Journal of Forecasting
J. Forecast. **25**, 537–559 (2006)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/for.1006

A Semi-Parametric Time Series Approach in Modeling Hourly Electricity Loads

JUN M. LIU,[1] RONG CHEN,*[2,3] LON-MU LIU[2] AND JOHN L. HARRIS[4]

[1] *Georgia Southern University, Statesboro, Georgia, USA*
[2] *University of Illinois at Chicago, Chicago, Illinois, USA*
[3] *Peking University, Beijing, China*
[4] *Progress Energy, Inc., Raleigh, North Carolina, USA*

ABSTRACT

In this paper we develop a semi-parametric approach to model nonlinear relationships in serially correlated data. To illustrate the usefulness of this approach, we apply it to a set of hourly electricity load data. This approach takes into consideration the effect of temperature combined with those of time-of-day and type-of-day via nonparametric estimation. In addition, an ARIMA model is used to model the serial correlation in the data. An iterative backfitting algorithm is used to estimate the model. Post-sample forecasting performance is evaluated and comparative results are presented. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS Electricity load forecasting; nonparametric regression; backfitting; time series; ARIMA model

INTRODUCTION

Short-term forecasting of electricity loads plays an important role in the day-to-day planning and operation of electricity power systems. The major operational activities such as plant scheduling, load dispatching, security assessment and reserve capacity allocation rely heavily on short-term forecasts. It is widely recognized that even a marginal improvement in the short-term forecast can result in significant gain in the reliability and profitability of power company operations.

It has been well recognized that a major portion of the total variation in electricity load time series can be attributed to the strong periodic behavior in the data. Meteorological factors, such as temperature, humidity, and wind speed, are important sources of variation in electricity load. Among these meteorological variables, temperature is found to be the most important in many studies (Al-Zayer and Al-Ibrahim, 1996). It is also observed that the temperature–load relationship is highly nonlinear. Engle *et al*. (1986) investigated this relationship and found it can be approximated by an asymmetric V-shaped function with a minimum at around 65°F. This minimum represents the transition point between the needs for heating and cooling. Mendenhall and Sincich (1996) think this

* Correspondence to: Rong Chen, University of Illinois at Chicago, 601 South Morgan Street (M/C 294), Chicago, IL 60607, USA. E-mail: rongchen@uic.edu

relationship may be better approximated by a U-shaped function because usually there is a temperature range in which neither heating nor cooling is needed. Consequently, this approach requires two transition points to be identified. One common method in practice is to transform temperature into degree-days or degree-hours based on the transition points and use piecewise regression to model the temperature effect (Gupta, 1985; Al-Zayer and Al-Ibrahim, 1996; Mendenhall and Sincich, 1996). It is also found that the hourly temperature–load relationship may be affected by factors such as time of day (Peirson and Henley, 1994), day of the week, season, location, income, price, holidays, and so on. Some researchers model the temperature–load relationship individually for different intraday period and day type (Ramanathan *et al*., 1997). Because of the nonlinear nature of the data, the identification of appropriate functional forms for different hour and day type is a complicated task. Instead of trying to control all potential factors, one can adopt the more flexible, data-driven nonparametric regression methods such as an artificial neural network (ANN) (Ho *et al*., 1992; Peng *et al*., 1992), smoothing splines (Engle *et al*., 1986; Harvey and Koopman, 1993), or other methods based on more general basis functions (Smith, 2000). In many of the studies, the strong periodic patterns in electricity load are modeled by decomposing the load into periodic components and including corresponding periodic terms in the models. The serial correlation in the residuals is modeled by relatively simple models (Engle *et al*., 1986; Smith, 2000). As an alternative, autoregressive integrated moving average (ARIMA) models can be used to model the periodic patterns and the serial correlation more adaptively.

In this paper, we combine the flexibility of nonparametric regression methods with the adaptive nature of ARIMA models and consider an additive semi-parametric regression model: we use a nonparametric regression component to model the nonlinear relationship and an ARIMA component to model the serial correlation in the noise. This model can be considered as the semi-parametric counterpart of the linear transfer function model (Box *et al*., 1994), with the transfer function modeled nonparametrically. As a result, this model is more flexible and can be used to model highly nonlinear relationships of unknown functional form. By modeling the noise as an ARIMA model, the serial correlation is removed and the transfer function can be estimated more efficiently. Additionally, the information about the correlation structure obtained in estimating the ARIMA parameters can be used to improve the forecasting performance. We apply this modeling methodology on a real electricity load dataset. This dataset contains the hourly electricity load data from 1998 to 2000 of an electric utility, whose service area is located in the eastern United States. The electricity load under consideration is the total system load including industrial, residential and commercial usage. In this study, for each hour of a certain type of the day (workday or non-workday), the temperature effect is modeled individually by local polynomial regression, thus allowing the temperature effect to vary according to time of day and type of day. In this respect, the proposed model is similar to the EGRV model developed by Engle, Granger, Ramanathan and Vahid-Arraghi (Electric Power Research Institute, 1993). The EGRV model is a multi-equation regression model, and its usefulness was demonstrated by Ramanathan *et al*. (1997) in a comparative study. But because the proposed model uses nonparametric regression methods to model the nonlinear temperature–load relationship, it can approximate the relationship more appropriately. Additionally, because of the *let-the-data-speak-for-themselves* property of nonparametric regression, the difficulty of identifying nonlinear parametric models for the 48 potentially different temperature–load relationships for different hour/day types can be avoided. Cottet and Smith (2003) also used flexible functional forms combined with Bayesian model averaging to model the temperature–load relationship. Their approach was based on a few basis functions, while the local polynomial method used in this paper is more general and is likely to produce more accurate estimates. In this study, the periodic load component and the serial correlation in the data are modeled by a multiplicative ARIMA model. We use a modified backfitting algorithm (Hastie and Tibshitrani, 1991) for model estimation. We use the data of the first two years (1998–1999) to build the model, and reserve the data of the third year to evaluate the forecasting performance of the model.

This paper is organized as follows. The next section describes the data in more detail and provides a preliminary local polynomial fitting. The motivation of the proposed two-component model is also given in this section. In the third section we introduce a semi-parametric two-component model motivated by the preliminary study. We apply this algorithm to the load data and the results are presented in the fourth section. In this section we also compare the within- and post-sample performances of the model with that of the EGRV model used by Ramanathan *et al*. (1997). In the fifth section we extend the nonparametric fitting to a two-dimensional model and fit load as a function of both temperature and time of day. A summary is provided in the final section.

PRELIMINARY ANALYSIS

Data under consideration

Figure 1 shows the time series plot of the hourly electricity load from 1998 to 2000. We observe the strong seasonality in the electricity usage: the loads are high in winter and summer, and low in spring and fall.

In the service area under study, electricity is a main source of energy for heating in the winter and cooling in the summer. Temperature is recorded hourly at four weather stations in the service area. The temperature used in this study is a weighted average of these four hourly temperatures, and the weights reflect the electricity usages in the areas covered by the four weather stations. Figure 2 is the time series plot of the hourly average temperature.

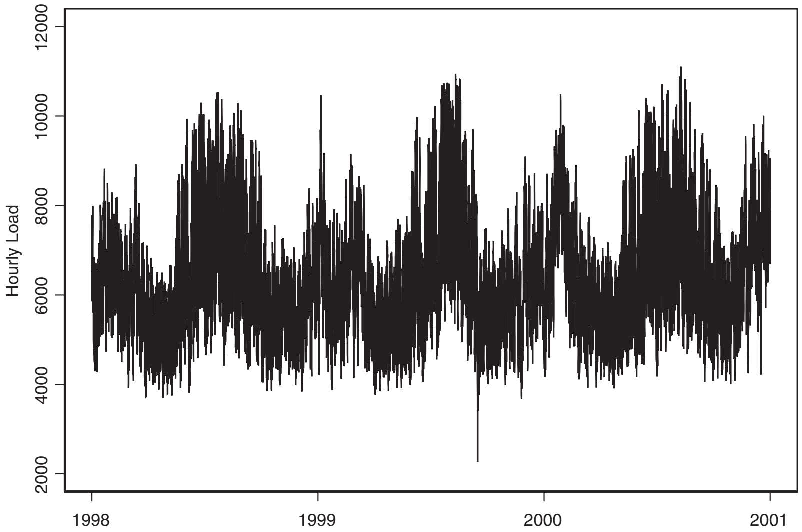

Figure 1. Time series plot of hourly load (1998–2000)

DOI: 10.1002/for

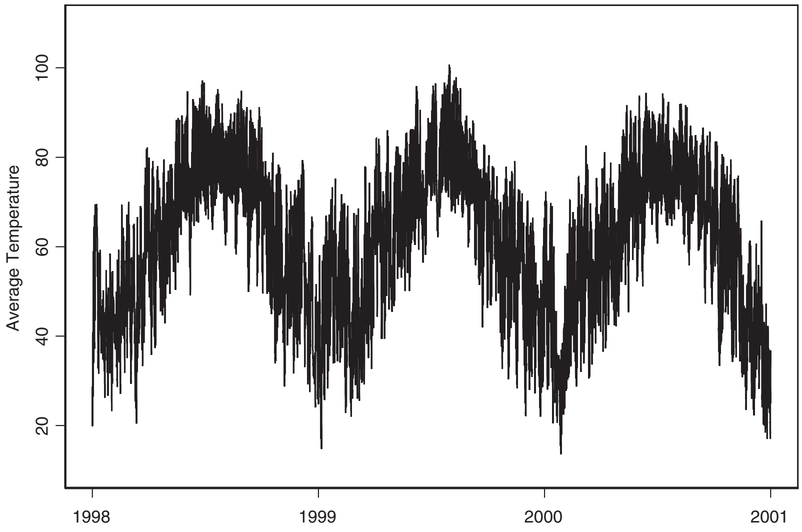

Figure 2. Time series plot of average temperature

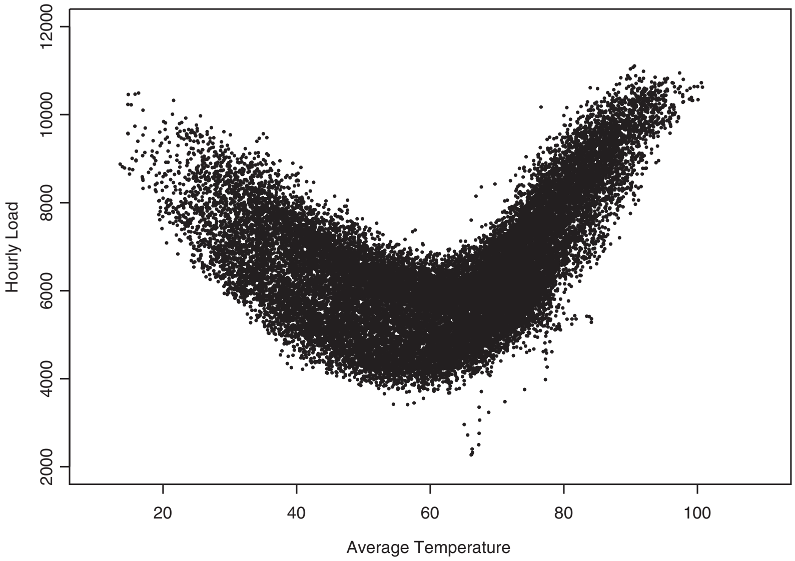

Figure 3. Load versus temperature

Figure 3 is the scatter plot of hourly load against average temperature. Even though the temperature–load relationship revealed in this plot can be reasonably approximated by either a U-shaped or a V-shaped function, the relationship is complex.

In addition to the yearly seasonal pattern shown in Figure 1, it is well known that hourly load data display daily and weekly periodic pattern (Harvey and Koopman, 1993; Smith, 2000). These periodic patterns can be modeled by seasonal ARIMA models. However, a pure seasonal ARIMA model

DOI: 10.1002/for

without incorporating the temperature effect would not be adequate. We shall start our analysis by studying the relationship between load and temperature.

From Figures 1 and 3, we see an unusual drop in electricity load around September 1999. A closer investigation reveals that this drop was the result of hurricane activity. To alleviate the potential impact of these extreme observations, we replaced the data with those of the following day. All subsequent analysis reported in this paper are based on this adjusted dataset. We also performed the same analysis based on the unadjusted data; the results obtained are similar to those based on the adjusted data. This is possibly due to that fact that the dataset is large and therefore the effect of a few adjusted data points is negligible.

Load and temperature

As mentioned earlier, the temperature–load relationship is highly nonlinear and is likely to be different according to many factors such as time of day and type of day. It is generally difficult to identify an appropriate functional form that can account for all these factors. As a result, we adopt the principle of 'let-the-data-speak-for-themselves' and use data-driven nonparametric smoothing methods to model the temperature effect. Specifically we use the LOcally WEighted Scatterplot Smoothing (LOWESS) algorithm by Cleveland (1979). The smoothing parameter λ is selected using the cross-validation (CV) algorithm (Wahba and Wold, 1975):

$$\lambda = \arg_{\lambda} \min\left\{\frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{f}_{\lambda,[-i]}(X_i)]^2\right\} \tag{1}$$

where (X_i, Y_i) is the ith observation, and $\hat{f}_{\lambda,[-i]}$ is the estimated function omitting the ith observation using bandwidth λ. As a preliminary analysis, we fit the electricity load against temperature for the years 1998–1999 using S-Plus function 'lowess'. In this case, the optimal CV bandwidth is found to be 0.10. The fitted curve is shown in the right panel of Figure 4.

The resulting root mean squared error (RMSE) of this model is 775.91. We also performed a post-sample forecast using the data of year 2000 with this model. In this paper, actual temperature data are used in all the forecasts. The RMSE for the post-sample forecasts is 893.18. (For notational simplicity, in what follows we will refer to the RMSE of the estimated model using data of years 1998–1999 as *within-sample RMSE* and the RMSE of the post-sample forecasts using data of year 2000 as *post-sample RMSE.*) As mentioned earlier, it is widely recognized in the load-forecasting literature that electricity consumption is usually different for different days and times, even under the same temperature. For example, under the same temperature, the hourly electricity usage at 8:00 am could be very different from that of 8:00 pm. To see this, we plot load against average temperature for 8 am and 8 pm in Figure 5. We observe that the data in each subset have less variation than in Figure 3. Similar results are observed in other hours. This partially explains the large RMSEs obtained in the above analysis. Better results could be obtained by taking the hour effect into consideration and modeling the temperature–load relationship individually for different hours. We also found the electricity usage pattern of workdays is different from that of non-workdays. This is reflected in the two-cluster pattern shown in the left display of Figure 5. This pattern is more pronounced in daytime hours than in night-time hours. Our study shows that non-workdays (here including Saturdays, Sundays, holidays and the day before most of the holidays) behave similarly, and we treat them the same in modeling. Hence we classify the days into two groups: workday and non-workday. The holidays we considered in this study include New Year's Day, Good Friday, Memorial Day, Independence Day, Labor Day, Thanksgiving and Christmas.

DOI: 10.1002/for

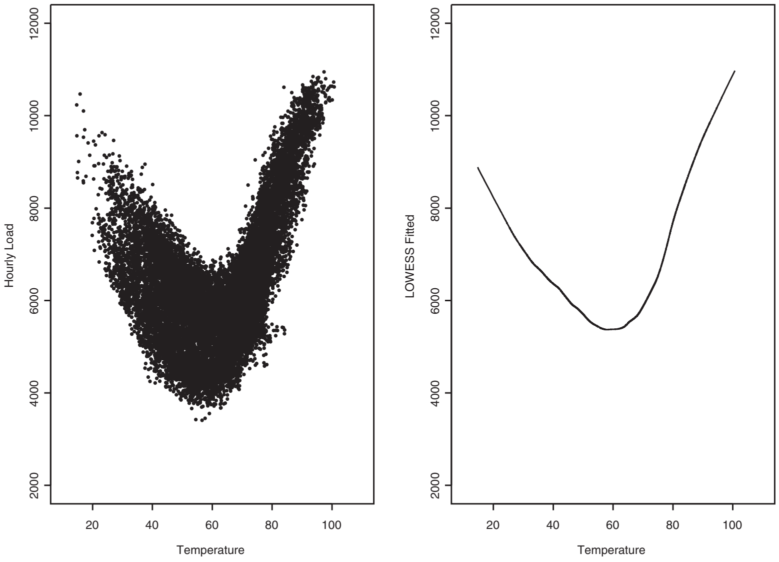

Figure 4. Load versus temperature (left) and estimated response function (right)

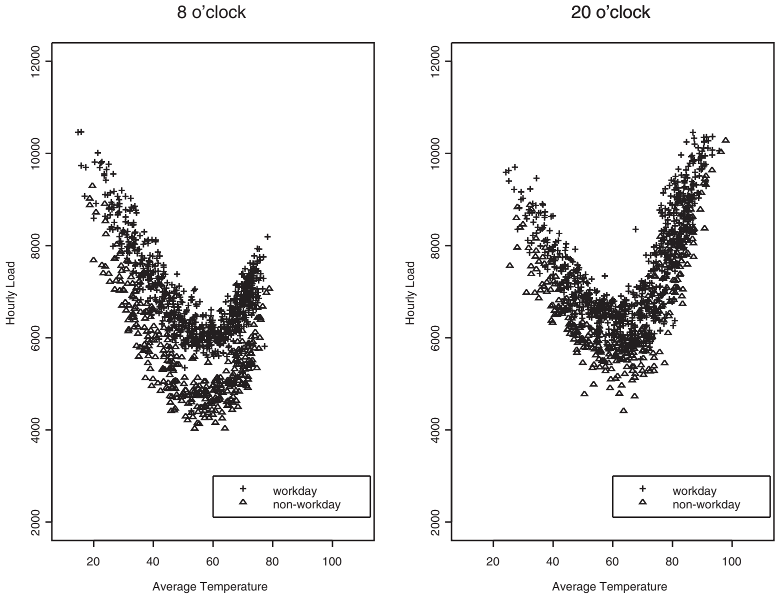

Figure 5. Workday/non-workday load against temperature for hours 8 and 20

Table I. Bandwidths for different hours of workdays and non-workdays

Workday	Hour	1	2	3	4	5	6	7	8	9	10	11	12
	λ	0.19	0.13	0.19	0.22	0.22	0.25	0.25	0.28	0.25	0.22	0.16	0.22
	Hour	13	14	15	16	17	18	19	20	21	22	23	24
	λ	0.28	0.31	0.28	0.25	0.28	0.28	0.13	0.25	0.19	0.22	0.19	0.19
Non-workday	Hour	1	2	3	4	5	6	7	8	9	10	11	12
	λ	0.13	0.22	0.31	0.22	0.22	0.25	0.22	0.25	0.31	0.34	0.31	0.34
	Hour	13	14	15	16	17	18	19	20	21	22	23	24
	λ	0.25	0.25	0.28	0.28	0.31	0.31	0.34	0.16	0.28	0.28	0.25	0.25

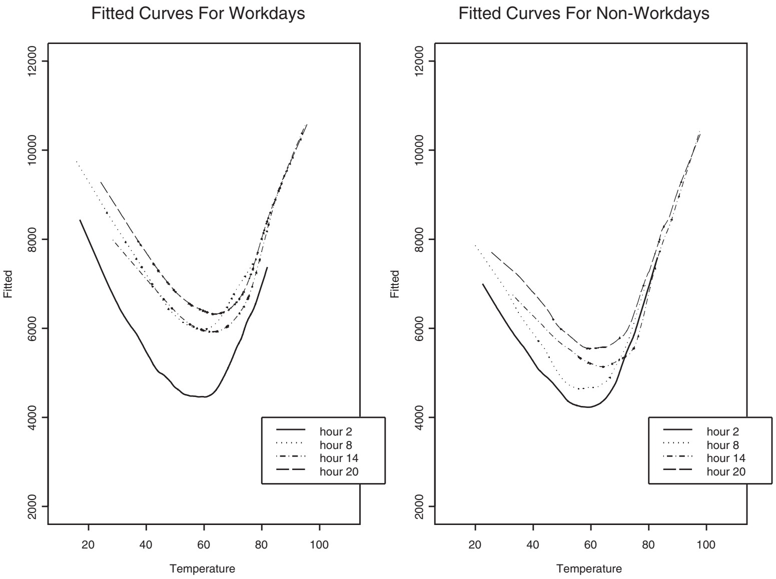

Figure 6. The estimated response functions of selected hours

Based on the above observations, we split the data by time of day and type of day. Then we fit each subset with a smooth function with LOWESS. The resulting overall within-sample RMSE is 369.59, and the post-sample RMSE is 539.92, which shows a significant improvement over the single-curve model. The smoothing parameters selected by CV for each LOWESS fitting are given in Table I.

The fitted curves of hours 2, 8, 14, 20 for workdays and non-workdays are shown in Figure 6. From these curves, we see different patterns among workday/non-workday and different hours. For example, the load usage levels of hours 8, 14, 20 are higher than that of hour 2 for workdays. Also, the load levels of hours 8, 14, 20 of workdays are higher than those of the same hours for non-workdays. Night-time hours (e.g., hour 2) are similar between workdays and non-workdays. This verifies our previous observation.

DOI: 10.1002/for

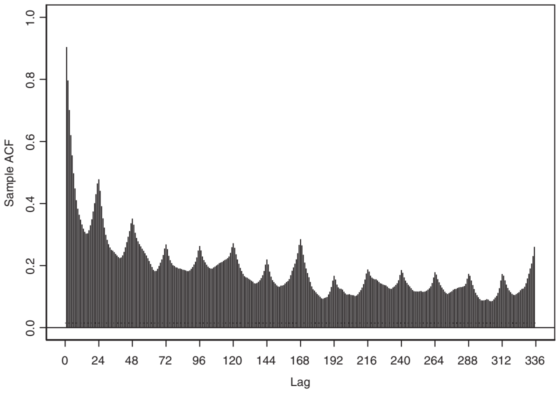

Figure 7. The ACF of LOWESS residuals

Serial correlation in the residuals

As all time series data, in this study serial correlation is an important issue. Figure 7 shows the autocorrelation function (ACF) plot of the residual series after removing the nonparametric mean curve estimated above. It is clear that autocorrelations are still prominent and show a strong periodicity of 24. To deal with this issue, we employ an ARIMA model to fit the residuals series. By examining the sample ACF and partial ACF (Box *et al.*, 1994), we identified a multiplicative ARIMA (1, 0, 0) × $(1, 1, 1)_{24}$ model, which can be written as

$$(1-\phi_1 B)(1-\phi_2 B^{24})(1-B^{24})\varepsilon_t = (1-\theta_1 B^{24})a_t$$

where ε_t is the residual from the nonparametric model, a_t is assumed to be i.i.d. $N(0, \sigma_a^2)$ and B is the backshift operator, $B^i\varepsilon_t \equiv \varepsilon_{t-i}$. By introducing this time series model, the within-sample RMSE decreases from 369.59 to 149.40, and the post-sample RMSE is reduced from 539.92 to 156.76.

A TWO-COMPONENT SEMI-PARAMETRIC MODEL

The model

Based on the above analysis, we consider the following two-component model:

$$Y_{ij} = f_i(X_{ij}) + \varepsilon_{ij} \tag{2}$$

$$(1-\phi_1 B)(1-\phi_2 B^{24})(1-B^{24})\varepsilon_t = (1-\theta_1 B^{24})a_t \tag{3}$$

Here $i = 1, \ldots, 48$ is the subscript denoting different time of day and type of day, Y_{ij} is the jth load in sub-series i, X_{ij} is the corresponding temperature and a_t follows i.i.d. $N(0, \sigma_a^2)$. Note that we use the time subscript t in equation (3) because we reassemble the time series in its original order for the ARIMA model. This is a semi-parametric model with two components. The nonparametric component (2) deals with the nonlinear temperature–load relationship, and the time series component (3) deals with the serial correlation in the data.

DOI: 10.1002/for

Estimation

The two-step estimation in the previous section provides a consistent estimator of the model. Note that, under a strong mixing condition for the residual series ε_t, local polynomial estimator is asymptotically consistent in estimating the mean function $f_i(\cdot)$ (Masry and Fan, 1997); therefore we expect the LOWESS estimator to perform similarly. As a consequence, the estimated residuals $\hat{\varepsilon}_t$ are also consistent to the residual series ε_t. The ARIMA model obtained using the estimated $\hat{\varepsilon}_t$ yields consistent estimates of the parameters, though the convergence rate will need a more careful study.

It is also important to note that when applying nonparametric smoothing methods to time series data special care must be taken because of the serial correlation. In this study we use the leave-one-out cross-validation method to select the bandwidth. Under the usual independence assumption, CV produces the optimal bandwidth. Hart and Vieu (1990) showed that the leave-one-out CV method remains optimal asymptotically under a strong mixing condition. However, in finite samples, the CV bandwidth obtained under the independent assumption can be misleading when serial correlation is present. In addition, the estimation will be more efficient when the serial correlation is taken into consideration. Hence, we propose to estimate both components simultaneously.

The model we considered (equations 2 and 3) has two components: one is a nonparametric model and the other is an ARIMA model. Our algorithm uses the backfitting idea (Hastie and Tibshitrani, 1991) and iteratively estimates the two components. Equation (3) can be written as

$$\varepsilon_t = \eta_t + a_t$$

where η_t is the linear projection of ε_t onto the σ-field generated by $\{\varepsilon_s, s < t\}$. It can be estimated by replacing the parameters and a_t with their estimated values in

$$\eta_t = [\phi_1 B + (1+\phi_2)B^{24} - \phi_1(1+\phi_2)B^{25} - \phi_2 B^{48} + \phi_1\phi_2 B^{49}]\varepsilon_t - \theta_1 B^{24} a_t$$

Then, equation (2) becomes

$$Y_{ij} - \eta_{ij} = f_i(X_{ij}) + a_{ij}$$

where a_{ij} are independent $N(0, \sigma_a^2)$. If η_{ij} is known, then estimating $f_i(\cdot)$ becomes a standard nonparametric regression problem with independent noises. On the other hand, if $f_i(\cdot)$ is given, then model (3) can be estimated using standard ARIMA procedures, with ε_t replaced by $Y_{ij} - f_i(X_{ij})$. Specifically, the proposed algorithm is given as follows:

- Let Y_t be the original data. Set the initial values $\hat{\eta}_t = 0$ and $Z_t = Y_t - \hat{\eta}_t$.
- Do the following until convergence:
  - — Split Z_t into sub-series Z_{ij}.
  - — Fit $Z_{ij} = f_i(X_{ij}) + \varepsilon_{ij}$, obtain $\hat{f}_i(\cdot)$.
  - — Combine $\hat{f}_i(X_{ij})$ into one series $\hat{f}(X_t)$ and let $\hat{\varepsilon}_t = Y_t - \hat{f}(X_t)$.
  - — Fit $(1 - \phi_1 B)(1 - \phi_2 B^{24})(1 - B^{24})\hat{\varepsilon}_t = (1 - \theta_1 B^{24})a_t$; obtain the residuals $\hat{a}_t$.
  - — Obtain the linear projection $\hat{\eta}_t = \hat{\varepsilon}_t - \hat{a}_t$.
  - — Set $Z_t = Y_t - \hat{\eta}_t$.
- After convergence, the fitted value of Y_{ij} is $\hat{Y}_{ij} = \hat{f}_i(X_{ij}) + \hat{\eta}_{ij}$ and the estimate of noise series a_t in (3) is $\hat{a}_t = Y_t - \hat{Y}_t$.

DOI: 10.1002/for

In each step, we used the leave-one-out cross-validation algorithm (1) to select the smoothing parameter λ, for the nonparametric estimation. By using the CV criterion, the nonparametric estimation is essentially based on optimizing the post-sample forecasting performance. However, the ARIMA estimation is based on minimizing the within-sample RMSE. As a consequence, the overall RMSE may not necessarily decrease in each iteration, and the algorithm may not converge to a single point. Our experience shows it usually converges to a limit cycle within a very small range which bears no practical difference.

RESULTS ON THE ELECTRICITY LOAD DATA

The algorithm presented above was run iteratively, and we continuously check the within-sample RMSEs for both the nonparametric component and the ARIMA component, as well as the estimated coefficients of the ARIMA model. Here the RMSEs of the nonparametric component and the ARIMA component are defined as

$$\sqrt{\frac{1}{n}\sum_{i,j}[Z_{ij}-\hat{f}_i(X_{ij})]^2} \text{ and } \sqrt{\frac{1}{n}\sum_t \hat{a}_t^2}$$

respectively. Figure 8 shows the evolution of within-sample RMSEs of the nonparametric and ARIMA steps across iterations. As we can see, the algorithm converges at around iteration 200.

The smoothing parameters are reselected by CV periodically during iterations. After the first several iterations the estimated smoothing parameters $\hat{\lambda}_i$ ($i = 1, \ldots, 48$) are quite stable and similar, as shown in Table II. Because of the similarity between $\hat{\lambda}_i$, we averaged $\hat{\lambda}_i$ at iteration 50 across the 48 curves and used it as the smoothing parameter in all subsequent iterations.

The within-sample RMSE at iteration 700 is 91.98. The fitted nonparametric curves of every 100 iterations up to iteration 700 for hours 8 and 20 are given in Figure 9. It is clear that the fitted curves converge in shape and location as the number of iterations increases. The curves at iteration 1 are very different from those of iteration 100, but when we run more iterations the curves become essentially indistinguishable.

The ACF of the final residuals $\hat{a}_t$ are given in Figure 10. From this figure, we can see that a periodic pattern of lag 168 is still present. This is because we did not take the weekly periodic pattern into consideration in model 3.

To remedy this, we re-identified the time series model using the estimated nonparametric functions at the final iteration and obtained the following model:

$$(1-\phi_1 B-\phi_2 B^2)(1-\phi_3 B^{168})(1-B^{24})\varepsilon_t = (1-\theta_1 B)(1-\theta_2 B^{24})(1-\theta_3 B^{168})a_t \tag{4}$$

This refined model is used in the final iteration. The identification and estimation of this model are carried out using the SCA software package developed by Liu *et al*. (1992).

The resulting within-sample RMSE is found to be 87.21. The estimates of the ARIMA model are shown in Table III.

The ACF of the residuals of this model, which is shown in Figure 11, indicates that the residual series is roughly a white noise process. (Note that in order to have a better presentation of the ACF pattern we set the vertical scale of Figures 10 and 11 to (–0.3, 0.3) because of the small magnitude of the ACF values).

DOI: 10.1002/for

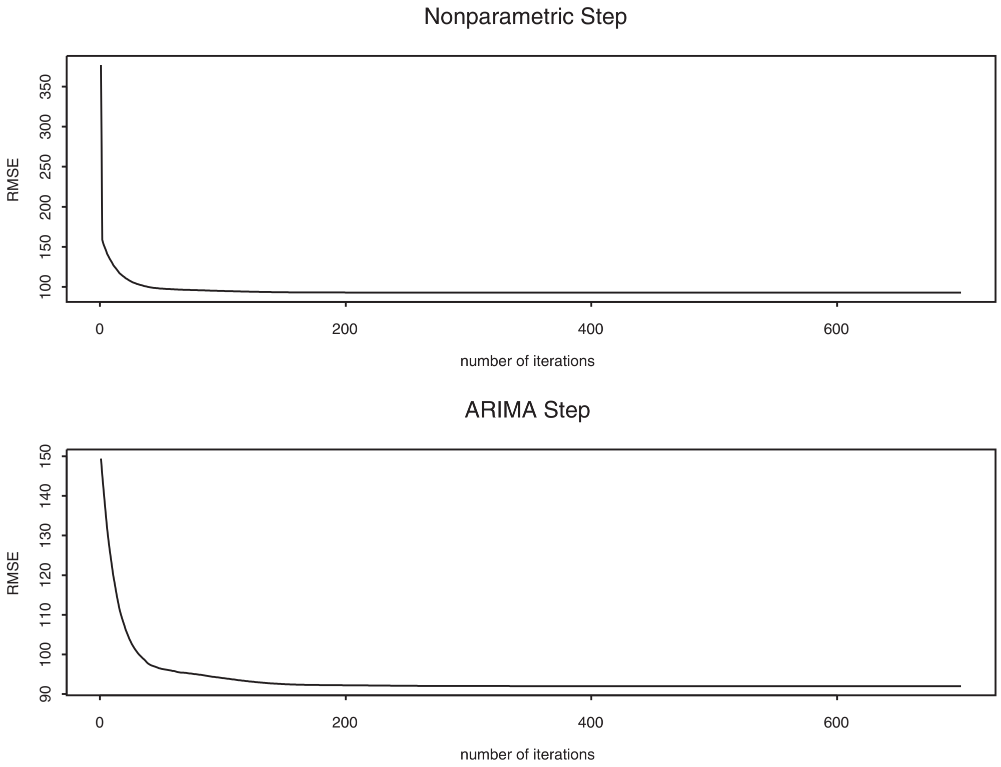

Figure 8. The within-sample RMSE versus number of iterations

Table II. Bandwidths for selected workday hours

Hour	3	6	9	12	15	18	21	24
Iteration 1	0.19	0.25	0.25	0.22	0.28	0.28	0.19	0.19
Iteration 2	0.13	0.10	0.13	0.10	0.16	0.19	0.10	0.16
Iteration 20	0.13	0.10	0.13	0.10	0.16	0.19	0.10	0.16
Iteration 50	0.16	0.13	0.19	0.16	0.16	0.16	0.10	0.16
Iteration 100	0.16	0.13	0.19	0.16	0.16	0.16	0.10	0.16

The fitted nonparametric curves at the final iteration for each subset are given in Figures 12 and 13.

Using the identified two-component model, we perform a post-sample forecast using data of year 2000. As mentioned previously, actual temperature data are used in the forecast. We find that the post-sample RMSE is 93.37. We also calculate the mean absolute percentage error (MAPE), which is defined as $\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|$, where Y_t is the actual observation and $\hat{Y}_t$ is the corresponding forecast. The within- and post-sample MAPEs are found to be 1.0094% and 1.0114%, respectively.

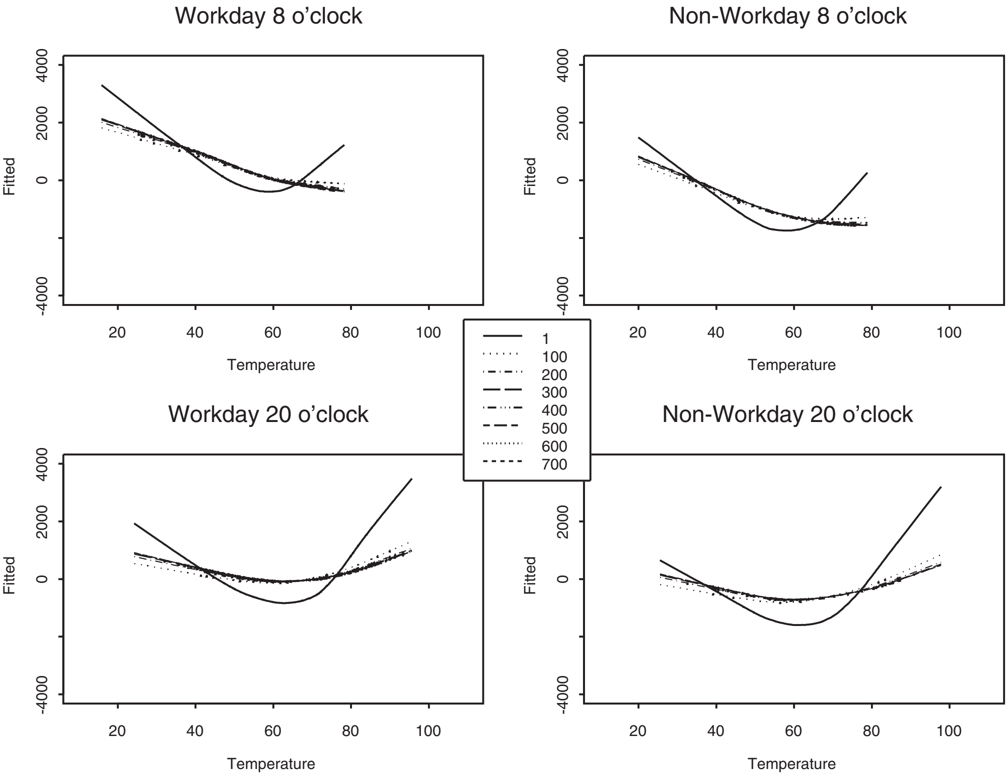

Figure 9. The nonparametric fitted curves for hours 8 and 20 across iterations

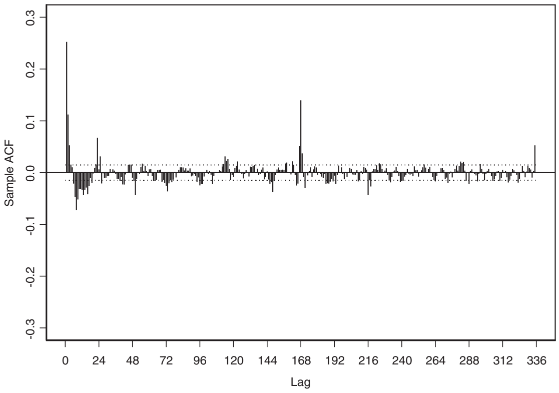

Figure 10. The ACF of $\hat{a}_t$

DOI: 10.1002/for

Table III. The model estimation of equation (4)

Variable	Value	*t*-Value	Type	Order	SE
ϕ_1	1.5340	69.40	AR	1	0.0221
ϕ_2	–0.5504	–25.40	AR	2	0.0217
ϕ_3	0.8880	135.57	AR	168	0.0065
θ_1	0.3250	12.91	MA	1	0.0252
θ_2	0.7762	156.54	MA	24	0.0050
θ_3	0.7753	85.02	MA	168	0.0091

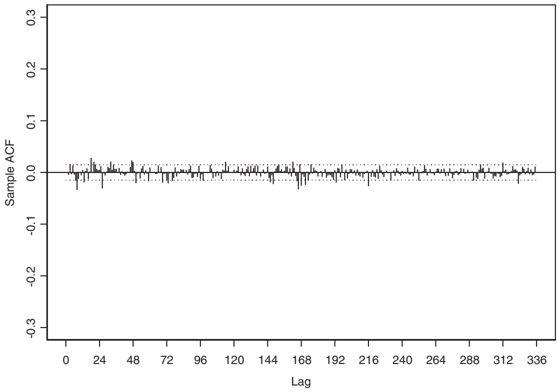

Figure 11. The ACF of $\hat{a}_t$ of the refined model (Model 4)

To visualize the forecasting performance of the proposed model, the post-sample forecasts in one winter week (1/10/00–1/16/00), one summer week (8/7/00–8/13/00) and one holiday period (12/22/00–12/31/00) are plotted in Figures 14–16.

The EGRV model (Electric Power Research Institute, 1993) performs very well in short-term electricity forecasting. It outperformed a wide range of alternative models in a comparative forecasting experiment hosted by the Puget Sound Power and Light Company (Ramanathan *et al.*, 1997). The EGRV model is a multi-equation model. In this model multiple regression functions with a dynamic error structure as well as adaptive adjustments are fit individually for each hour in weekday and weekend. In the proposed model we also treat different hour and day type individually, but we adopt nonparametric regression methods to capture the temperature–load relationship. The EGRV selects variables from 31 candidate variables of four categories: deterministic (e.g., year trend, binary variables for months, Monday, Friday and day after holiday), temperature-related (e.g., temperature and its square, temperature peaks, moving average of temperature), load-related (e.g., current load) and past error. For more details about the EGRV model, please refer to Ramanathan *et al.* (1997). Here we compare the within- and post-sample RMSE and MAPE of the EGRV model with those of the proposed two-component model (which will be called the *semi-parametric transfer function model* (SPTF) in what follows). Actual temperature is used in both forecasts. The results are summarized in Table IV.

As shown in Table IV, even though the EGRV model has better within-sample performance, the SPTF model has better post-sample forecasting performance than the EGRV model. We further

DOI: 10.1002/for

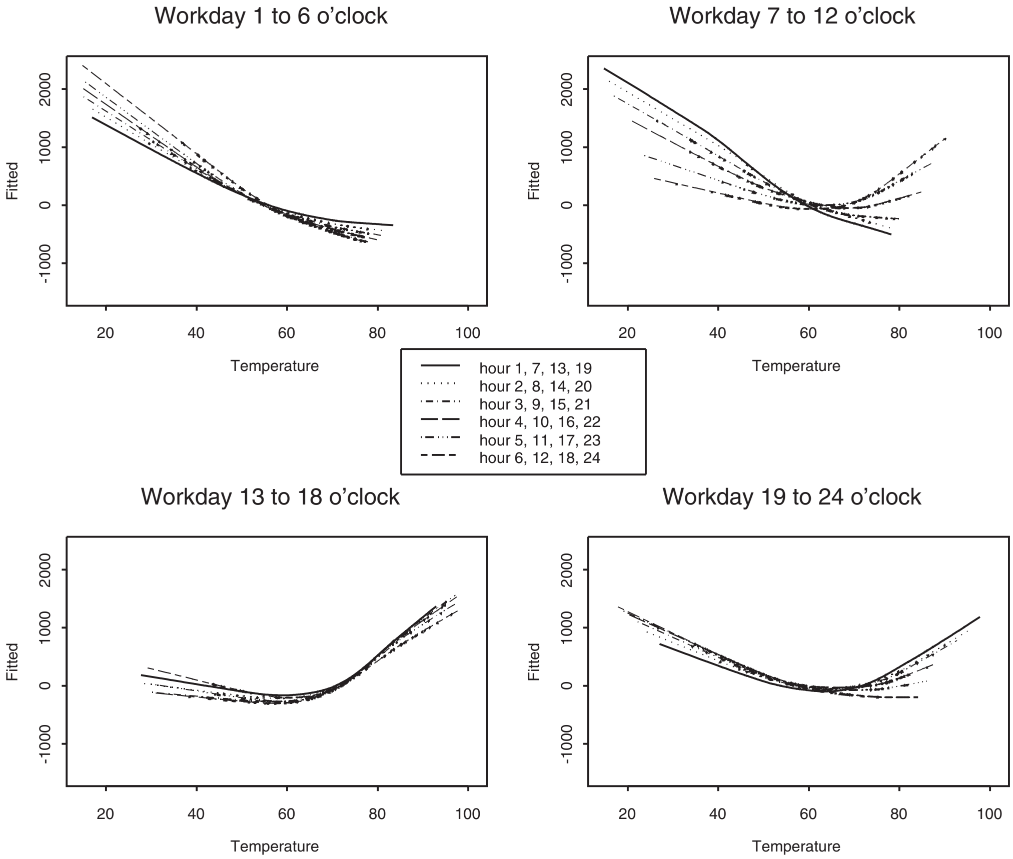

Figure 12. The nonparametric fitted curves of workdays

Table IV. The overall performances of the SPTF and the EGRV models

	Within RMSE	Post RMSE	Within MAPE	Post MAPE
SPTF	87.21	93.37	1.009%	1.011%
EGRV	82.64	108.21	0.918%	1.110%

compare the post-sample forecasting performances of these two models in individual hours and days of the week. The comparison of MAPE is summarized in Table V, in which '1's denote the cases when the SPTF model performs better (i.e., has smaller MAPE) than the EGRV model, '0's otherwise. The column sums denote the number of hours in which the SPTF model outperforms the EGRV model for each day of the week, and the row sums denote the number of days in which the SPTF model outperforms the EGRV model for each hour of the day. Overall, we found that in 50% (i.e., 84 out of 168) of the cases the SPTF model outperforms the EGRV model. However, it is interesting to see that the SPTF model outperforms the EGRV model mostly in the usage-intensive hours (namely, almost all cases in 7 am, 10 am to 2 pm, and 7 pm to 10 pm). The same comparison of RMSEs shows a similar pattern; the details are omitted here.

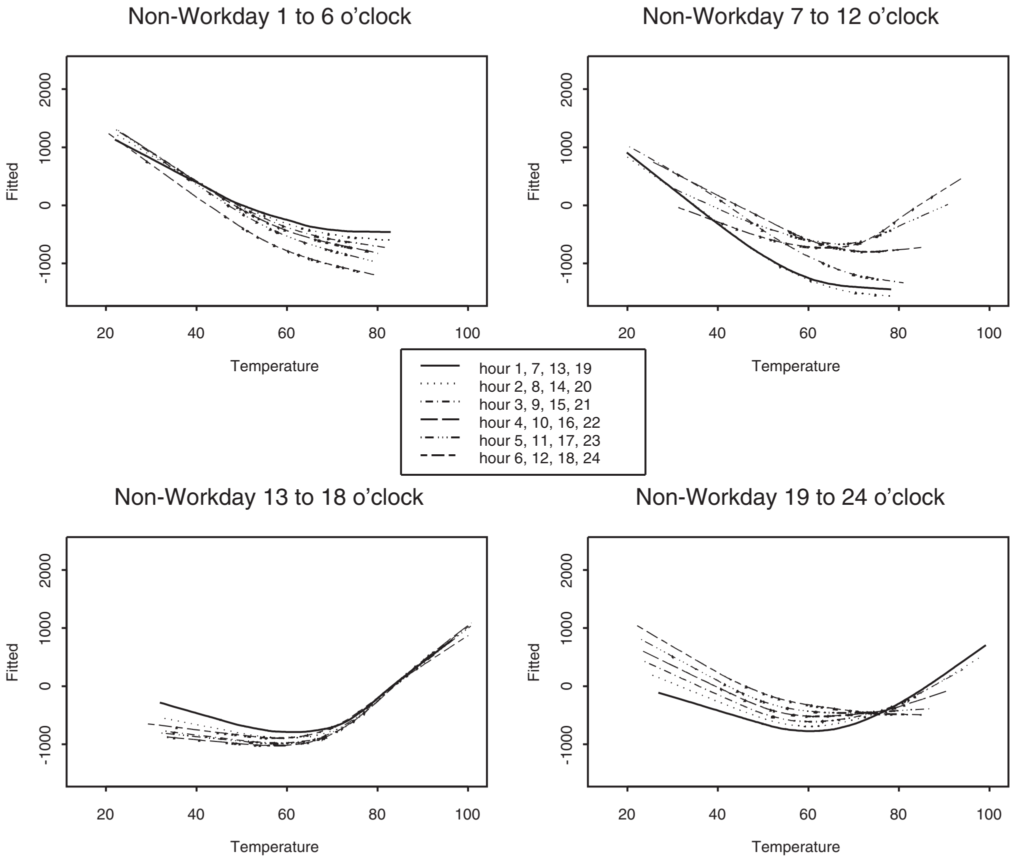

Figure 13. The nonparametric fitted curves of non-workdays

Remark: Note that the previous results were obtained by setting the initial value of $\hat{\eta}_t$ to zero and estimating the nonparametric functions first. One can also start the iteration with the other component; e.g., set $\hat{f}(\cdot) = 0$ and estimate the ARIMA model first. If the iterative procedure has a stable solution, both approaches should provide similar solutions. To verify this, we ran the procedure again for 700 iterations, starting with $\hat{\eta}_t = 0$, and fit the ARIMA model first. Figure 17 shows the within-sample and post-sample RMSE for each iteration. Again, we see that the RMSEs stabilize at approximately iteration 200, and the ARIMA RMSE and nonparametric RMSE stabilize at about the same level. The within-sample RMSE at iteration 700 is 92.74 and MAPE is 1.096%, similar to what we have obtained before. Again, to take care of the weekly effect, we used model (4) in the final iteration. The resulting within-sample RMSE and MAPE are 87.99 and 1.019%, respectively. Based on this refined model, we also performed the post-sample forecast using data of year 2000. The post-sample RMSE and MAPE are 93.99 and 1.104%, respectively. Hence changing the order of fitting gives us roughly the same result, which suggests that the procedure does converge in this case.

The above post-sample forecast is essentially one-step-ahead forecast. Forecasts of such short horizon are important for decisions such as immediate dispatch or system stability. But other

DOI: 10.1002/for

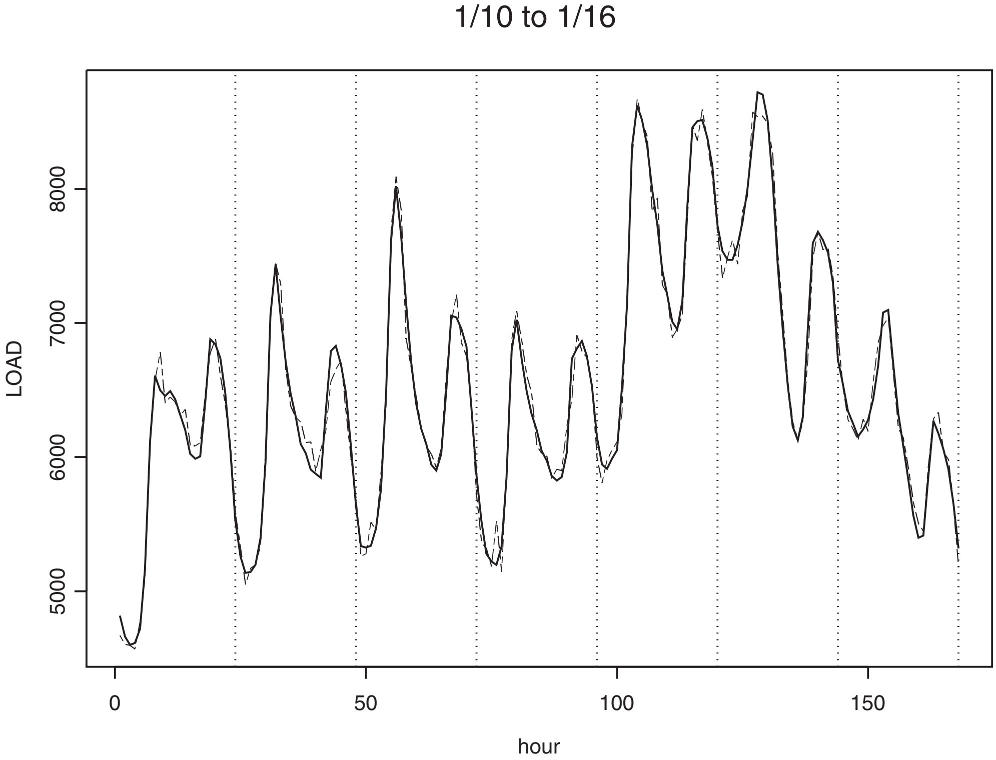

Figure 14. The actual observations (solid line) and the forecasts (dashed line) of Jan. 10 to Jan. 16, 2000

decisions such as plant scheduling and reserve capacity allocation usually require longer-horizon forecasts. The proposed approach can be used for this purpose. To illustrate this, we use the proposed approach to make 24-step (i.e., one-day)-ahead forecasts. Note that because of the nonlinear nature of the problem we need to identify a different model for the different forecasting horizon. We identify the following two-component model:

$$Y_{ij} = f_i(X_{ij-24}) + \varepsilon_{ij} \tag{5}$$

$$(1-B)(1-B^{24})(1-B^{168})\varepsilon_t = \frac{(1-\theta_1 B^{24})(1-\theta_2 B^{168})}{1-\phi_1 B} a_t \tag{6}$$

This model is estimated using the modified backfitting algorithm described above. To save space, the detailed estimation results are not given here. A post-sample 24-step-ahead forecast is performed using the identified model and the data of year 2000. Similarly, as a benchmark, the EGRV model is used to perform a post-sample 24-step-ahead forecast. The within- and post-sample performances of the two models are given in Table VI. We can see that in this case the SPTF model has better overall performance than the EGRV model.

DOI: 10.1002/for

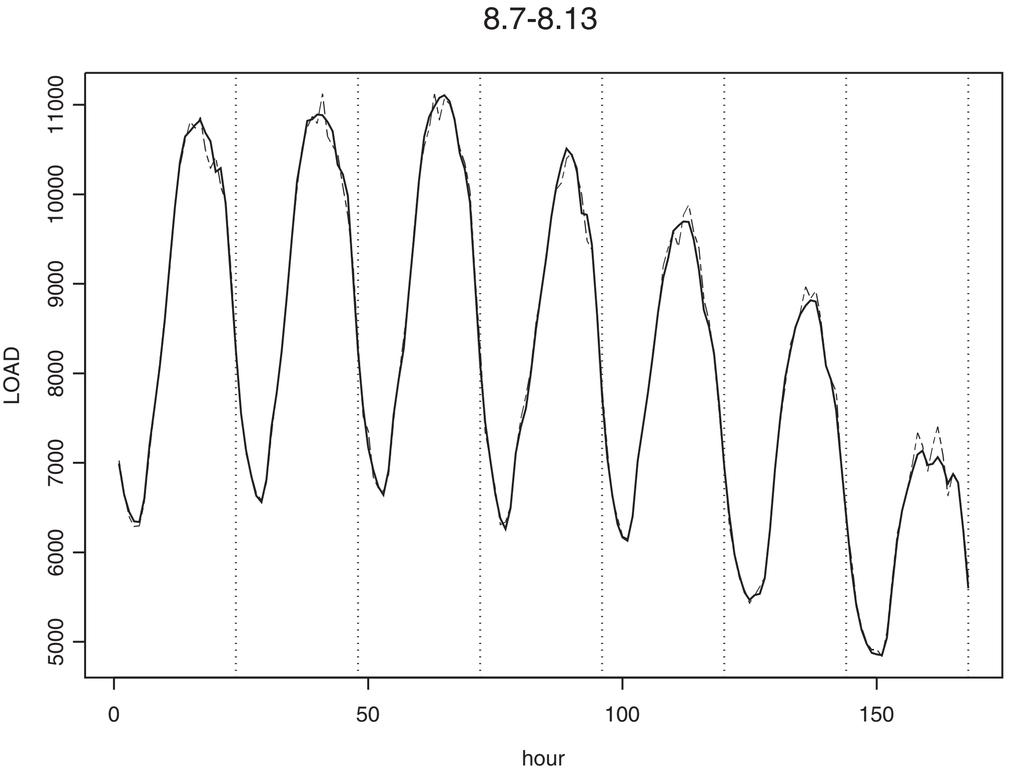

Figure 15. The actual observations (solid line) and the forecasts (dashed line) of Aug. 7 to Aug. 13, 2000

A TWO-DIMENSIONAL EXTENSION

The estimated one-dimensional functions (Figures 12 and 13) are summarized in three-dimensional plots shown in Figures 18 and 19. The changes between neighboring curves are small and smooth. This observation leads us to consider the two-dimensional version of the proposed model:

$$Y_t = f_{w_t}(X_t, T_t) + \varepsilon_t, \tag{7}$$

$$(1-\phi_1 B)(1-\phi_2 B^{24})(1-B^{24})\varepsilon_t = (1-\theta_1 B)(1-\theta_2 B^{24})a_t \tag{8}$$

where $w_t \in \{1, 2\}$ indicates whether observation t belongs to a workday or a non-workday and $T_t = \{t \bmod 24\}$ represents the time of the day for observation t. The function f is a two-dimensional smooth function with T_t being a circular variable (i.e. hour 1 and hour 24 are considered to be neighbors). As a consequence, we consider two two-dimensional functions, one for workdays and one for non-workdays, instead of considering all 48 non-parametric functions in (2). The estimation is carried out with the LOESS procedure (Cleveland and Devlin, 1988), which is the multidimensional version of LOWESS. We used the GCV criteria (Craven and Wahba, 1979) to select the bandwidth. Note that the LOESS bandwidth is based on Euclidean distance and treats all dimensions symmetrically.

DOI: 10.1002/for

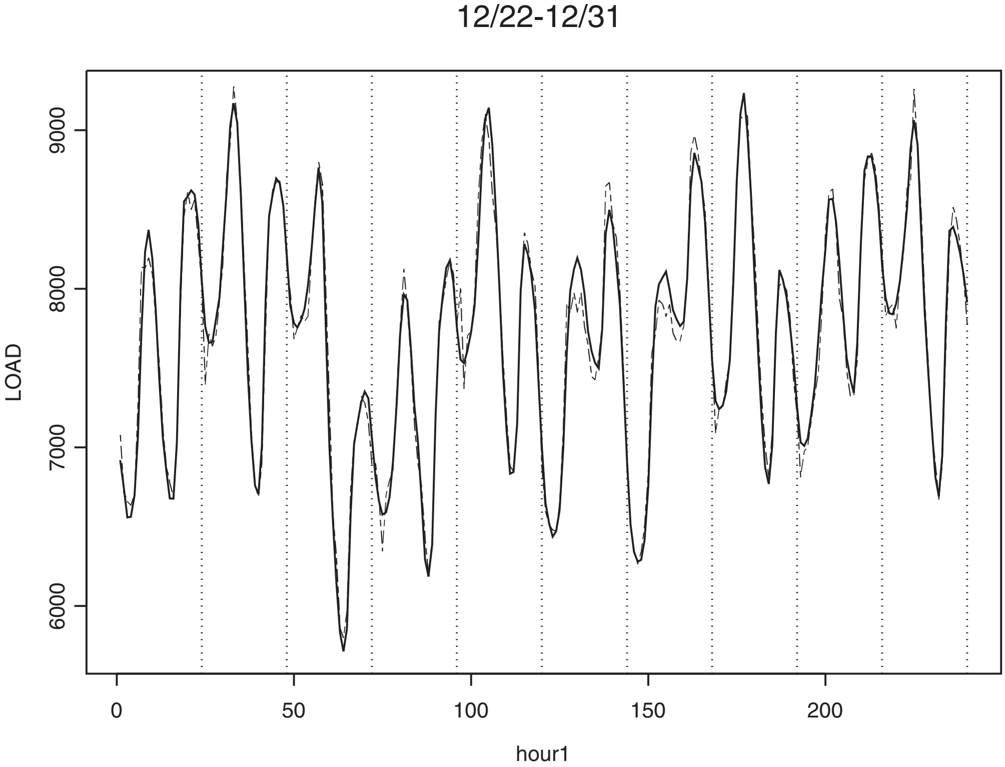

Figure 16. The actual observations (solid line) and the forecasts (dashed line) of Dec. 22 to Dec. 31, 2000

However, in this dataset data are more densely distributed along the temperature dimension than along the time dimension, so different amounts of smoothing may be required on different dimensions. One way to achieve this is to increase the time scale to a multiple of the original scale; i.e., define $T_t^* = m * T_t$, ($m \geq 1$) and substitute T_t^* for T_t in equation (7). m can be treated as an unknown parameter and selected together with the bandwidth λ, by GCV, i.e.:

$$(m,\lambda) = \arg_{m,\lambda} \min \frac{RSS_{m,\lambda}}{n\{1 - tr(\mathbf{H}_{m,\lambda})/n\}^2}$$

where RSS is the residual sum of squares of the model and **H** is the smoothing matrix. The bandwidth λ and m selected by GCV are 0.05 and 5, respectively. Repeat the entire analysis, including the final refinement using model (4); the resulting within-sample and post-sample RMSE are 91.93 and 102.95, respectively. Figures 20 and 21 show the estimated two-dimensional surfaces. The two-dimensional model performance is slightly inferior to that of the one-dimensional model. One possible explanation can be seen in Figures 12 and 13, where the data range for each curve can be very different. This 'boundary effect' of the nonparametric estimation may cause part of the accuracy problem.

DOI: 10.1002/for

Table V. Comparison between the SPTF model the and EGRV model

Hour	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Sun.	Row total
1	0	0	0	0	0	0	0	0
2	0	0	1	0	0	0	1	2
3	1	0	0	0	0	1	1	3
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	1	1
6	0	0	0	1	0	0	0	1
7	1	1	1	1	1	0	0	5
8	0	1	0	0	0	0	0	1
9	0	0	0	0	0	1	0	1
10	1	1	1	1	1	1	1	7
11	1	1	1	1	1	1	1	7
12	1	1	1	1	1	1	1	7
13	1	1	1	1	1	1	1	7
14	1	0	1	1	1	0	1	5
15	0	0	1	1	0	1	0	3
16	1	1	0	0	0	0	1	3
17	0	0	0	0	0	0	0	0
18	0	0	0	0	0	1	0	1
19	1	1	1	1	1	0	1	6
20	1	1	1	1	1	1	1	7
21	1	1	1	1	1	1	1	7
22	1	1	1	1	1	1	1	7
23	0	0	0	1	0	0	0	1
24	0	0	0	0	0	1	1	2
Col. total	12	11	12	13	10	12	14	84

Table VI. The 24-step-ahead performance of the SPTF and the EGRV models

	Within RMSE	Post RMSE	Within MAPE	Post MAPE
SPTF	96.29	108.61	1.12%	1.18%
EGRV	250.23	388.79	2.82%	3.74%

SUMMARY AND DISCUSSION

In this paper we considered a semi-parametric two-component modeling procedure for nonlinear time series data. The model consists of a nonparametric component and a parametric ARIMA component. The model estimation is carried out using a modified backfitting procedure. Both one-dimensional and two-dimensional models are considered. The proposed modeling methodology reveals a possibility of modeling data of the form $Y_t = f(X_t) + e_t$, where $f(\cdot)$ is an unknown smooth function and e_t follows an ARIMA model. Because $f(\cdot)$ is modeled nonparametrically, this model is very flexible and can be used to model highly nonlinear relationships of unknown functional form. By modeling e_t explicitly, the serial correlation is removed and the transfer function $f(\cdot)$ can be estimated more efficiently; additionally, the estimated ARIMA parameters can be used to improve forecasting performance. We applied this modeling procedure to model and forecast hourly electricity

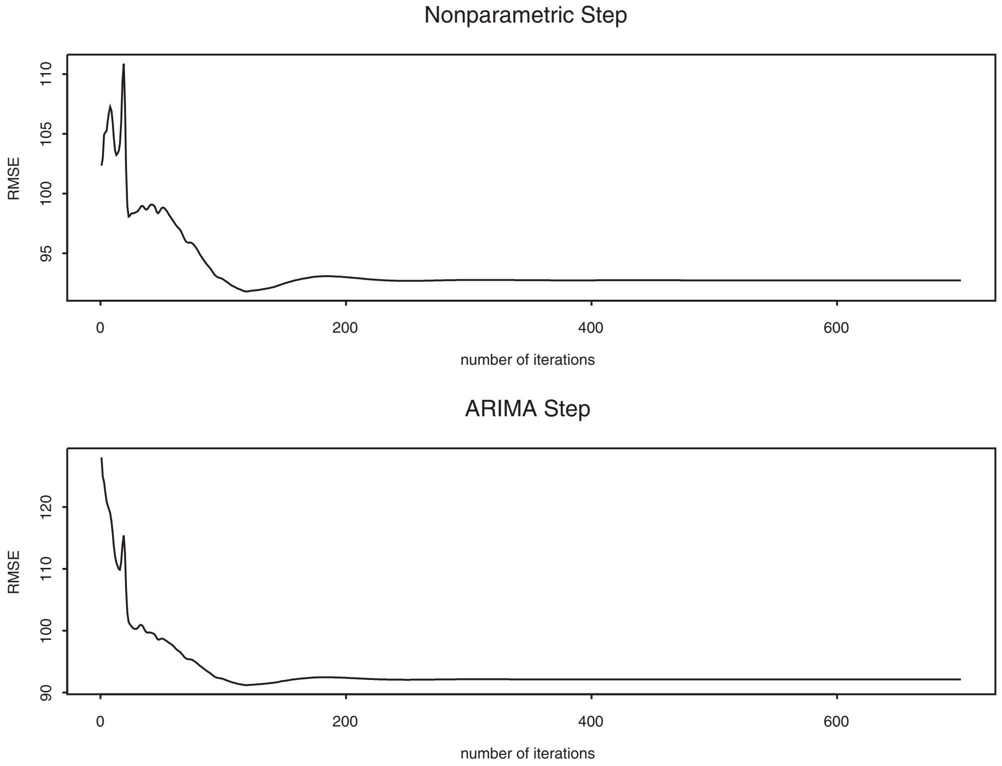

Figure 17. The within-sample RMSE versus number of iterations

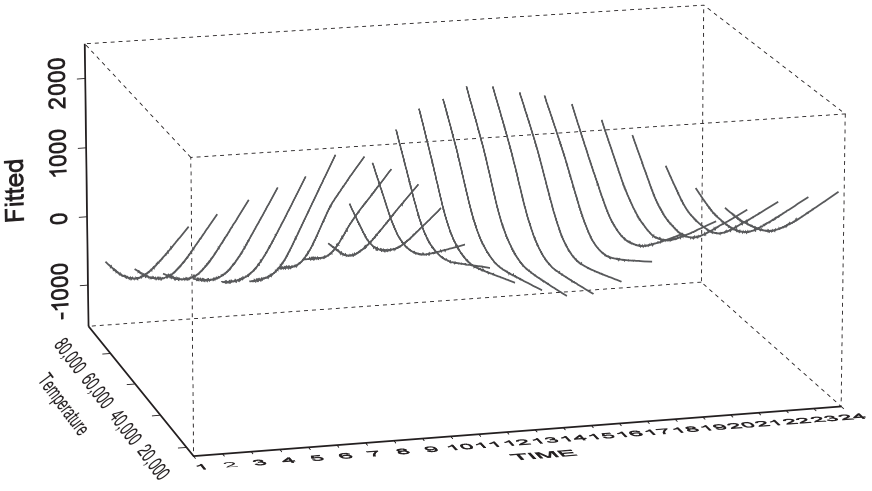

Figure 18. The nonparametric estimated response functions for workdays

DOI: 10.1002/for

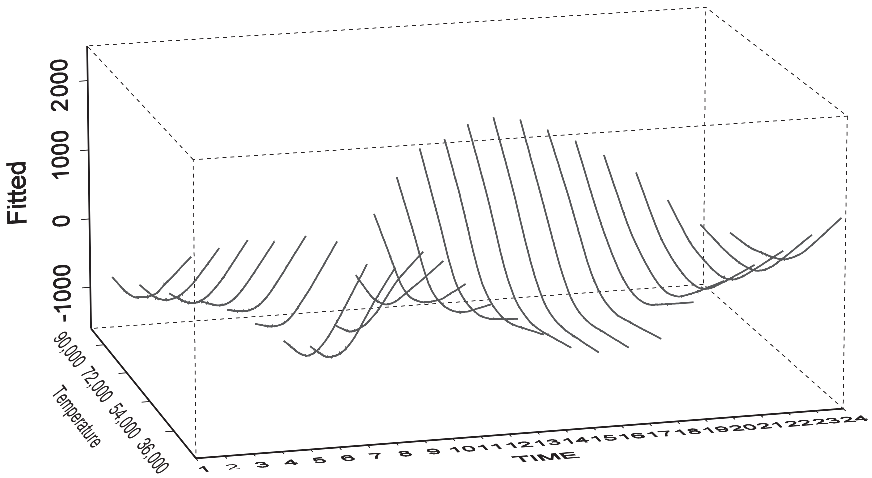

Figure 19. The nonparametric estimated response functions for non-workdays

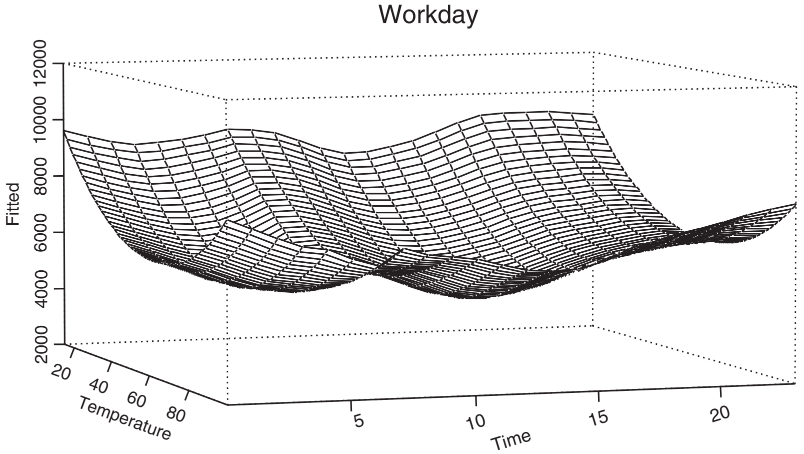

Figure 20. The nonparametric estimated surface for workdasys

load and found it performed very well. The proposed approach also has some disadvantages and remaining issues. First, it is more computationally intensive in estimation than its parametric counterparts; this is partly due to the nonparametric nature of the model and partly due to the iterative nature of the estimating algorithm. Second, because of the nonlinear nature of the model, the multi-step-ahead forecasts are more sophisticated; usually different models are needed for different forecasting horizons. Nonparametric methods with parametric functional forms, such as regression splines, can provide better solutions for these issues but the local asymptotic properties are more difficult to establish. Third, like other nonparametric methods, the proposed approach suffers from the *curse of dimensionality* (e.g., Hastie and Tibshitrani, 1991), which practically restricts the model to low dimensions. We need to consider more restrictive models, such as the additive model, to solve

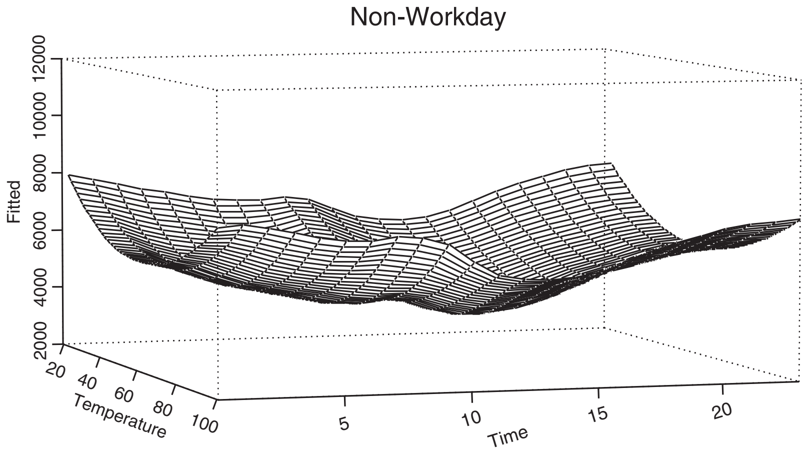

Figure 21. The nonparametric fitted surface for non-workdays

this problem. Finally, the theoretical properties of the model and the estimation procedure still require more careful and rigorous study.

ACKNOWLEDGEMENTS

Rong Chen's research is sponsored in part by NSF grant DMS 0244541 and NIH grant R01 Gm068958. The authors would like to thank the referee for helpful comments.

REFERENCES

Al-Zayer J, Al-Ibrahim AA. 1996. Modeling the impact of temperature on electricity consumption in the Eastern Province of Saudi Arabia. *Journal of Forecasting* **15**: 97–106.

Box GEP, Jenkins GM, Reinsel GC. 1994. *Time Series Analysis: Forecasting and Control* (3rd edn). Prentice-Hall: Englewood Cliffs, NJ.

Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statisical Association* **74**: 829–836.

Cleveland WS, Devlin SJ. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the Americian Statisical Association* **83**: 596–610.

Cottet R, Smith M. 2003. Bayesian modeling and forecasting of intraday electricity load. *Journal of the Americian Statistical Association* **98**: 839–849.

Craven P, Wahba G. 1979. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross validation. *Numerische Mathematik* **31**: 377–403.

Electric Power Research Institute. 1993. *Probabilistic Methods in Forecasting Hourly Loads* (EPRI TR-101902). *Palo Alto*, CA.

Engle RF, Granger CWJ, Rice J,Weiss A. 1986. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* **81**: 310–320.

Gupta PC. 1985. Adaptive short-term forecasting of hourly loads using weather information. In *Comparative Models for Electrical Load Forecasting*, Bunn DW, Farmer ED (eds). Wiley: New York; 43–56.

Hart JD, Vieu P. 1990. Data-driven bandwidth choice for density estimation based on dependent data. *Annals of Statistics* **18**: 873–890.

DOI: 10.1002/for

Harvey A, Koopman, SJ. 1993. Forecasting hourly electricity demand using time-varying splines. *Journal of the American Statistical Association* **88**: 1228–1236.

Hastie TJ, Tibshitrani RJ. 1991. *Generalized Additive Models*. Chapman & Hall: London.

Ho K, Hsu Y, Yang C. 1992. Short-term load forecasting using an multilayer neural network with an adaptive learning algorithm. *IEEE Transactions on Power Systems* **7**: 141–149.

Liu LM, Hudak GB, Box GEP, Muller ME, Tiao GC. 1992. *Forecasting and Time Series Analysis Using the SCA Statistical System*, Vol. 1. Scientific Computing Associates Corp: River Forest, IL.

Masry E, Fan J. 1997. Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics* **24**: 165–179.

Mendenhall W, Sincich T. 1996. *A Second Course in Statistics–Regression Analysis*. Prentice-Hall: Englewood Cliffs, NJ.

Peirson J, Henley A. 1994. Electricity load and temperature. *Energy Economics* **16**: 235–243.

Peng T, Hubele N, Karady G. 1992. Advancement in the application of neural networks for short-term load forecasting. *IEEE Transactions on Power Systems*, Series B **57**: 99–138.

Ramanathan R, Engle R, Granger CWJ, Vahid-Araghi F, Brace C. 1997. Short-run forecasts of electricity loads and peaks. *International Journal of Forecasting* **13**: 161–174.

Smith M. 2000. Modeling and short-term forecasting of New South Wales electricity system load. *Journal of Business and Economic Statistics* **18**: 465–478.

Wahba G, Wold S. 1975. A completely automatic French curve: fitting spline functions by cross-validation. *Communications in Statistics*, Series A **4**: 1–17.

Authors' biographies:

Jun M. Liu is Assistant Professor at the Department of Finance and Quantitative Analysis, Georgia Southern University. He received his PhD in Business Administration (with an inquiry in Business Statistics) from the University of Illinois at Chicago. His research interest includes linear and nonlinear/nonparametric time series analysis, forecasting methods, and statistical applications in business.

Rong Chen is Professor of Statistics at the Department of Information and Decision Sciences, University of Illinois at Chicago. During 2002–2005, he was also professor and department head of Department of Business Statistics and Econometrics, Peking University, China. Dr Chen received his BS (1985) in Mathematics from Peking University, and PhD (1990) in Statistics from Carnegie Mellon University. His main research interests are in nonlinear/nonparametric time series analysis, statistical computing and statistical applications. He is a fellow of American Statistical Association, and has served as an Associate Editor for four leading statistical journals.

Lon-Mu Liu is Professor of Information and Decision Sciences at the University of Illinois at Chicago. He received his PhD in Statistics from the University of Wisconsin–Madison. His research interest includes time series analysis, forecasting methods, econometric modeling and software development. He has written a book, *Time Series Analysis and Forecasting*, which focuses on practical applications of time series analysis.

John L. Harris is a principal in Enterprise Risk Management at Progress Energy. Prior to his current duties, he has held positions in economic assessment, financial forecasting, financial analysis, and strategic planning. He received his MA and PhD in Economics from the University of Illinois at Chicago Circle, an MS in Business from George Williams, and a BS in Physics from William & Mary. His research interests include forecasting, econometric modeling, time series analysis, and dynamic systems. He has written numerous articles on economic and financial matters.

Authors' addresses:

Jun M. Liu, Department of Finance and Quantitative Analysis, Georgia Southern University, PO Box 8151, Statesboro, GA 30460–8151, USA.

Rong Chen and Lon-Mu Liu, Department of Information and Decision Sciences, University of Illinois at Chicago, 601 South Morgan Street (M/C 294), Chicago, IL 60607, USA.

John L. Harris, Enterprise Risk Management Department, Progress Energy, Inc., Raleigh, NC 27609, USA.

DOI: 10.1002/for