.966	0.006
30-minute ahead prediction	5.511	5.381	40.448	0.022
40-minute ahead prediction	6.419	6.163	48.481	0.014
50-minute ahead prediction	7.085	6.838	50.298	0.005
60-minute ahead prediction	7.712	7.361	50.399	0.001

TABLE XVI
RELATIVE ERROR STATISTICS OVER DIFFERENT 10-min TIME INTERVALS

Relative Error (%)	Mean	Standard Deviation	Maximum	Minimum
10-minute ahead prediction	8.035	8.217	74.332	0.004
20-minute ahead prediction	12.719	14.069	105.534	0.011
30-minute ahead prediction	16.561	19.342	167.549	0.027
40-minute ahead prediction	19.419	22.936	188.542	0.014
50-minute ahead prediction	21.671	26.069	187.518	0.006
60-minute ahead prediction	23.791	29.596	248.166	0.001

VI. CONCLUSION

In this paper, four time series models for different prediction horizons were built by data mining algorithms. Two 10-min time series models were constructed, one for the wind speed and the other for the wind farm power. The third model was developed to predict wind farm power on hourly intervals. In the fourth (integrated) model, the wind speed predicted by the 10-min time series model was used as an input to the kNN model to predict the wind farm power. A comprehensive comparative analysis of the four models was reported in the paper.

The multiperiod ahead predictions produced by the first three time series models were accurate. The integrated model for power prediction turned out to be less accurate and stable than the time series models.

The time series prediction models accurately predict the wind speed and more importantly the wind farm power at different time scales. The models are applicable to the wind farm and electricity market management and predictive control of individual turbines, both leading to wind power generation optimization.

The ultimate goal of the research presented in this paper was to further improve the accuracy of power prediction of wind farms. This is a fundamental issue in the wind industry. One avenue to be pursued in future research is the transformation of time series data, e.g., using wavelets or Kalman filters. It is likely that other existing or expanded data mining algorithms will further enhance the power prediction performance. The short-term time series prediction model may become a basis for predictive control aimed at optimizing the wind turbine control settings to maximize the power captured from the wind.

One disadvantage of the proposed approach is that the time series model uses its own previously predicted values. As the number of prediction steps increases, the errors get accumulated. A possible approach for improving prediction accuracy is to build a set of prediction models for each time step. These models would not need their own predicted values as inputs. For example, one prediction model could be built for making 10-min ahead predictions; a second model for making 20-min ahead predictions using the current and historical data.

REFERENCES

[1] A. Kusiak and Z. Song, "Combustion efficiency optimization and virtual testing: A data-mining approach," *IEEE Trans. Ind. Inf.*, vol. 2, no. 3, pp. 176–184, Aug. 2006.

[2] D. A. Spera, *Wind Turbine Technology: Fundamental Concepts of Wind Turbine Engineering*. New York: ASME, 1994, 638 pp.

[3] A. Sfetsos, "A novel approach for the forecasting of the mean hourly wind speed time series," *Renew. Energy*, vol. 27, pp. 163–174, 2002.

[4] A. J. Smola and B. Schoelkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, pp. 199–222, 2004.

[5] B. G. Brown, R. W. Katz, and A. H. Murphy, "Time series models to simulate and forecast wind speed and wind power," *J. Appl. Meteorol.*, vol. 23, pp. 1184–1195, 1984.

[6] C. W. Potter and M. Negnevitsky, "Very short-term wind forecasting for Tasmanian power generation," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 965–972, May 2006.

[7] E. Frank, Y. Wang, S. Ingis, G. Holmes, and I. H. Witten, "Using model trees for classification," *Mach. Learn.*, vol. 32, pp. 63–76, 1998.

[8] I. G. Damousis, M. C. Alexiadis, J. B. Therocharis, and P. S. Dokopoulos, "A fuzzy model for wind speed prediction and power generation in wind parks using spatial correlation," *IEEE Trans. Energy Convers.*, vol. 19, no. 2, pp. 352–361, Jun. 2004.

[9] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005, 525 pp.

[10] J. A. Harding, M. Shahbaz, S. Srinivas, and A. Kusiak, "Data mining in manufacturing: A review," *ASME Trans.: J. Manuf. Sci. Eng.*, vol. 128, pp. 969–976, 2006.

[11] G. Box and G. M. Jenkins, *Time Series Analysis Forecasting and Control*. San Francisco, CA: Holden-Day, 1976.

[12] J. Espinosa, J. Vandewalle, and V. Wertz, *Fuzzy Logic, Identification and Predictive Control*. London, U.K.: Springer-Verlag, 2005.

[13] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, pp. 367–378, 2002.

[14] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, pp. 1189–1232, 2001.

[15] M. J. A. Berry and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 2nd ed. New York: Wiley, 2004.

[16] P. Backus, M. Janakiram, S. Mowzoon, G. C. Runger, and A. Bhargava, "Factory cycle-time prediction with data-mining approach," *IEEE Trans. Semicond. Manuf.*, vol. 19, no. 2, pp. 252–258, May 2006.

[17] P. Seidel, A. Seidel, and O. Herbarth, "Multilayer perceptron tumor diagnosis based on chromatography analysis of urinary nucleoside," *Neural Netw.*, vol. 20, pp. 646–651, 2007.

[18] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to the SMO algorithm for SVM regression," *IEEE Trans. Neural Netw.*, vol. 11, no. 5, pp. 1188–1193, Sep. 2000.

[19] S. Li, D. C. Wunsch, E. O'Hair, and M. G. Giesselmann, "Comparative analysis of regression and artificial neural network models for wind turbine power curve estimation," *ASME Trans.: J. Solar Energy Eng.*, vol. 123, pp. 327–332, 2001.

[20] T. G. Barbounis, J. B. Theocharis, M. C. Alexiadis, and P. S. Dokopoulos, "Long-term wind speed and power forecasting using local recurrent neural network models," *IEEE Trans. Energy Convers.*, vol. 21, no. 1, pp. 273–284, Mar. 2006.

[21] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, pp. 139–157, 2000.

[22] T. Hothorn and B. Lausen, "Bundling classifiers by bagging trees," *Comput. Statist. Data Anal.*, vol. 49, pp. 1068–1078, 2005.

[23] J. L. Torres, A. Garcia, M. D. Blas, and A. D. Francisco, "Forecast of hourly average wind speed with ARMA models in Spain," *Solar Energy*, vol. 79, pp. 65–77, 2005.

[24] [Online]. Available: http://en.wikipedia.org/wiki/Time_series, Accessed Dec. 1, 2008.

[25] A. Kusiak, H. Y. Zheng, and Z. Song, "Models for monitoring and predicting wind farm power," *Renew. Energy*, vol. 34, no. 3, pp. 583–590, 2009.

[26] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, pp. 273–324, 1997.

Authorized licensed use limited to: The University of Iowa. Downloaded on February 19, 2009 at 20:53 from IEEE Xplore. Restrictions apply.

[27] Y. Wang and I. H. Witten, "Induction of model trees for predicting continuous classes," presented at the 9th Eur. Conf. Mach. Learn., Univ. Econ., Faculty Informat. Statist., Prague, Czech Republic, 1997, Poster Paper.
[28] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth, 1984.
[29] M. Martin, L. M. Cremades, and J. M. Santababara, "Analysis and modeling of time series of surface wind speed and direction," *Int. J. Climatol.*, vol. 19, pp. 197–209, 1999.

Andrew Kusiak (M'90) received the B.S. and M.S. degrees in engineering from the Warsaw University of Technology, Warsaw, Poland, in 1972 and 1974, respectively, and the Ph.D. degree in operations research from the Polish Academy of Sciences, Warsaw, in 1979.

He is currently a Professor at the Intelligent Systems Laboratory, Department of Mechanical and Industrial Engineering, The University of Iowa, Iowa City. He speaks frequently at international meetings, conducts professional seminars, and does consultation for industrial corporations. He has served on the editorial boards of over 30 journals. He is the author or couathor of numerous books and technical papers in journals sponsored by professional societies, such as the Association for the Advancement of Artificial Intelligence, the American Society of Mechanical Engineers, etc. His current research interests include applications of computational intelligence in automation, wind and combustion energy, manufacturing, product development, and healthcare.

Prof. Kusiak is the Institute of Industrial Engineers Fellow and the Editor-in-Chief of the *Journal of Intelligent Manufacturing*.

Haiyang Zheng received the B.S. degree from Beihang University, Beijing, China, in 2005. He is currently a Graduate Student in industrial engineering at The University of Iowa, Iowa City, where he is also a Member of the Intellligent Systems Laboratory.

He was with the ultrahigh molecular weight polyethylene fiber industry for two years. His current research interests include data mining, computational intelligence, and process optimization applied in wind power, high-voltage alternating current, and manufacturing industry.

Zhe Song (S'08) received the B.S. and M.S. degrees from the China University of Petroleum, Dong Ying City, China, in 2001 and 2004, respectively, and the Ph.D. degree from The University of Iowa, Iowa City, in 2008.

He is currently a Postdoctoral Researcher at the Intelligent Systems Laboratory, Department of Mechanical and Industrial Engineering, The University of Iowa. He is the author or coauthor of numerous technical papers in journals sponsored by IEEE and the International Federation of Production Research. His current research interests include modeling energy systems, control and optimization, data mining, computational intelligence, decision theory, control theory, and statistics.

Authorized licensed use limited to: The University of Iowa. Downloaded on February 19, 2009 at 20:53 from IEEE Xplore. Restrictions apply.