Modeling and Short-Term Forecasting of New South Wales Electricity System Load

Michael SMITH
Econometrics and Business Statistics, University of Sydney, Sydney, NSW 2006, Australia (*mikes@econ.usyd.edu.au*)

This article employs Bayesian semiparametric regression methodology to model intraday electricity load data and obtain short-term load forecasts. The role of such forecasts in the New South Wales wholesale electricity market is discussed and the method applied to New South Wales system load data. The semiparametric regression model used identifies daily periodic, weekly periodic, and temperature-sensitive components of load. Each component is decomposed as a linear combination of basis functions, with a nonzero probability mass that the corresponding coefficients are exactly zero. Three possible models for the errors are also considered, including independent, autoregressive, and first-differenced autoregressive models. A moving window of data is used to overcome the slow time-varying nature of the temperature and periodic effects. The entire model is estimated using a Bayesian Markov chain Monte Carlo approach, and forecasts are obtained using a Monte Carlo sample from the joint predictive distribution of future system load. It is demonstrated how accurate temperature forecasts can result in accurate intraday system load forecasts for even quite long forecast horizons.

KEY WORDS: Bayesian semiparametric regression; Intraday load profiles; Markov chain Monte Carlo; Model averaging; Smoothing; Wholesale electricity market.

Electricity utilities and associated regulatory bodies throughout the world have long had a keen interest in forecasting the load of electricity systems. Such load forecasts can be roughly divided into two types. The first is produced for periods of several months or years into the future and usually uses data aggregated to the daily, weekly, or monthly level; for example, see Engle, Granger, and Hallman (1989). These forecasts are often considered medium or longer term and are useful for staffing, maintenance, and capital investment planning. The second type is produced on a shorter term basis using intraday data. These are traditionally produced between five minutes and two days ahead and are used to balance generating capacity (the components of which have a variety of fixed start-up times) with system demand. Accurate short-term forecasts have long been recognized as important; otherwise system security could be breached and blackouts or equipment damage occur. However, in recent years short-term forecasting has become more important with the advent of electricity reform.

During this decade the electricity industries in many of the developed nations of the world have begun a major restructuring process. The main emphasis of this change has been the replacement of traditional "vertically integrated" local electricity monopolies with wholesale electricity markets. Such markets involve several participant utilities that bid against each other for the sale and purchase of electricity. In the long run, it is hoped that this will reduce the cost of electricity to end-users. The commercial success of market participants largely depends on their ability to place bids to supply (or purchase) electricity that are competitive enough to result in generation (or distribution) levels that meet any regulatory obligations and commercial requirements and minimize costs. At the same time, generators would like to sell their electricity at the maximum price possible, while distributors would like to purchase it at the minimum price possible. Constructing an appropriately priced bid depends heavily on the expected level of system load, and therefore the ability of market participants to accurately forecast this load is essential.

This article highlights the role of short-term load forecasting in the newly introduced New South Wales (NSW) wholesale electricity market and develops a methodology to provide forecasts that meet the commercial needs of the participants. A semiparametric regression model is used for system load that captures the periodic "load profile" and temperature-sensitive component. The periodic behavior of system load is decomposed into weekly and daily components. An additive temperature effect is considered, along with a more general bivariate interaction with the time of day. The temperature-sensitive component is modeled separately for working days and nonworking days. Each of the unknown functions is represented as a linear combination of basis functions, where the user is free to select from a wide variety of potential bases, with several alternatives examined here. To overcome the slow time-varying nature of the periodic components, a moving window of data is used for in-sample purposes. However, information from previous periods on the functional form of each of these effects is incorporated into the model by including previous estimates as basis functions. Three distributional assumptions for the disturbances are examined, including a flexible autoregression, a flexible autoregression in the first-difference, and independent errors.

Bayesian semiparametric regression methodology, similar to that discussed by Smith and Kohn (1996) in the independent error case and Smith, Wong, and Kohn (1998) in the dependent-error case, is employed to estimate the

Journal of Business & Economic Statistics
October 2000, Vol. 18, No. 4

model. This uses a Bayesian hierarchical model to parameterize the nonzero possibility that each basis coefficient is zero and the corresponding basis term redundant. The hierarchical model is estimated using Markov chain Monte Carlo methods, including a correction of the "focused sampler" introduced by Wong, Hansen, Kohn, and Smith (1997) to overcome the computational burden associated with the high dimension of the basis representation of the regression functions. A proof that the focused sampler generates iterates from the correct invariant distribution is included in Appendix B. The resulting estimator is both automatic and applicable to complex nonparametric regression models with large sample sizes and a combination of smooth and nonsmooth functions. Forecasted loads are obtained from Monte Carlo samples drawn from the marginal predictive densities of each future half-hourly observation. From these, a full range of inference can be constructed, including predictive mean estimates and corresponding prediction intervals.

Many other authors have examined the problem of short-term load forecasting, with Moghram and Rahman (1989) summarizing a variety of early solutions. Most recent authors use artificial neural networks (ANN) to identify the nonlinear load profile and, in some research, nonlinear weather effects. For example, Lee, Cha, and Park (1992) applied the back-propagation algorithm to identify the periodic component of load but did not consider a weather input. Ho, Hsu, and Yang (1992) and Hsu and Yang (1995) considered the issue as two subproblems. They used an ANN to identify an intraday load profile normalized to [0, 1] and a second ANN to relate daily weather variables to maximum and minimum daily load. External forecasts for the weather variables were used to forecast maximum and minimum daily load, which in turn were combined with the estimated normalized load profile to obtain total system load forecasts. This approach effectively models load as a multiplication of separate nonlinear weather-sensitive and periodic intraday components, which are estimated separately. Peng, Hubele, and Karady (1992) also used an ANN with temperature extremes as input. Although all these authors have had some success in capturing the nonlinear mean periodic profile, they did not propose statistical models. An upshot of this is that forecast error intervals were not obtained and time series error processes were not considered. Only very short-term forecast horizons were considered—at most 24 hours and usually only a few hours ahead. However, as discussed in the Section 1, forecast horizons of at least 36 hours are required for use in the clearance mechanism of the NSW wholesale electricity market.

Another approach would be to use a seemingly unrelated regression model in which a separate parametric regression is used for each intraday period and the errors are correlated across equations. Fiebig, Bartels, and Aigner (1991) and Bartels and Fiebig (1998) employed such models to analyze electricity demand profiles for a variety of end usages, although it would be straightforward to extend these to system load models. Harvey and Koopman (1993) employed a semiparametric regression model for intraday load data. They represented the periodic component of load using two time-varying periodic natural smoothing spline models. The first uses a daily periodic component for weekdays, along with periodic correction factors for weekends, and the second model uses a single spline for an entire week. This work is extended here in several directions. First, it is demonstrated how other periodic bases, not just natural smoothing splines, can be employed to model the periodic component of system load. Second, Harvey and Koopman carefully preselected a set of basis terms to produce a parsimonious representation of the periodic load profiles. The authors stressed that this process is difficult, and the approach suggested here overcomes the problem by placing knots at all possible design locations and ascribing a nonzero probability mass that the corresponding coefficients are exactly zero. Third, because a small number of basis terms do not have to be preselected, the periodic component can be decomposed into full daily and weekly periodic effects rather than employing parsimonious correction factors. Moreover, it is viable to consider separate temperature effects for working and nonworking days, as well as a general interaction of temperature with time of day. Fourth, the smoothing methodology is locally adaptive (Smith and Kohn 1996), which is important because the component functions require different degrees of smoothing—something that is not provided by global bandwidth based procedures. Fifth, the approach allows information on the functional form of the periodic profiles from previous years to be employed. Last, dependent error structures can be incorporated into the model and estimated simultaneously with the signal.

The article is structured as follows. Section 1 details the role of short-term forecasting in the NSW wholesale electricity market and discusses the data. Section 2 proposes a variety of semiparametric regression models for load data, including possible error distributions and different temperature effects. Section 3 outlines the Bayesian estimation methodology and how to obtain samples from the marginal predictive densities. Section 4 provides in-sample and forecasting results when the methodology is applied to NSW load data, and Section 5 concludes the article. Appendix A contains calculations required to implement the sampling schemes, and Appendix B contains a proof that the focused sampler has the correct invariant distribution.

1. THE NSW MARKET AND LOAD DATA

1.1 The NSW Wholesale Electricity Market

New South Wales (NSW) is the most populous state in Australia, with a population of 6.2 million people, 3.9 million of whom are resident in the state capital of Sydney (Source: Australian 1996 Census, summarized in *Australian Demographic Statistics,* Australian Bureau of Statistics, series 3101.0). A statewide electricity grid has been established for many years, and electricity generating capacity is located throughout the state, although the majority of demand is located in the Sydney metropolitan region. The NSW state government has been at the forefront in implementing a wholesale electricity market that, it is envisaged, will eventually encompass all Australian states and territo-

ries. As of 1998 this market is regulated by the National Electricity Market Management Corporation (NEMMCO).

The wholesale market currently operates as follows (this is the market as of early 1999). The suppliers and distributors electronically lodge schedules of bids for the sale and purchase of electricity with NEMMCO at 12:30 the day prior to actual dispatch of the electricity. NEMMCO compiles this bid data and uses it, along with a short-term forecast of system demand and grid capacity, to determine an expected dispatch amount and a dispatch order of generators. At the beginning of each five minute actual dispatch period, an automated estimate of system demand is calculated for the end of the five-minute period and fed into an automatic generation control (AGC) system electronically. The AGC system then adjusts actual electricity dispatch (given the standing bid data and rate of change constraints) automatically.

Load forecasts play a central role in this market. First, forecasts with at least a 36-hour horizon are required by NEMMCO for the purposes of determining a preliminary dispatch schedule. These forecasts are recalculated at regular intervals and made available to market participants, such as the utilities and energy traders who provide liquidity in the wholesale market. Second, there is a major incentive for both suppliers and distributors to assess the quality of these forecasts. If these utilities can construct a significantly superior forecasting methodology, then they can use this to optimize their bids to better meet their generation, distribution, and income aims.

1.2 NSW Load Data

The historical data available to market participants include half-hourly observations on the following two variables:

L_t: Mean NSW system load during the half-hour ending at time t (megawatts)

T_t: Sydney temperature at time t as measured at the Sydney Weather Bureau (degrees Celsius, accurate to one decimal place)

As noted by previous analysts, system load exhibits a remarkable level of daily, weekly, and yearly periodic behavior. For example, Figure 1(a) plots the half-hourly system load for the summer week starting Monday, January 6, 1992, to Sunday, January 12, 1992. Daily periodic behavior is apparent, with about 04:00 being the daily load low and 10:30 and 23:30 being daily load peaks on any day. There is also a weekly periodic behavior, with Saturdays and Sundays having very different load profiles to weekdays. Figure 1, (b)–(d), shows the system loads for the

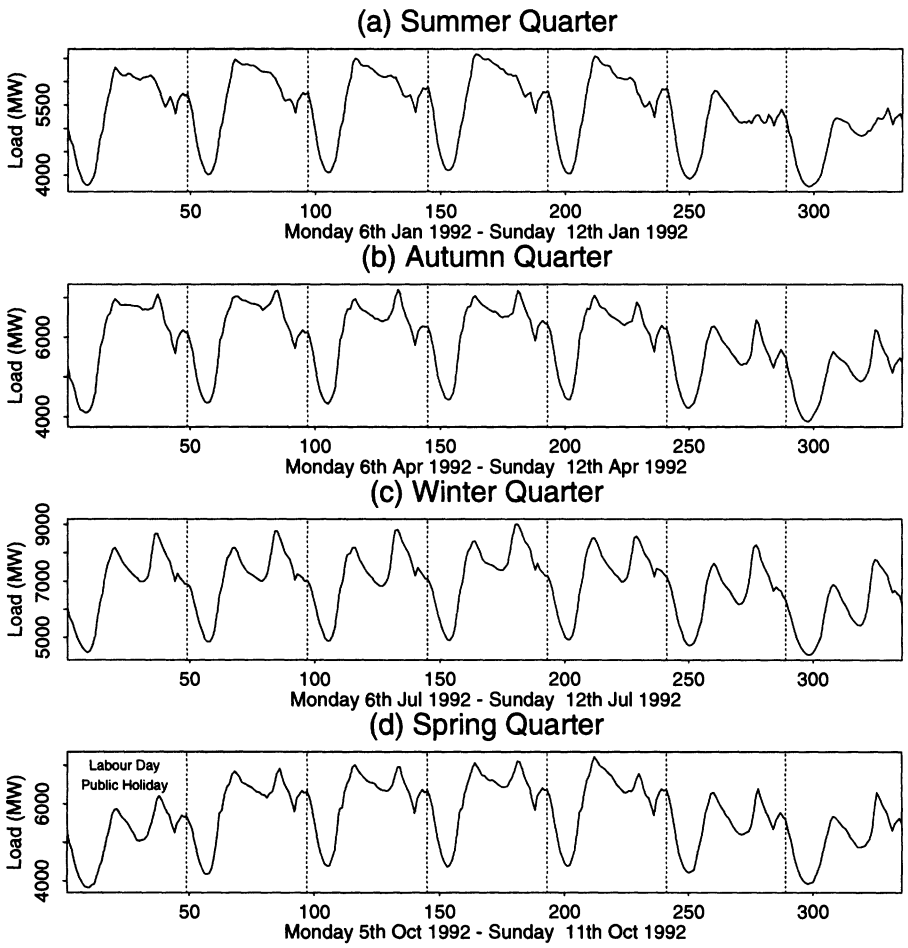

Figure 1. NSW System Load for Four Weeks in 1992. Each week roughly corresponds to each of the four southern hemisphere seasons. In each panel, the bold line is the system load, and the vertical lines are at 24:00 and separate the days. In panel (d) the Monday was a Labour Day public holiday and exhibits a profile similar to a Sunday.

13th, 26th, and 39th subsequent weeks. Together, the four weeks roughly correspond to examples from the southern hemisphere summer, autumn, winter, and spring quarters. These demonstrate that the system load profiles (and the daily and weekly periodic components) can vary over the year. For example, the daily peak occurs around 17:30 in the winter [Fig. 1(c)] compared to 10:30 in summer [Fig. 1(a)].

The most popular way to account for this seasonal evolution in periodic profiles is to use a moving window of sample data, over which the periodic components are effectively static, to estimate the periodic components. For example, Saha and Senimi (1996), Hsu and Yang (1995), and Ho et al. (1992) all use windows of ten previous similar days to train their ANN forecasting procedures. Following these authors and the advice of industry forecasters, a moving window of three weeks of half-hourly load and Sydney temperature data is used, resulting in a sample size of $n = 1,008$ observations. (I am grateful for discussions with Transgrid staff with regard to this issue.) Other window widths can easily be employed by potential users, although high quality forecasts are obtained with this window width for the NSW data.

Figure 1(d) also provides the load profile of the 1992 Labour Day public holiday, which fell on Monday, October 5. This profile is very similar to that of the following Sunday (October 11). Therefore, for the purposes of this analysis, public holidays are treated as Sundays, although a more complete analysis would treat each of these days as distinct day types.

Weather variables, such as cloud cover, wind speed, temperature, and humidity, are also likely to affect system load. However, there is evidence that in many locations in the world temperature is the most important weather variable (Hsu and Yang 1995; Hyde and Hodnett 1997). Engle, Granger, Rice, and Weiss (1986) explored the form of the relationship using aggregate load data from U.S. utilities and found that it is similar to a V shape, with a minimum around 18.3°C (65°F). However, it is possible that the relationship flattens out at extreme temperature ranges and that the location of the minimum may be different for NSW data and may vary over the seasons.

2. MODELS FOR INTRADAY SYSTEM LOAD

2.1 Semiparametric Regression Models

Semiparametric regression is used in this article to model the sample window of intraday system load data. The periodic load profile is decomposed additively into a daily periodic component P_{dy}, weekly periodic component P_{wk}, and an interweek trend I. Such a decomposition is more parsimonious than employing only a weekly effect and interweek trend as Harvey and Koopman (1993) did because the basic diurnal pattern can be captured once by P_{dy}, rather than repeated seven times in P_{wk}. The end result is that the periodic profile can be estimated using less basis terms in a semiparametric regression model.

Following previous authors, such as Harvey and Koopman (1993), Hsu and Yang (1995), and Wong (1994), a temperature effect on system load, over and above the periodic profile, is also considered. Other weather variables, such as humidity, can be incorporated into the model in a similar manner. Unlike in the work of previous authors, the form of the temperature-sensitive component of load is allowed to differ for working and nonworking days. The rationale behind this distinction is that workplace reactions to climate changes are likely to be very different to residential reactions. The simplest temperature component considered is additive, resulting in the semiparametric regression model

$$\text{Model 1: } L_t = I(t) + P_{\text{dy}}(\text{TOD}_t) + P_{\text{wk}}(\text{TOW}_t) + W_t(\alpha_1 + f_1(T_t)) + (1 - W_t)(\alpha_2 + f_2(T_t)) + u_t.$$

Here, TOD_t and TOW_t are the "time of day" and "time of week" of the tth half-hourly observation normalized to (0, 1]. The dummy variable W_t is equal to 1 if observation t occurs on a working day and 0 otherwise. The scalers α_1 and α_2 are working- and nonworking-day intercepts, and f_1 and f_2 are additive working- and nonworking-day temperature effects, respectively.

A more general model is to consider the response of system load to changes in temperature as an interaction with the time of day, so that

$$\text{Model 2: } L_t = I(t) + P_{\text{dy}}(\text{TOD}_t) + P_{\text{wk}}(\text{TOW}_t) + W_t(\alpha_1 + f_3(T_t, \text{TOD}_t)) + (1 - W_t)(\alpha_2 + f_4(T_t, \text{TOD}_t)) + u_t.$$

The functions f_3 and f_4 are general bivariate response surfaces for working and nonworking days, respectively; they are more general than the multiplicative model suggested by Hsu and Yang (1995). The three following possible distributions for the random error u_t are considered:

(A1) *Independent errors,* in which u_t are iid $\text{N}(0, \sigma^2)$.

(A2) *Stationary autocorrelated errors,* in which $u_t = \sum_{i=1}^{r} \phi_i u_{t-i} + e_t$ and e_t are iid $\text{N}(0, \sigma^2)$. The true order of the autoregression is unknown and is estimated, given the maximal order r. In the empirical work, the maximal order is $r = 12$.

(A3) *Autocorrelated first-differenced errors,* in which $v_t = u_t - u_{t-1}$, $v_t = \sum_{i=1}^{r} \phi_i v_{t-i} + e_t$, and e_t are iid $\text{N}(0, \sigma^2)$. Without loss of generality, the preperiod error is fixed so that $u_0 = 0$ because there is a mean in the regression model. As in (A2), the order of the autocorrelated series is estimated from a maximal order r, where again $r = 12$ in the empirical work.

2.2 Linear Basis Representations

In the preceding semiparametric regressions, each of the unknown functions $P_{\text{dy}}, P_{\text{wk}}, f_1, \ldots, f_4$ is modeled as a linear combination of preselected basis terms. For any component function, f say,

$$f(\mathbf{x}) = \sum_i \beta_i b_i(\mathbf{x}),$$

where $\mathbf{x}$ is a vector of independent variable values, β_i are unknown coefficients, and $b_i \in \mathbf{B}$ are basis functions from a basis $\mathbf{B}$.

For the univariate periodic functions P_{dy} and P_{wk}, bases should be used that are periodic on (0, 1]; for example, Har-

vey and Koopman (1993) employed periodic natural cubic splines. However, in this study the following two periodic radial bases suggested by Luo and Wahba (1997) are used:

1. *Quartic Periodic Radial Basis* ($\mathcal{B}_1$): Here, the basis terms are $b_i(x) = R(x, k_i)$, for $i = 1, \ldots, m$, where

$$R(s,t) = \left(\left(-(|s-t|) - \frac{1}{2}\right)^4 + \left(|s-t| - \frac{1}{2}\right)^2/2 - \frac{7}{240}\right)/24.$$

2. *Quadratic Periodic Radial Basis* ($\mathcal{B}_2$): Here, $b_i(x) = ((|x-k_i| - \frac{1}{2})^2 - 1/12)/2$ for $i = 1, \ldots, m$.

The basis terms require location points (often called "knots"), which are taken as all possible values of the independent variables and denoted $k_1, \ldots, k_m$. In the case of the daily periodic component, there are $m = 48$ half-hours in a day, so that $k_i = i/48$, for $i = 1, \ldots, 48$. For the weekly periodic component, there are $m = 336$ half-hours in a week, so $k_i = i/336$, for $i = 1, \ldots, 336$.

The additive temperature effects f_1 and f_2 are thought to be smooth functions, for which there are many potential bases. Here, natural cubic splines (Wahba 1990) are used to model these functions, where

$$b_i(x) = \begin{cases} \frac{1}{2}x^2\left(k_i - \frac{1}{3}x\right) & \text{if } x \le k_i \\ \frac{1}{2}k_i^2\left(x - \frac{1}{3}k_i\right) & \text{if } x > k_i \end{cases} \quad \text{for } i = 1, \ldots, m. \quad (1)$$

As undertaken by Smith and Kohn (1996), a grid of $m = 20$ knots is chosen that follows the densities of the in-sample observed values of temperature on working days for f_1 and on nonworking days for f_2. Following Wahba (1990), the basis is augmented with a linear term $b_{21}(x) = x$ to make the representation of the regression functions more parsimonious.

The bivariate surfaces f_3 and f_4 are defined over the bivariate variable $\mathbf{x} = (T, \text{TOD})'$. To retain periodicity of the regression mean in TOD, the bases used should ensure that $f_3(\cdot, 0) = f_3(\cdot, 1)$ and $f_4(\cdot, 0) = f_4(\cdot, 1)$. One way to achieve this is to employ a bivariate in which the terms are multiplications of univariate quartic periodic radial basis functions so that $b_i(\mathbf{x}) = R(T, k_{1,i})R(\text{TOD}, k_{2,i})$. Here, the knot sequence of the basis terms is $\mathbf{k}_1, \ldots, \mathbf{k}_m$, where $\mathbf{k}_i = (k_{1,i}, k_{2,i})'$. Because the two bivariate surfaces should not be periodic in temperature, this basis is augmented with a linear term in the first argument, T. The knots $\mathbf{k}_i$ are located at all the observations, although this is not always necessary in empirical work as demonstrated by Smith and Kohn (1997). Other bases formed by multiplying together different periodic univariate basis terms, or alternative periodic radial bases, could just as easily be employed.

Lastly, a simple linear interweekly trend is employed, so that $I(t) = \delta t$. A more sophisticated analysis may include nonlinear trends obtained from the analysis of several years of data.

2.3 Incorporating Historical Information on System Load Profiles

Figure 2 plots NSW load for the first Thursday in May from 1991 to 1994. It demonstrates that information about the functional form of the periodic components P_{dy} and P_{wk} can be gained from their estimates from the equivalent period last year (say, $\hat{P}_{\text{dy}}^{\text{last year}}$ and $\hat{P}_{\text{wk}}^{\text{last year}}$). Moreover, because a running window of in-sample data is used each day to reestimate the regression model, the current estimates of P_{dy} and P_{wk} are unlikely to be very different from those uncovered the previous day (say, $\hat{P}_{\text{dy}}^{\text{prev. day}}$ and $\hat{P}_{\text{wk}}^{\text{prev. day}}$). This information can be incorporated by augmenting the bases for the periodic effects with their historical estimates, so that

$$P_{\text{dy}}(x) = \sum_{i=1}^{48} \beta_i^d b_i(x) + \beta_{49}^d \hat{P}_{\text{dy}}^{\text{prev. day}}(x) + \beta_{50}^d \hat{P}_{\text{dy}}^{\text{last year}}(x) \quad (2.1)$$

and

$$P_{\text{wk}}(x) = \sum_{i=1}^{336} \beta_i^w b_i(x) + \beta_{337}^w \hat{P}_{\text{wk}}^{\text{prev. day}}(x) + \beta_{338}^w \hat{P}_{\text{wk}}^{\text{last year}}(x). \quad (2.2)$$

Here, β_i^d and β_i^w are regression coefficients that require estimation from the data. Notice that because all the basis terms, including the historical estimates, are periodic on (0, 1], P_{dy} and P_{wk} will be also.

3. METHODOLOGY

3.1 Introduction

Both semiparametric regression models in Section 2.1 can be written as a linear model, $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{u}$. Here, $\mathbf{y} = (L_1, \ldots, L_n)', \mathbf{u} = (u_1, \ldots, u_n)', X$ is an $n \times p$ design matrix made up of all the basis terms of the component functions evaluated at each of the n observations, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ contains the unknown regression coefficients. Table 1 provides the number of basis terms for each of the component functions, along with the total number in the two regression models. To estimate these regressions the uncertainty that each basis term is redundant must

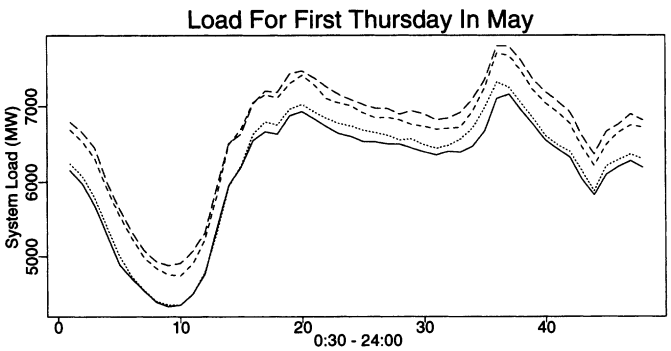

Figure 2. Plot of the System Load for the First Thursday in May From 1991 to 1994: ——, May 2, 1991; - - -, May 7, 1992; – – –, May 6, 1993; — — —, May 5, 1994.

Table 1. Number of Basis Terms in Each of the Component Functions for the Regression Models, Both Augmented and Not Augmented With Historical Estimates of P_{dy} and P_{wk}

Model	$\delta, \alpha_1, \alpha_2$	P_{dy}	P_{wk}	f_1	f_2	f_3	f_4	Total
			Augmented					
1	3	50	338	21	21	—	—	p = 433
2	3	50	338	—	—	721	289	p = 1,401
			Not augmented					
1	3	48	336	21	21	—	—	p = 429
2	3	48	336	—	—	721	289	p = 1,397

NOTE: For Model 2, basis terms for the temperature effects f_3 and f_4 are located at all observations during a three-week window with no public holidays, for working and nonworking days, respectively.

be taken into account. One alternative is to use a search algorithm, such as that discussed by Friedman (1991), to identify a single subset of basis terms, the coefficients of which are estimated using least squares. However, this does not account for the probability associated with all the other subsets, and is hard to extend to the flexible autoregressions at (A2) and (A3). Instead, a Bayesian hierarchical model is used to explicitly parameterize the uncertainty that each basis term enters the regression. Markov chain Monte Carlo procedures are used to estimate the model, given a distributional assumption for the errors. Estimates of the parameters are obtained by averaging over the subsets of basis terms generated by the sampling scheme, resulting in what are sometimes called "model averages" (Raftery, Madigan, and Hoeting 1997). Smith and Kohn (1996) developed such an estimator for estimating functions when the errors are independent, and Smith et al. (1998) developed one when the errors are autocorrelated. The reader is referred to these articles for an in-depth discussion, and descriptions of the Bayesian model, priors, and sampling schemes will be provided in Sections 3.2 and 3.3 for the independent and dependent error cases, respectively.

3.2 Independent Errors

A vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ of binary indicator variables is introduced to denote the nonzero elements of $\boldsymbol{\beta}$, with $\beta_i = 0$ iff $\gamma_i = 0$ and $\beta_i \neq 1$ iff $\gamma_i = 1$. Denoting the nonzero regression parameters as $\boldsymbol{\beta}_\gamma$ and corresponding columns of the design matrix as X_γ, the linear model can be rewritten conditional on $\boldsymbol{\gamma}$ as

$$\mathbf{y} = X_\gamma \boldsymbol{\beta}_\gamma + \mathbf{u}. \tag{3.1}$$

Following O'Hagan (1995) a fractional conditional prior for $\boldsymbol{\beta}_\gamma$ is used to construct the hierarchical model, with $p(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \sigma^2) \propto p(\mathbf{y}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, \sigma^2)^{1/n}$ so that $\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \sigma^2 \sim \mathrm{N}(\boldsymbol{\mu}_\gamma, n\sigma^2(X_\gamma' X_\gamma)^{-1})$, where $\boldsymbol{\mu}_\gamma = (X_\gamma' X_\gamma)^{-1} X_\gamma' \mathbf{y}$. The prior $p(\sigma^2|\boldsymbol{\gamma}) \propto 1/\sigma^2$, but the prior on the vector of binary indicator variables $\boldsymbol{\gamma}$ is different from the flat prior used by Smith and Kohn (1996). Because each unknown function may require a different number of basis terms, the prior differs depending on which function the indicator γ_i corresponds. If γ_i corresponds to function j, say, then the prior $p(\gamma_i|\omega_j) = \omega_j$, where ω_j is a hyperparameter with a uniform hyperprior on [0, 1]. Counting the intercepts and linear trend as functions, there are seven functions in both regression models, $(\delta, \alpha_1, \alpha_2, P_{\mathrm{dy}}, P_{\mathrm{wk}}, f_1, f_2)$ and $(\delta, \alpha_1, \alpha_2, P_{\mathrm{dy}}, P_{\mathrm{wk}}, f_3, f_4)$, which correspond to the hyperparameters $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_7)$. The elements of $\boldsymbol{\omega}$ are assumed a priori independent.

Because the parameter $\boldsymbol{\gamma}$ has support on 2^p different possible values, direct enumeration of the posterior distribution of the parameters is computationally infeasible. To overcome this problem, the following Markov chain Monte Carlo sampling scheme can be used:

Sampling Scheme.

(0): Select initial state $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{[0]}$.

(1): Generate from $\gamma_i|\boldsymbol{\gamma}_{j\neq i}, \mathbf{y}$ for $i = 1, \ldots, p$.

Step (1) is repeated many times, resulting in a Markov chain that converges to its invariant distribution $\boldsymbol{\gamma}|\mathbf{y}$. The sampler is run for a warmup period of 2,000 iterations, followed by a sampling period of $J = 1,500$ iterations in which the Monte Carlo sample $\{\boldsymbol{\gamma}^{[1]}, \ldots, \boldsymbol{\gamma}^{[J]}\}$ is collected.

Generation directly from the conditional distribution at step (1) results in the Gibbs sampling scheme introduced by Smith and Kohn (1996). When p is large, however, this can be computationally burdensome, and a "focused sampling" step is used instead. This method employs a transition kernel that is faster to compute than generating directly from the conditional distribution at step (1) but has the same invariant distribution. It uses the Bayesian decomposition of the posterior

$$\underbrace{p(\gamma_i|\boldsymbol{\gamma}_{j\neq i}, \mathbf{y})}_{\pi^*(\gamma_i)} \propto \underbrace{p(\mathbf{y}|\boldsymbol{\gamma})}_{l(\gamma_i)} \underbrace{p(\gamma_i|\boldsymbol{\gamma}_{j\neq i})}_{\pi(\gamma_i)}, \tag{3.2}$$

where π^*, l, and π denote the posterior, likelihood, and conditional prior, respectively. Appendix A.1 calculates these quantities, and it can be seen that evaluating l involves substantial computations, whereas π involves only minimal computation. The focused sampling step to generate an iterate of γ_i is as follows:

Focused Sampling Step. Generate u from a uniform distribution on (0, 1), then

1. if currently $\gamma_i = 1$ and $u > \pi\ (\gamma_i = 0)$, then set $\gamma_i = 1$;
2. if currently $\gamma_i = 1$ and $u < \pi\ (\gamma_i = 0)$, then generate γ_i from the density

$$\Pr(\gamma_i = 0) = \frac{l\ (\gamma_i = 0)}{l\ (\gamma_i = 1) + l\ (\gamma_i = 0)};$$

3. if currently $\gamma_i = 0$ and $u > \pi\ (\gamma_i = 1)$, then set $\gamma_i = 0$;
4. if currently $\gamma_i = 0$ and $u < \pi\ (\gamma_i = 1)$, then generate γ_i from the density

$$\Pr(\gamma_i = 1) = \frac{l\ (\gamma_i = 1)}{l\ (\gamma_i = 1) + l\ (\gamma_i = 0)}.$$

Appendix B contains a proof that this sampling step has the posterior distribution $\pi^*(\gamma_i)$ as its invariant distribution. Note that in cases 1 and 3 only $\pi(\gamma_i)$ has to be calculated. In semiparametric regression problems with function decompositions involving many basis terms, most of these terms are likely to be superfluous so that $\pi\ (\gamma_i = 1)$ is usually

close to zero. Therefore, case (3) is likely to occur most frequently and substantial computational savings are obtained over generating directly from the posterior, which requires $l(\gamma_i)$ to be calculated for all p indicators.

A mixture estimate of the posterior mean $E(\boldsymbol{\beta}|\mathbf{y})$ can be calculated using the Monte Carlo sample as $\hat{\boldsymbol{\beta}} = 1/J \sum_{j=1}^{J} E(\boldsymbol{\beta}|\boldsymbol{\gamma}^{[j]}, \mathbf{y})$, which is a model average. Estimates of the component functions in the regression models in Section 2.1 are then easily calculated at any point z in the domain of the independent variables by evaluating the corresponding basis terms at z and multiplying them by the relevant elements of $\hat{\boldsymbol{\beta}}$.

3.3 Autocorrelated Errors

Smith et al. (1998) extended the hierarchical model at Equation (3.1) to the case in which the errors are modeled as a flexible autoregression [see (A2)]. This framework is employed here and also holds for the autoregression in the first-differenced errors considered at (A3). The resulting procedure simultaneously estimates (a) the component functions, (b) the autoregressive coefficients $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_r)'$, and (c) the order of the autoregression, given the maximum order $r = 12$.

This approach reparameterizes the autoregressive parameters $\boldsymbol{\phi}$ in terms of the partial autocorrelations $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_r)'$. This is possible because there is a one-to-one transformation between $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ and allows for the convenient enforcement of stationarity by restricting the domain of the partial autocorrelations so that $-1 < \psi_i < 1$ for $i = 1, \ldots, r$.

To facilitate selecting the autoregressive order, indicator variables $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_r)'$ are introduced to denote the nonzero partial autocorrelations. In the same spirit as subset selection for regression parameters, a conditional prior $\boldsymbol{\psi}|\boldsymbol{\kappa}, \sigma^2$ can be defined, where $\psi_1, \ldots, \psi_r$ are a priori conditionally independent with

$$\begin{aligned} \psi_i|\kappa_i = 0, \sigma^2 &= 0 \\ \psi_i|\kappa_i = 1, \sigma^2 &\sim \text{uniform}(-1, 1). \end{aligned}$$

The prior probabilities $p(\kappa_i = 1), i = 1, \ldots, r$, are prescribed by the user. In the empirical work presented here, a descending prior on the order is used, where $p(\kappa_1 = 1) = .5, p(\kappa_2 = 1) = .4, p(\kappa_3 = 1) = .3, p(\kappa_4 = 1) = .2$, and $p(\kappa_i = 1) = .1$ for $i = 5, \ldots, r$. This prior favors lower-order autoregressions, although Smith et al. (1998) demonstrated that empirical results appear to be insensitive to this prior.

The formulation in terms of $\boldsymbol{\psi}$ allows the error variance to be explicitly calculated as $\text{var}(\mathbf{u}) = \sigma^2 \Omega_\psi$, where Ω_ψ is a matrix depending only on $\boldsymbol{\psi}$ and the data $\mathbf{y}$. Note, however, that the form of Ω_ψ will be different depending on whether an autoregression is assumed in the errors u_t, as in (A2), or in the first-differenced error sequence $v_t = u_t - u_{t-1}$, as in (A3). To extend the hierarchical model in Section 3.2 to the time-dependent error case, a fractional prior for $\boldsymbol{\beta}_\gamma$ is used that conditions on $\boldsymbol{\psi}$—that is, where $p(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \sigma^2, \boldsymbol{\psi}, \boldsymbol{\kappa}) \propto p(\mathbf{y}|\boldsymbol{\gamma}, \sigma^2, \boldsymbol{\psi}, \boldsymbol{\beta}_\gamma, \boldsymbol{\kappa})^{1/n}$, which results in the prior $\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \sigma^2, \boldsymbol{\psi}, \boldsymbol{\kappa} \sim \text{N}(\boldsymbol{\mu}_{\gamma,\psi}, n\sigma^2(X_\gamma' \Omega_\psi^{-1} X_\gamma)^{-1})$, where $\boldsymbol{\mu}_{\gamma,\psi} = (X_\gamma' \Omega_\psi^{-1} X_\gamma)^{-1} X_\gamma' \Omega_\psi^{-1} \mathbf{y}$. Finally, the priors for $\sigma^2, \boldsymbol{\gamma}$, and $\boldsymbol{\omega}$ are unchanged from those in Section 3.2, and $\boldsymbol{\kappa}, \boldsymbol{\gamma}$, and σ^2 are assumed to be a priori independent.

Estimation of this model is undertaken using the sampling scheme discussed in Smith et al. (1998) but with the focused sampling step outlined in Section 3.2 for generating $\boldsymbol{\gamma}$ from its conditional posterior distribution being implemented. The reader is referred to this work for details on the sampling scheme, the end result of which is a Monte Carlo sample $\{\boldsymbol{\beta}^{[j]}, \boldsymbol{\kappa}^{[j]}, \boldsymbol{\psi}^{[j]}, \boldsymbol{\gamma}^{[j]}\}_{j=1}^{J}$ from the joint posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\psi}, \boldsymbol{\gamma}|\mathbf{y})$ on which inference can be based. The posterior means $E(\boldsymbol{\beta}|\mathbf{y})$, $E(\boldsymbol{\psi}|\mathbf{y})$, and $E(\kappa_i|\mathbf{y}) = p(\kappa_i = 1|\mathbf{y})$ are estimated using the mixture and histogram estimates

$$\hat{\boldsymbol{\beta}} = \frac{1}{J} \sum_{j=1}^{J} E(\boldsymbol{\beta}|\boldsymbol{\gamma}^{[j]}, \boldsymbol{\psi}^{[j]}, \boldsymbol{\kappa}^{[j]}, \mathbf{y}), \qquad \hat{\boldsymbol{\psi}} = \frac{1}{J} \sum_{j=1}^{J} \boldsymbol{\psi}^{[j]},$$

and

$$\hat{\kappa}_i = \frac{1}{J} \sum_{j=1}^{J} p(\kappa_i = 1|\psi_{k\neq i}^{[j]}, \kappa_{k\neq i}^{[j]}, \boldsymbol{\gamma}^{[j]}, \boldsymbol{\beta}^{[j]}, \mathbf{y}), \qquad i = 1, \ldots, r,$$

which can be calculated as in Smith et al. (1998). The estimate $\hat{\boldsymbol{\phi}}$ of the autoregressive parameters is obtained from $\hat{\boldsymbol{\psi}}$ using the one-to-one relationship. As with the independent error case, the function estimates are calculated by multiplying the appropriate elements of $\hat{\boldsymbol{\beta}}$ with the corresponding basis terms. Histogram estimates of the uncorrelated residuals and fitted values can also be calculated from the Monte Carlo iterates as

$$\hat{\mathbf{e}} = \frac{1}{J} \sum_{j=1}^{J} \Omega_{\psi^{[j]}}^{-1/2} (\mathbf{y} - X_{\gamma^{[j]}} \boldsymbol{\beta}_\gamma^{[j]}), \qquad \hat{\mathbf{y}} = \mathbf{y} - \hat{\mathbf{e}}.$$

As in the independent error case, 2,000 iterates are used for the warmup and $J = 1,500$ iterations as a sampling period.

3.4 Constructing Forecasts

Bayesian forecasts can be obtained by evaluating the joint predictive density $p(\tilde{\mathbf{y}}|\mathbf{y})$ for n_2 subsequent future observations of system load $\tilde{\mathbf{y}} = (y_{n+1}, \ldots, y_{n+n_2})'$, with $y_t = L_t$. This cannot be evaluated analytically and is therefore calculated in a Monte Carlo fashion. In the independent error case, this can be achieved by appending the following steps to the sampling scheme in Section 3.2.

Forecasting Sampling Scheme (Independent Errors).

1. Generate an iterate from $p(\sigma^2|\boldsymbol{\gamma}, \mathbf{y})$.
2. Generate an iterate from $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma}, \mathbf{y})$.
3. Generate an iterate from $p(\tilde{\mathbf{y}}|\sigma^2, \boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, \mathbf{y}) = p(\tilde{\mathbf{y}}|\sigma^2, \boldsymbol{\beta}_\gamma)$.

After convergence of the sampling scheme in $\boldsymbol{\gamma}$, this will provide an iterate from the joint distribution $p(\tilde{\mathbf{y}}, \boldsymbol{\beta}, \sigma^2|\mathbf{y})$ so that the iterates $\tilde{\mathbf{y}}^{[j]}$ have density $p(\tilde{\mathbf{y}}|\mathbf{y})$. Calculation of the conditional distributions is outlined in Appendix A.2. Monte Carlo mixture estimates of the predictive marginal densities, $p(y_{n+i}|\mathbf{y})$, along with the marginal

mean, $E(y_{n+i}|\mathbf{y})$, are calculated as

$$\hat{p}(y_{n+i}|\mathbf{y}) = \frac{1}{J}\sum_{j=1}^{J} p(y_{n+i}|\boldsymbol{\beta}_{\gamma}^{[j]}, (\sigma^2)^{[j]}),$$

$$i = 1, \ldots, n_2 \quad (3.3)$$

and

$$\hat{y}_{n+i} = \frac{1}{J}\sum_{j=1}^{J} E(y_{n+i}|\boldsymbol{\beta}_{\gamma}^{[j]}, (\sigma^2)^{[j]}), \quad i = 1, \ldots, n_2, \quad (3.4)$$

where the conditional density at (3.3) and mean at (3.4) are as calculated in Appendix A.2.

To construct forecasts when the errors are autocorrelated, as in (A2) and (A3), the sampling scheme for dependent errors outlined by Smith et al. (1998) is augmented with the following steps.

Forecasting Sampling Scheme (Dependent Errors).

1. Generate an iterate from $p(\sigma^2|\boldsymbol{\gamma}, \boldsymbol{\beta}_{\gamma}, \boldsymbol{\kappa}, \boldsymbol{\psi}, \mathbf{y})$.
2. Generate an iterate from $p(\tilde{\mathbf{y}}|\sigma^2, \boldsymbol{\beta}_{\gamma}, \boldsymbol{\gamma}, \boldsymbol{\kappa}, \boldsymbol{\psi}, \mathbf{y}) = p(\tilde{\mathbf{y}}|\sigma^2, \boldsymbol{\beta}_{\gamma}, \boldsymbol{\psi}, \mathbf{y})$.

The calculation of these conditional densities is outlined in Appendix A.3. After the sampling scheme in the model parameters converges, this will provide iterates from $p(\sigma^2, \tilde{\mathbf{y}}|\mathbf{y})$ so that iterates $\tilde{\mathbf{y}}^{[j]}$ have density $p(\tilde{\mathbf{y}}|\mathbf{y})$. A histogram estimate of each predictive mean $E(y_{n+t}|\mathbf{y})$ is $\hat{y}_{n+t} = 1/J\sum_{j=1}^{J} y_{n+t}^{[j]}$, and a mixture estimate of the predictive marginal density $p(y_{n+t}|\mathbf{y})$ is provided by

$$\hat{p}(y_{n+t}|\mathbf{y}) = \frac{1}{J}\sum_{j=1}^{J} p(y_{n+t}|\boldsymbol{\beta}_{\gamma}^{[j]}, \boldsymbol{\psi}^{[j]}, (\sigma^2)^{[j]}, \tilde{\mathbf{y}}_{-t}^{[j]}, \mathbf{y}),$$

where $\tilde{\mathbf{y}}_{-t} = (y_{n+1}, \ldots, y_{n+t-1}, y_{n+t+1}, \ldots, y_{n+n_2})'$.

Under either independent or dependent error assumptions, $100(1-\alpha)\%$ prediction intervals for each forecasted value $\hat{y}_{n+i}$ can be calculated by ranking the Monte Carlo iterates $\{y_{n+i}^{[1]}, \ldots, y_{n+i}^{[J]}\}$ and counting off $(\alpha J)/100$ of the highest and lowest iterates.

4. EMPIRICAL ESTIMATES

The three-week window of data from 12:30, May 27, to 12:00, June 17, 1994, is fit to study the empirical performance of the methodology and models. From this period

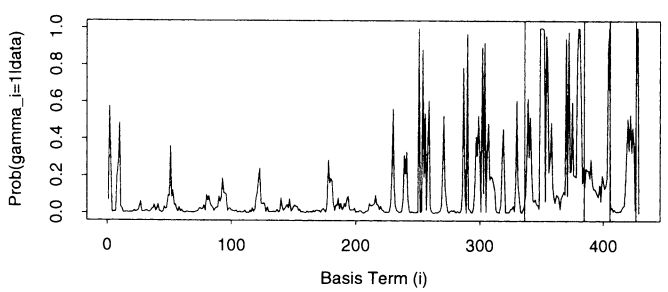

Figure 3. Marginal Posterior Probabilities That Each Regression Coefficient Is Nonzero for Model 1 With Independent Errors and a Basis That Is Not Augmented With Historical Estimates. Basis terms 1–336, 337–384, 385–405, 406–426, 427, 428, and 429 correspond to P_{wk}, P_{dy}, f_1, f_2, α_1, α_2, and δ, respectively.

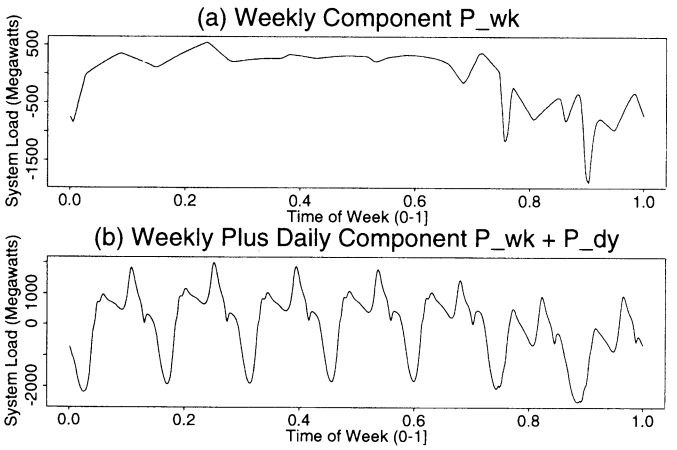

Figure 4. Estimated Periodic Profiles From the Three-Week Window of Data 12:30, May 27–12:00, June 17, 1994. Panel (a) is the intraweekly periodic component $\hat{P}_{wk}$ and panel (b) is the sum of the daily and weekly periodic components $\hat{P}_{wk} + \hat{P}_{dy}$, respectively.

$n_2 = 336$ forecasted half-hourly load values are calculated and compared to the true values of system load for 12:30, June 17, to 12:00, June 24. This process of forecasting at least 36 hours forward at 12:30 mimics that required by bidders in the wholesale market.

Model 1 with an assumption of independent errors (A1) is fit to this data without the augmented historical basis terms. Figure 3 plots the histogram estimate of the marginal posterior probability that each regression coefficient is nonzero, which is calculated as $\hat{p}\ (\gamma_i = 1|\mathbf{y}) = 1/J\sum_{j=1}^{J}\gamma_i^{[j]}$. There are 103 terms with $.1 < \hat{p}\ (\gamma_i = 1|\mathbf{y}) < .9$ and only 15 terms with $\hat{p}\ (\gamma_i = 1|\mathbf{y}) > .9$. This is because the mass of the posterior density $p(\boldsymbol{\gamma}|\mathbf{y})$ is spread over a wide range of possible subsets, so that there is no single high probability subset of basis terms. In such a situation, estimating a regression based on a single set of basis terms is not likely to be as effective as using estimates that average over the distribution of $\boldsymbol{\gamma}$. Bayesian posterior mean estimates provide a convenient framework for such model average estimates to be calculated in a methodologically coherent fashion.

The histogram estimate $\hat{\boldsymbol{\omega}} = 1/J\sum_{j=1}^{J}\boldsymbol{\omega}^{[j]} = (0,\ 1,\ 1,\ .369,\ .076,\ .240,\ .156)'$, indicating that basis terms corresponding to different functions have different probabilities of having nonzero coefficients. In particular, only 7.6% of the 336 basis terms for the weekly periodic component are likely to have nonzero coefficients and 36.9% of the 48 basis terms for the daily periodic component. In such situations, the flat prior for $\boldsymbol{\gamma}$ given by Smith and Kohn (1996), where $\omega_1 = \omega_2 = \cdots = \omega_7 = 1/2$, is not appropriate.

Model 1 is reestimated, but this time using a basis that is augmented with historical terms as outlined in Section 2.3. The result is that a much more parsimonious representation is now possible, with $\hat{\omega} = (.033, 1, 1, .021, .0032, .256, .158)'$. Figures 4(a) and 5(a) plot the resulting estimates of the weekly and daily periodic components. Although intrinsic daily or weekly effects cannot be separated using this model, a basic diurnal pattern appears to be captured by the daily periodic component, and only the sharp individual day deviations (mainly at the weekend) from this norm are captured by the estimated

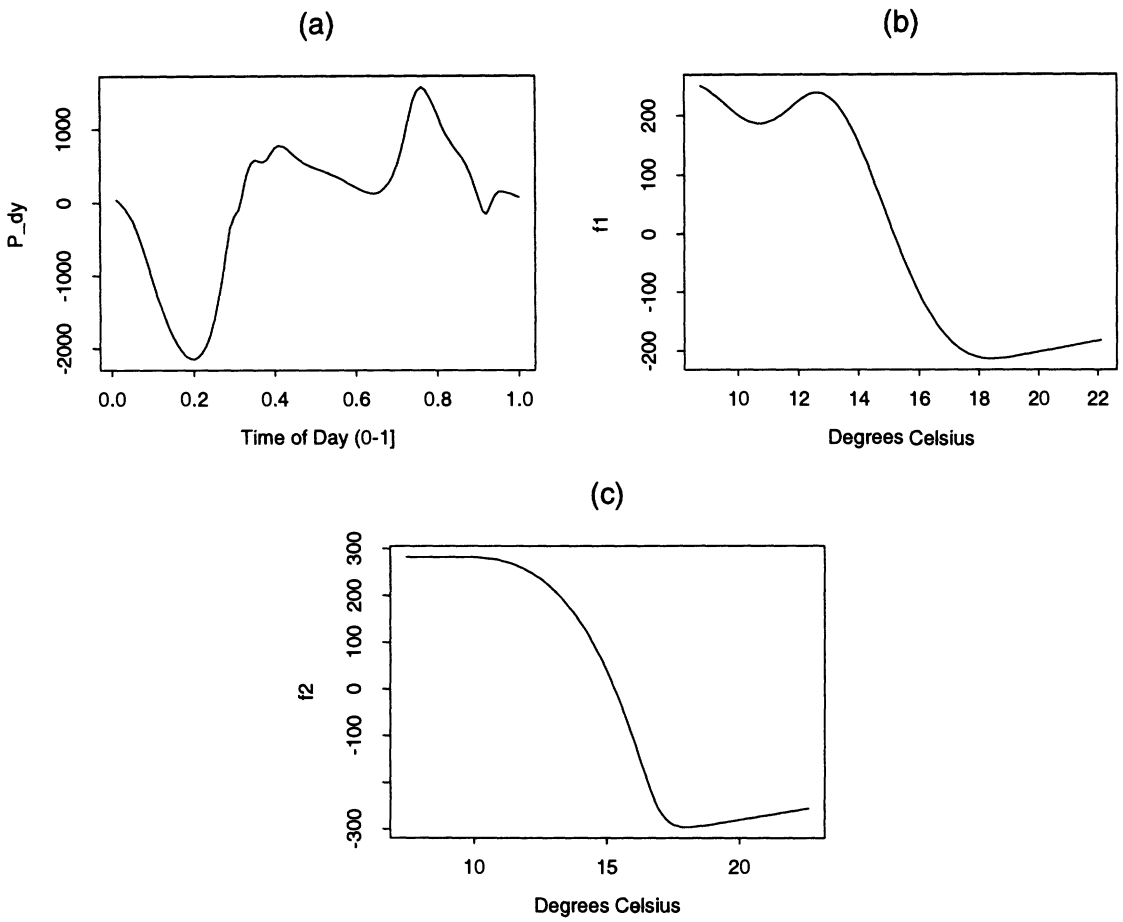

Figure 5. Estimates of P_{dy}, f_1, and f_2 From Model 1 With Independent Errors Fitted to the Three-Week Period 12:30, May 27–12:00, June 17, 1994.

weekly periodic component. Figure 4(b) plots the estimate of the total intraweekly periodic component $(P_{\text{wk}} + P_{\text{dy}})$, which is an estimate of an early winter periodic profile similar to that apparent in Figure 1(c).

The autocorrelation function of the residuals from this independent error model suggests that they are autocorrelated. There are two potential causes for this autocorrelation. First, a wide variety of variables that affect load are not incorporated into the load model, including special events, television programming, and other weather variables. Second, temperature is unlikely to affect load in a strictly additive manner so that Model 1 is likely to be somewhat misspecified.

To account for autocorrelation, Model 1 is fit again but with the errors assumed to follow the flexible autoregressive process outlined in (A2). Autocorrelation functions of the resulting residuals, ê, suggest that they are uncorrelated and that the in-sample autocorrelation has been captured effectively. Table 2 provides estimates of the autoregressive parameters $(\hat{\phi})$, the partial autocorrelations $(\hat{\psi})$, and posterior probabilities that each partial is nonzero $(\hat{\kappa})$. The estimate $\hat{\kappa}$ identifies that partial autocorrelations of lag greater than three are highly likely to be zero so that an appropriate lag length for the autoregression is three. The estimate for $\psi_1 = .988$ indicates that the error series may be better modeled in-sample with an autoregressive process in the first difference. Table 2 also contains estimates of ϕ, ψ, and κ when the errors follow the process outlined in (A3), which has estimated lag length two. Rerunning the dependent error regressions at different seasons suggests that these lag lengths are appropriate throughout the year.

Estimates of the temperature-sensitive components, f_1 and f_2, for the independent error model are given in Figure 5, (b) and (c), respectively. Care must be taken in interpreting these functions because temperature is highly related to

Table 2. Estimated Parameter Values for the Flexible Autoregressive Error (A2) and First-Differenced Error (A3) Models

	Error distribution A2			Error distribution A3		
Lag i	$\hat{\phi}_i$	$\hat{\psi}_i$	$\hat{\kappa}_i$	$\hat{\phi}_i$	$\hat{\psi}_i$	$\hat{\kappa}_i$
1	**1.406**	**.988**	**1.000**	**.414**	**.498**	**1.000**
2	**−.241**	**−.523**	**1.000**	**.171**	**.167**	**1.000**
3	**−.183**	**−.186**	**.999**	.009	−.002	.159
4	−.006	.000	.097	−.007	−.018	.256
5	−.004	.006	.100	−.023	−.026	.304
6	.010	.009	.105	.001	−.003	.073
7	−.013	.001	.023	−.008	−.009	.129
8	.010	.010	.111	−.001	−.002	.059
9	.001	.001	.023	.001	.000	.045
10	−.002	.000	.013	−.001	−.001	.041
11	.000	.001	.021	−.002	−.002	.039
12	.001	.001	.020	.000	.000	.036

NOTE: These values include the estimates of the autoregressive parameters $\hat{\phi}_i$, partial autocorrelations $\hat{\psi}_i$, and probabilities that each partial is nonzero $\hat{\kappa}_i$.

TOD and some of the variation in load captured by P_{dy} is due to temperature variation. June is early winter in Australia, and the temperatures are relatively low. The range 12°C–18°C appears to be the range of greatest response, with decreases in the temperature resulting in higher load due to increased heating. However, increases in temperature beyond 18°C result in an increase in load. On working days this could be caused by preprogrammed office air conditioning, while the increase on nonworking days is more slight. Both function estimates have minimums close to 18.3°C, as much of the energy literature would suggest, although the response of load to temperature does appear to differ on working and nonworking days. Both function estimates are different from the parametric form suggested by Doulai and Manoharan (1995) and do not have the simple V shape uncovered using monthly data by Engle et al. (1986).

The Bayesian framework also allows for the calculation of the posterior probability that a particular function, f say, is "null" with $f = 0$. This occurs when the corresponding indicator variables γ_f have a high posterior probability of being simultaneously zero; that is, $p(\gamma_f = \mathbf{0}|\mathbf{y})$. This is estimated using the histogram estimate $\hat{p}(f = 0|\mathbf{y}) = 1/J \sum_{j=1}^{J} \mathcal{I}\,(\gamma_f^{[j]} = \mathbf{0})$, where $\mathcal{I}(B)$ is an indicator function equal to one if B is true, zero otherwise. The posterior probability that each of the four nonlinear components of Model 1 are null were calculated for all three error distributions. All components were found to have zero posterior probability of being null, except for the temperature-sensitive components in the dependent error models. Here, $\hat{p}(f_1 = 0|\mathbf{y}) = .973$ and $\hat{p}(f_2 = 0|\mathbf{y}) = .96$ for error assumption (A2), while $\hat{p}(f_1 = 0|\mathbf{y}) = .962$ and $\hat{p}(f_2 = 0|\mathbf{y}) = .97$ for error assumption (A3). Therefore, in the dependent error models, variation in load caused by temperature (when controlling for TOD) is captured by the error processes at (A2) and (A3).

Overall, the dependent error model (A3) appears to provide the best in-sample fit. This does not, however, ensure that it is an effective model for forecasting load. First, although estimates from three week windows of load data may suggest a nonstationary component to system load, this does not ensure that such a component exists, in which

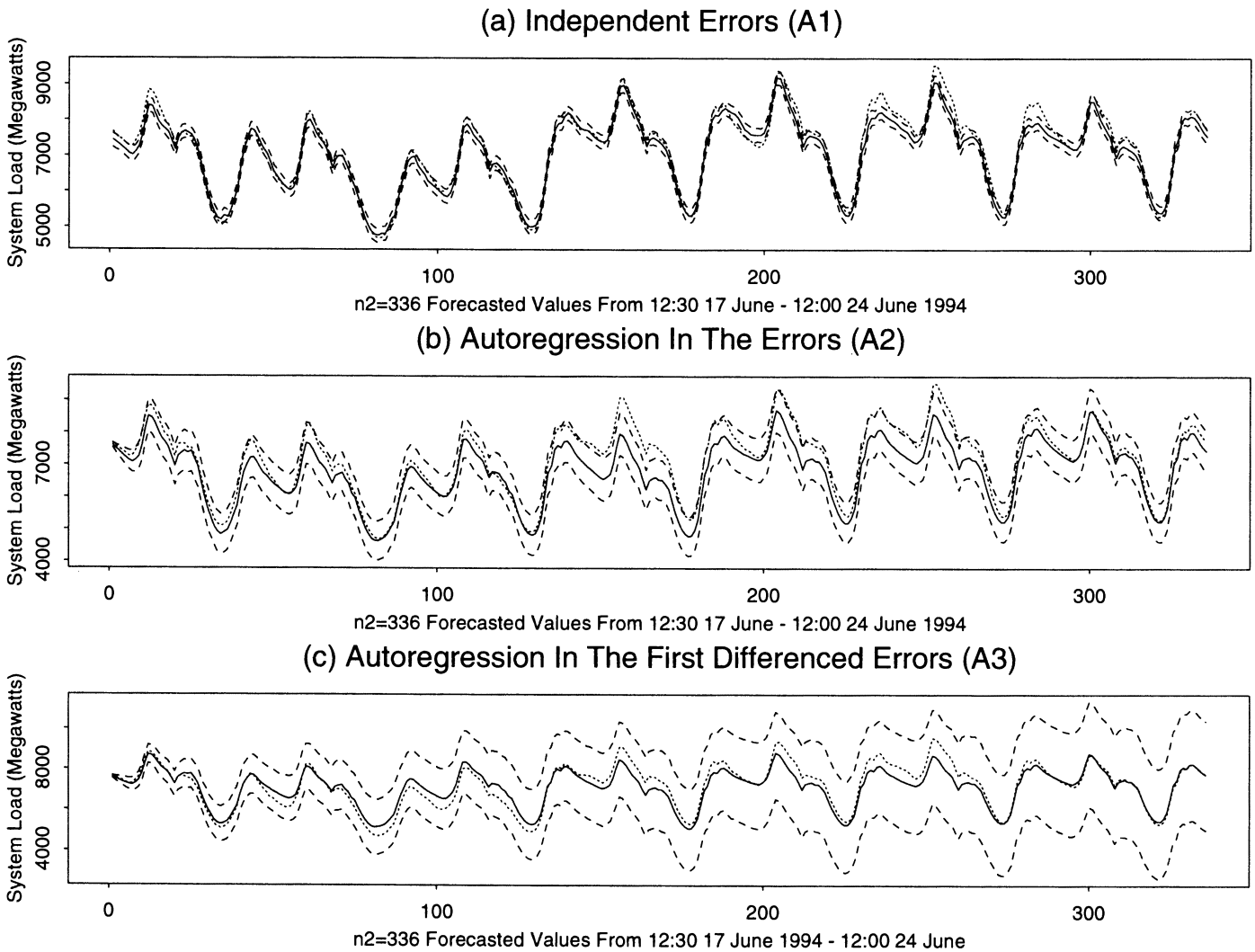

Figure 6. Forecasted Values of Half-Hourly System Load for the n_2 = 336 Periods during 12:30, June 17, 1994, to 12:00, June 24, 1994. The bold line is the forecasted mean $\hat{L}_{n+t}$, the long dashed line is the 90% upper and lower prediction intervals, and the dotted line is actual observed system load for the week. Panel (a) is for the independent error model, panel (b) is for the autoregressive error model, and panel (c) is for the first-differenced error autoregression.

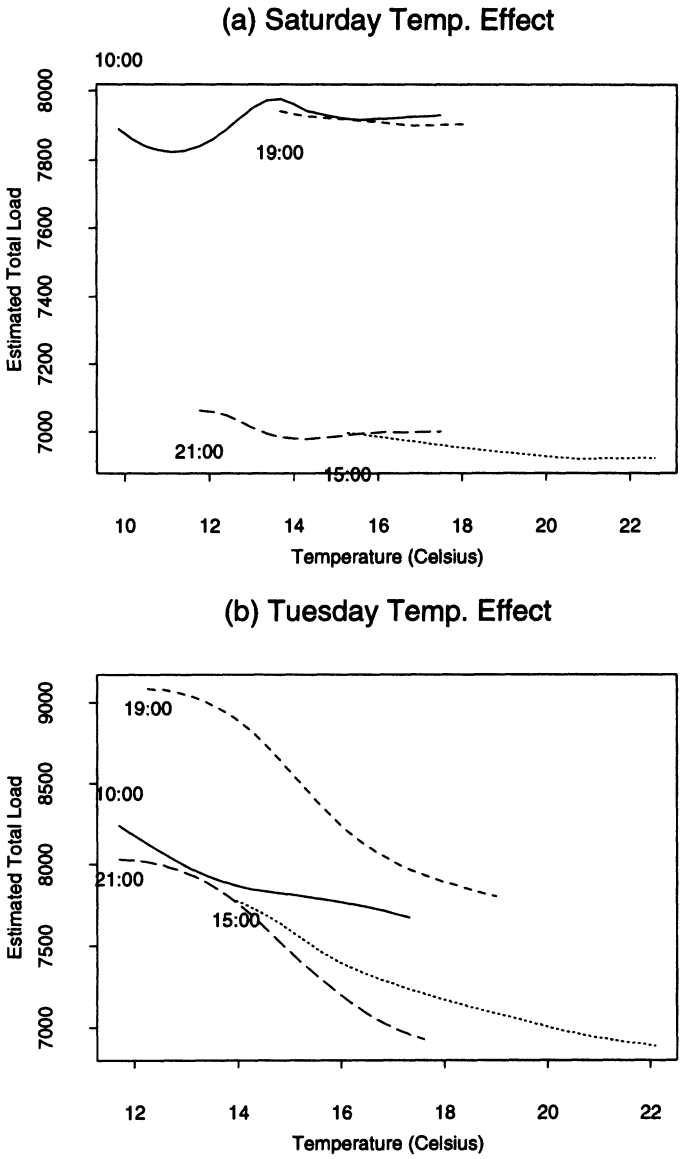

Figure 7. Estimated Response in Total Load to Changes in Temperature at Different Times of the Day for Model 2 With Independent Errors. Panel (b) is for a working day (Tuesday) and panel (a) is for a nonworking day (Saturday).

case unrealistically wide forecast intervals will result. Second, both dependent error models incorporate the temperature effect into the time series process, inhibiting the use of external temperature forecasts. To demonstrate these two points, Figure 6 plots the forecasted values $\hat{L}_{n+t}$ for model 1 with the three different error assumptions $n_2 = 336$ periods into the future. The estimated 90% prediction intervals are also plotted, along with the actual system load L_{n+t} for the week. The wide prediction intervals in panel (c) are because prediction intervals for a first-differenced autoregressive process increase at an approximately linear rate with respect to the forecast horizon (Box, Jenkins, and Reinsel 1994, chap. 5). A more realistic prediction interval is provided by the stationary autoregressive error model in panel (b). Moreover, accurate temperature forecasts can improve load forecasts dramatically. For example, panel (a) features the forecasts for the independent error model with temperature components, as estimated in Figure 5, (b) and (c), and the true temperature assumed known. [Forecasts from models incorporating weather effects are only as good as the weather forecasts employed. Some authors develop their own weather forecasts; for example, Wong (1994) and Doulai and Manoharan (1995) forecast temperature using time series models at a particular location. Accurate weather forecasting is a complex spatial problem, however, and to assess a load forecasting procedure it is useful to isolate it from the issues surrounding weather forecasting.]

Model 2 was also fit to this data with independent errors and the bivariate interaction surfaces f_3 and f_4 estimated. Figure 7 provides the results in "surface slice" form, plotting the estimated reaction in total system load to temperature changes at the high load periods 10:00, 15:00, 19:00, and 21:00 for a working day [panel (b)] and a nonworking day [panel (a)]. The estimated temperature-sensitive component of load varies at different times of the day, and the estimates for working days and nonworking days differ.

Temperature is not random over even quite lengthy periods of time, however, which limits the ability to estimate a temperature-sensitive component using three-week windows of data. First, the in-sample design of temperature data is poor, which is particularly troublesome in the estimation of the bivariate functions f_3 and f_4, which suffer from boundary value effects to a greater extent than estimates of univariate functions. Second, forecast or actual temperatures during the forecast period can fall outside the range of those experienced in-sample. Again, this is more

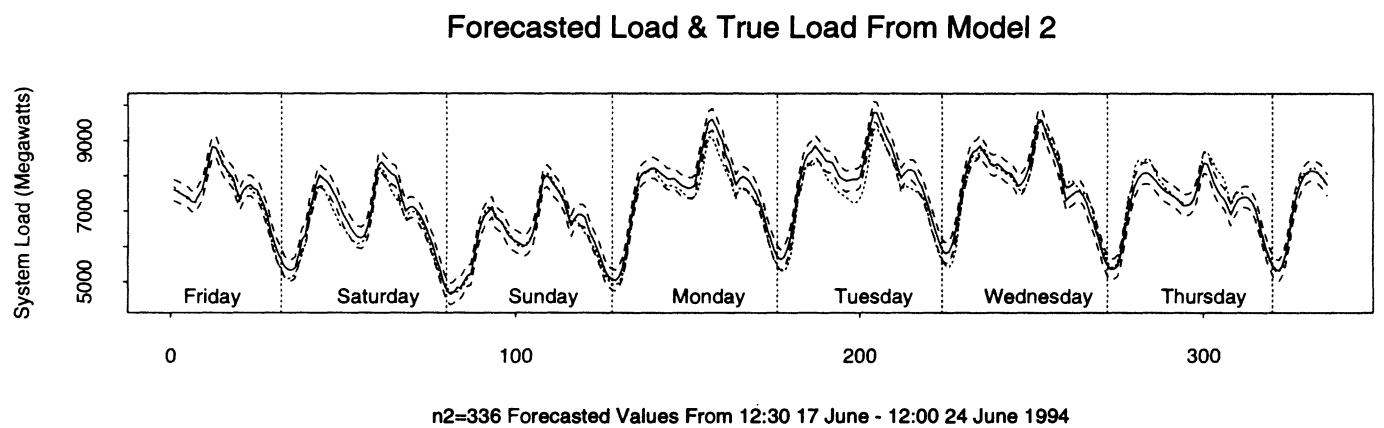

Figure 8. Forecast Results for 12:30, June 17, 1994–12:00, June 24, 1994, Using Model 2 With Independent Errors. The forecasted load is given by the bold line, actual load by the dotted line, and 90% confidence intervals by the dashed lines.

severe in the bivariate case because there is a different temperature range for each of the intraday periods. The problem is partially addressed here by setting forecast temperatures outside the observed range to the in-sample maximum or minimum.

Figure 8 plots the actual values for system load and forecasted values using Model 2. These forecasts, along with those from Model 1 found in Figure 6(a), highlight that intraday load can be forecast reliably for even quite long forecast horizons when accurate temperature forecasts are exploited in the load model. The prediction intervals in Figure 8 are slightly wider than those in Figure 6(a) because of the higher posterior variance of the bivariate temperature-sensitive functions.

Comparing the effectiveness of load forecasting procedures current in the literature is difficult. This is because authors use different forecast horizons, examine data arising from different localities and, for methods that incorporate temperature inputs, use temperature forecasts with an unknown degree of accuracy. For example, the articles discussed in the introduction provide forecasts for a variety of horizons between one hour and one day ahead, but for structuring initial bids in the NSW wholesale market, forecasts at least 36 hours ahead are required. Nevertheless, to demonstrate the potential of the method, forecasts were produced at 12:30 on each of the 100 days between June 24 and October 2, 1994, for a horizon of $n_2 = 336$ observations. This includes the high load season for NSW, during which the wholesale price of electricity is at its annual maximum. Model 1 was used with an assumption of independent errors that, when coupled with known temperature values, provided the most reliable forecasts of the alternatives investigated here. Table 3 provides the mean, median, minimum, and maximum average percentage error (APE) for each day type. As found in previous intraday load studies, the highest APE values occur on Sundays and public holidays. One reason for this is that, although public holidays have similar load profiles to Sundays, the assumption that they are exactly the same is incorrect. A separate load model for each public holiday, such as Christmas and Easter, based on load profiles exhibited in previous years is likely to be more accurate.

5. CONCLUSION

This article has proposed a load forecasting procedure to meet the needs of participants in the NSW wholesale electricity market. The approach employs recently developed Bayesian methods to estimate semiparametric regression models for intraday load. It provides several significant improvements over previous semiparametric regression intraday load models. These include model average estimates, incorporation of historical information on load profiles, and the flexibility to investigate a variety of models for both the mean and error structures. Empirical analysis of NSW data leads to several conclusions. First, including a nonstationary component of load can produce unrealistically poor forecasts for horizons typically required in the NSW wholesale market. Second, there is strong evidence for serial dependence, but it is difficult to separate in-sample from a temperature-sensitive component. Third, models that incorporate temperature components have potential to produce substantially more accurate forecasts for quite long forecast horizons. Last, although temperature does appear to affect load as an interaction with the time of day, identification of a general bivariate surface from short windows of data is difficult because of poor design and boundary value effects.

Table 3. Summary Statistics of the APE's From Forecasts Made at 12:30 Each Day Between June 24 and October 2, 1994, for a Forecast Horizon of n_2 = 336 Half-Hourly Observations

	Day type							
APE	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.	Pub. hol.
Mean	3.15	2.80	2.44	2.60	2.63	3.28	4.04	9.85
Median	2.75	2.33	2.11	2.14	2.20	2.94	3.27	9.54
Minimum	.00	.00	.00	.00	.00	.00	.00	.09
Maximum	23.51	13.95	11.76	11.34	11.82	14.70	19.92	27.51

ACKNOWLEDGMENTS

I thank two anonymous referees, Denzil Fiebig, Eva Knight, Rodney Strachan, Chi-Ming Wong, and, in particular, Robert Kohn for helpful comments. This research was partially supported by an Australian Research Council grant and a faculty grant while I was at the Department of Econometrics and Business Statistics, Monash University. I thank Robert Hyndman and Maxwell King at Monash University for ready access to computing facilities and staff at Pacific Power for providing the load and temperature data.

APPENDIX A: CALCULATIONS REQUIRED TO IMPLEMENT THE SAMPLING SCHEMES IN SECTION 3

A.1 Focused Sampling Scheme

To implement the focused sampling step in Section 3.2, the likelihood $l(\gamma_i)$ requires calculation (up to a constant), as well as the conditional prior $\pi(\gamma_i)$:

$$\begin{aligned} l(\gamma_i) &= p(\mathbf{y}|\boldsymbol{\gamma}) \\ &= \int\int p(\mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\beta}_\gamma, \sigma^2) p(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma}, \sigma^2) p(\sigma^2)\, d\boldsymbol{\beta}_\gamma\, d\sigma^2 \\ &\propto (n+1)^{-q_\gamma/2} S(\boldsymbol{\gamma})^{-n/2}, \end{aligned}$$

where $q_\gamma = \sum_{i=1}^{p} \gamma_i$ and $S(\boldsymbol{\gamma}) = \mathbf{y}'\mathbf{y} - \mathbf{y}' X_\gamma (X_\gamma' X_\gamma)^{-1} X_\gamma' \mathbf{y}$. If γ_i is associated with the jth function, then integrating out the hyperprior ω_j,

$$\begin{aligned} \pi(\gamma_i) &\propto p(\boldsymbol{\gamma}) \\ &= \int p(\boldsymbol{\gamma}|\boldsymbol{\omega}) p(\boldsymbol{\omega})\, d\boldsymbol{\omega} \\ &\propto \int_0^1 \omega_j^{q_j} (1-\omega_j)^{(p_j - q_j)}\, d\omega_j \\ &= \text{beta}(q_j + 1, p_j - q_j + 1). \end{aligned} \tag{A.1}$$

Here, p_j is the number of basis terms associated with the jth function, q_j is the number of these that are active (i.e., have corresponding binary indicator equal to 1), and beta$(\cdot,\cdot)$ is the beta function. The exact density π can therefore be calculated by evaluating (A.1) for $\gamma_i = 1$ and $\gamma_i = 0$ and normalizing.

A.2 Independent Errors

To implement the Monte Carlo estimate of the predictive density for independent errors in Section 3.4, the following conditional densities have to be calculated. The density

$$p(\sigma^2|\boldsymbol{\gamma},\mathbf{y}) \propto \int p(\mathbf{y}|\boldsymbol{\gamma},\sigma^2,\boldsymbol{\beta}_\gamma)p(\boldsymbol{\beta}_\gamma|\sigma^2,\boldsymbol{\gamma})p(\sigma^2)\,d\boldsymbol{\beta}_\gamma$$

$$\propto (\sigma^2)^{-(n/2+1)} \exp\left\{\frac{-S(\boldsymbol{\gamma})}{2\sigma^2}\right\}$$

so that $\sigma^2|\boldsymbol{\gamma},\mathbf{y} \sim$ inverse gamma $(n/2, S(\boldsymbol{\gamma})/2)$. The density

$$p(\boldsymbol{\beta}_\gamma|\sigma^2,\boldsymbol{\gamma},\mathbf{y}) \propto \int p(\mathbf{y}|\boldsymbol{\gamma},\sigma^2,\boldsymbol{\beta}_\gamma)p(\boldsymbol{\beta}_\gamma|\sigma^2,\boldsymbol{\gamma})p(\sigma^2)\,d\sigma^2$$

$$\propto \exp\left\{-\frac{n+1}{2n\sigma^2}\,(\boldsymbol{\beta}_\gamma-\boldsymbol{\mu}_\gamma)'X_\gamma'X_\gamma(\boldsymbol{\beta}_\gamma-\boldsymbol{\mu}_\gamma)\right\}$$

so that $\boldsymbol{\beta}_\gamma|\sigma^2,\boldsymbol{\gamma},\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}_\gamma,(\sigma^2 n)/(n+1)(X_\gamma'X_\gamma)^{-1})$, where $\boldsymbol{\mu}_\gamma$ is as defined in Section 3.2. The predictive density $p(\tilde{\mathbf{y}}|\boldsymbol{\beta}_\gamma,\sigma^2) = \prod_{i=1}^{n_2} p(y_{n+i}|\boldsymbol{\beta}_\gamma,\sigma^2)$, where $y_{n+i}|\boldsymbol{\beta}_\gamma,\sigma^2 \sim \mathrm{N}(\mathbf{x}_{\gamma,n+i}'\boldsymbol{\beta}_\gamma,\sigma^2)$ and $\mathbf{x}_{\gamma,n+i}$ is the independent variable vector corresponding to subset $\boldsymbol{\gamma}$ at the $n+i$ future period.

A.3 Dependent Errors

To obtain the Monte Carlo estimate of the predictive density for the dependent error case outlined in Section 3.4, the following conditional posterior densities have to be calculated. The density

$$p(\sigma^2|\boldsymbol{\gamma},\boldsymbol{\beta}_\gamma,\boldsymbol{\kappa},\boldsymbol{\psi},\mathbf{y})$$

$$\propto p(\mathbf{y}|\boldsymbol{\gamma},\sigma^2,\boldsymbol{\beta}_\gamma,\boldsymbol{\psi},\boldsymbol{\kappa})p(\boldsymbol{\beta}_\gamma|\sigma^2,\boldsymbol{\gamma},\boldsymbol{\psi})p(\sigma^2)$$

$$\propto (\sigma^2)^{-(n+q_\gamma)/2-1} \exp\left\{\frac{-1}{2\sigma^2}\,[S(\boldsymbol{\gamma},\boldsymbol{\psi})\right.$$

$$\left.+\frac{(n+1)}{n}\,(\boldsymbol{\beta}_\gamma-\boldsymbol{\mu}_{\gamma,\psi})'X_\gamma'\Omega_\psi^{-1}X_\gamma(\boldsymbol{\beta}_\gamma-\boldsymbol{\mu}_{\gamma,\psi})\right]\Big\},$$

where $S(\boldsymbol{\gamma},\boldsymbol{\psi}) = \mathbf{y}'\Omega_\psi^{-1}\mathbf{y} - \mathbf{y}'\Omega_\psi^{-1}X_\gamma(X_\gamma'\Omega_\psi^{-1}X_\gamma)^{-1}X_\gamma'\Omega_\psi^{-1}\mathbf{y}$ and $\boldsymbol{\mu}_{\gamma,\psi}$ is as defined in Section 3.3. This is an inverse gamma density in σ^2. In generating from this density, it is important to note that the form of Ω_ψ (and its known Cholesky factor $\Omega_\psi^{-1/2}$) differs depending on whether the autoregression is in the errors u_t or the first-differenced errors $v_t = u_t - u_{t-1}$. The conditional predictive density $p(\tilde{\mathbf{y}}|\sigma^2,\boldsymbol{\beta}_\gamma,\boldsymbol{\psi},\mathbf{y}) = \prod_{t=n+1}^{n+n_2} p(y_t|\sigma^2,\boldsymbol{\beta}_\gamma,\boldsymbol{\psi},y_{t-1},\ldots,y_{t-r-1})$. Generation from this density can therefore be achieved by generating sequentially from the conditional distributions $y_t|\sigma^2$, $\boldsymbol{\beta}_\gamma$, $\boldsymbol{\psi}$, $y_{t-1},\ldots,y_{t-r-1} \sim \mathrm{N}(\theta_t,\sigma^2)$, for $t = 1,\ldots,n_2$. Here, when the autoregression is in u_t [as in (A2)], then

$$\theta_t = \mathbf{x}_{\gamma,t}'\boldsymbol{\beta}_\gamma + \sum_{j=1}^{r}\phi_j u_{t-j},$$

but with an autoregression in the first difference $u_t - u_{t-1}$ [as in (A3)],

$$\theta_t = \mathbf{x}_{\gamma,t}'\boldsymbol{\beta}_\gamma + u_{t-1} + \sum_{j=1}^{r}\phi_j(u_{t-j} - u_{t-j-1}).$$

APPENDIX B: INVARIANT DISTRIBUTION OF THE FOCUSED SAMPLER

This appendix provides a proof that the invariant distribution of the focused sampler outlined in Section 3.2 is the posterior distribution $\pi^*(\gamma_i)$ defined at Equation (3.2). Because γ_i is a binary variable, all that has to be shown is that the transition kernel Q of the focused sampling step satisfies the detailed balance criterion. That is,

$$\pi^*\ (\gamma_i = 1)Q\ (\gamma_i = 1 \to \gamma_i = 0)$$

$$= \pi^*\ (\gamma_i = 0)Q\ (\gamma_i = 0 \to \gamma_i = 1). \quad \text{(B.1)}$$

Proof. The transition kernel of the focused sampler is given by

$$Q\ (\gamma_i = 1 \to \gamma_{i,j} = 0)$$

$$= \text{(prob. of case (2))} \times \text{(prob. of a zero in case (2))}$$

$$= \pi\ (\gamma_{i,j} = 0)\frac{l\ (\gamma_i = 0)}{l\ (\gamma_i = 1) + l\ (\gamma_i = 0)}$$

$$Q\ (\gamma_i = 0 \to \gamma_{i,j} = 1)$$

$$= \text{(prob. of case (4))} \times \text{(prob. of a one in case (4))}$$

$$= \pi\ (\gamma_{i,j} = 1)\frac{l\ (\gamma_i = 1)}{l\ (\gamma_i = 1) + l\ (\gamma_i = 0)}.$$

From (3.2), the posterior distribution can be written as

$$\pi^*\ (\gamma_i = 1)$$

$$= \frac{l\ (\gamma_i = 1)\pi\ (\gamma_i = 1)}{l\ (\gamma_i = 1)\pi\ (\gamma_i = 1) + l\ (\gamma_i = 0)\pi\ (\gamma_i = 0)}$$

$$\pi^*\ (\gamma_i = 0)$$

$$= \frac{l\ (\gamma_i = 0)\pi\ (\gamma_i = 0)}{l\ (\gamma_i = 1)\pi\ (\gamma_i = 1) + l\ (\gamma_i = 0)\pi\ (\gamma_i = 0)}.$$

Substituting the transition kernel of the focused sampler and the expansion for π^* just given into the detailed balance criterion at (B.1) shows that the criterion is satisfied. Therefore, π^* is the invariant distribution of the focused sampling step.

[Received August 1998. Revised December 1999.]

REFERENCES

Bartels, R., and Fiebig, D. (1998), "Residential End-use Electricity Demand: Results From a Designed Experiment," unpublished manuscript.

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

Doulai, P., and Manoharan, M. (1995), "State Estimation/Kalman's Algorithm: Short Term Load Forecasting," in *Proceedings of the Australasian Universities Power Engineering Conference*, Perth, pp. 175–180.

Engle, R., Granger, C., and Hallman, J. (1989), "Merging Short and Long Run Forecasts: An Application of Seasonal Cointegration to Monthly Electricity Sales Forecasting," *Journal of Econometrics*, 40, 45–62.

Engle, R., Granger, C., Rice, J., and Weiss, A. (1986), "Semiparametric Estimates of the Relation Between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310–320.

Fiebig, D., Bartels, R., and Aigner, D. (1991), "A Random Coefficient Approach to the Estimation of Residential End-use Load Profiles," *Journal of Econometrics*, 50, 297–327.

Friedman, J. (1991), "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 19, 1–67.

Harvey, A., and Koopman, S. (1993), "Forecasting Hourly Electricity Demand Using Time-Varying Splines," *Journal of the American Statistical Association*, 88, 1228–1253.

Ho, K., Hsu, Y., and Yang, C. (1992), "Short Term Load Forecasting Using a Multilayer Neural Network With an Adaptive Learning Algorithm," *IEEE Transactions on Power Systems*, 7, 141–149.

Hsu, Y., and Yang, C. (1995), "Electrical Load Forecasting," in *Applications of Neural Networks*, ed. A. Murray, Boston: Kluwer, pp. 157–189.

Hyde, O., and Hodnett, P. (1997), "An Adaptable Automated Procedure for Short Term Electricity Forecasting," *IEEE Transactions on Power Systems*, 12, 84–94.

Lee, K., Cha, Y., and Park, J. (1992), "Short-term Load Forecasting Using an Artificial Neural Network," *IEEE Transactions on Power Systems*, 7, 124–132.

Luo, Z., and Wahba, G. (1997), "Hybrid Adaptive Splines," *Journal of the American Statistical Association*, 92, 107–116.

Moghram, I., and Rahman, S. (1989), "Analysis and Evaluation of Five Short-term Load Forecasting Techniques," *IEEE Transactions on Power Systems*, 4, 1484–1491.

O'Hagan, A. (1995), "Fractional Bayes Factors for Model Comparison" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 99–138.

Peng, T., Hubele, N., and Karady, G. (1992), "Advancement in the Application of Neural Networks for Short-term Load Forecasting," *IEEE Transactions on Power Systems*, 7, 250–257.

Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Saha, T., and Senini, N. (1996), "Short Term Load Forecasting Using an Artificial Neural Network," in *Proceedings of the Australasian Universities Power Engineering Conference*, Melbourne (vol. 3), pp. 661–666.

Smith, M., and Kohn, R. (1996), "Nonparametric Regression via Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–344.

——— (1997), "A Bayesian Approach to Nonparametric Bivariate Regression," *Journal of the American Statistical Association*, 92, 1522–1535.

Smith, M., Wong, C., and Kohn, R. (1998), "Additive Nonparametric Regression With Autocorrelated Errors," *Journal of the Royal Statistical Society*, Ser. B, 2, 311–331.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: Society for Industrial and Applied Mathematics.

Wong, C. (1994), "A State Space Approach to Nonparametric Regression," unpublished Ph.D. thesis, University of New South Wales, the Australian Graduate School of Management.

Wong, F., Hansen, M., Kohn, R., and Smith, M. (1997), "Focused Sampling and its Application to Nonparametric Regression," unpublished manuscript.