Aust. N. Z. J. Stat. 2022 doi: 10.1111/j.1467-842X.XXX

Short-term forecasting with a computationally efficient nonparametric transfer function model

Jun. M. Liu[1*]

Georgia Southern University

Summary

In this paper a semi-parametric approach is developed to model nonlinear relationships in time series data using polynomial splines. Polynomial splines require very little assumption about the functional form of the underlying relationship, so they are very flexible and can be used to model highly nonlinear relationships. Polynomial splines are also computationally very efficient. The serial correlation in the data is accounted for by modeling the noise as an Autoregressive Integrated Moving Average (ARIMA) process, by doing so, the efficiency in nonparametric estimation is improved and correct inferences can be obtained. The explicit structure of the ARIMA model allows the correlation information to be used to improve forecasting performance. An algorithm is developed to automatically select and estimate the polynomial splines model and the ARIMA model through backfitting. This method is applied on a real-life data set to forecast hourly electricity usage. The nonlinear effect of temperature on hourly electricity usage is allowed to be different at different hours of the day and days of the week. The forecasting performance of the developed method is evaluated in post-sample forecasting and compared with several well-accepted models. The results show the performance of the proposed model is comparable with a long short-term memory deep learning model.

Key words: short-term electricity forecasting; nonparametric smoothing; backfitting; time series; ARIMA model

1. Introduction

Curiosity about the future is a constant interest of human kind, and predicting the future has been a continuous effort since the beginning of human history. Accurate forecast can be based on the understanding of the relationship among different variables. Due to its

*Author to whom correspondence should be addressed.
[1]Department of Enterprise Systems & Analytics, Parker College of Business, Georgia Southern University, P.O. Box 7998, Statesboro, GA 30460, USA
Email: jliu@georgissouthern.edu

Acknowledgment. The author gratefully acknowledges the valuable input, advice and feedback from the Associate Editor and two anonymous referees, which lead to significant improvement of the paper. The author is solely responsible for any remaining errors or omissions.

importance, a lot of research has been conducted in the area of quantitative forecasting and numerous methods are developed from the perspectives of cross-sectional data, time series data, and longitudinal data. In the area of time series forecasting, Box and Jenkins (1976) developed the linear transfer function models, which have been well studied and proven successful in many fields (e.g., Tiao & Box 1981, Poskitt 1989). However, in practice, we often encounter nonlinear relationships that cannot be well approximated by linear functions. Nonlinear parametric models are developed as a response to this challenge, successful examples include the *nonlinear transfer function model* (Chen & Tsay 1996, the *threshold models* (Tong 1983), the *functional-coefficient autoregressive model* (FAR) (Chen & Tsay 1993a, Cai, Fan & Yao 2000), the *nonlinear additive autoregressive model* (Chen & Tsay 1993b), the *adaptive functional-coefficient model* (Xia & Li 1999, Fan, Yao & Cai 2003), the *single index model* (Härdle, Hall & Ichimura 1993, Carroll et al. 1997, Newey & Stoker 1993, Wang & Yang 2009) and the *partially linear models* (Härdle, Liang & Gao 2000). Douc, Moulines & Stoffer (2014) provides an in-depth review of nonlinear time series modeling. Chen & Bunn (2014) uses a finite mixture regime-switching model to forecast daily electricity prices. The threshold models are extended to high-dimensional time series under a factor structure (Lam & Yao 2012, Liu & Chen 2016, Liu & Chen 2020).

For a nonlinear relationship, there are usually a very large number of different parametric functions that can be used to approximate it, therefore it is usually difficult to choose the 'correct' parametric function, and the model-selection process usually involves certain degree of subjectivity. Nonparametric smoothing methods 'let the data speak for themselves' to determine the appropriate functional form for the underlying relationship, therefore require very little assumption. As a result, the subjectivity in choosing the functional form is avoided. Nonparametric smoothing methods are flexible and suitable for modeling highly nonlinear time series, pioneering works in this area include Robinson (1983), Auestad & Tjöstheim (1990), Lewis & Stevens (1991), Masry (1996b and 1996a), Fan & Gilbels (1996), Smith, Wong & Kohn (1998), Hart & Vieu (1990), and many others. Fan & Yao (2003) provides an excellent review of the nonlinear and nonparametric models in time series analysis. Wang & Yang (2007) proposes the spline-backfitted kernel smoothing model which benefits from the computational efficiency of polynomial splines and the well-established local properties of the kernel smoother. Aneiros-Pérez & Vieu (2008) proposes a functional version of the partial linear model for nonparametric time series prediction. Liu, Cai & Chen (2015) use local linear estimators to estimate the trend and seasonal effect functions in the functional coefficient model. Tsay & Chen (2018) provides a comprehensive review of the commonly used nonlinear, nonparametric, and machine learning models for time series analysis.

Demand forecasting plays an important role in the operations of electricity power systems. Daily operational activities such as plant scheduling, load dispatching, security

assessment, reserve capacity allocation, purchasing/selling decisions, all rely heavily on short-term demand forecasts. Due to the shear volume of the business, even a small improvement in the forecasting performance can generate significant gains in the reliability, safety, and profitability of power companies. Accurate demand forecast improves the efficiency in the operation of power plants, which helps to conserve energy, reduce fossil fuel usage and the resulting carbon and pollutant emission, therefore has a long-lasting impact on building a clean and sustainable economy. A lot of work has been done in electricity usage forecasting and several earlier successful models can be found in Gupta (1985). Probabilistic forecasting of electricity has gained much interest in recent years, the performance of various models are evaluated in the Global Energy Forecast Competitions and the performance documented in e.g., Hong et al. (2016) and Hong, Xie & Black (2019). Hong & Fan (2016) provides a review on probabilistic load forecasting. Kezunovic et al. (2020) surveys big data analytics applications and associated implementation issues in future electricity grids.

Electricity usage is highly seasonal, and seasonality accounts for a large portion of the short-term variation. A seasonal ARIMA model can be used to model the seasonal variation, but an univariate ARIMA model by itself is not likely to work well. The seasonal variation can also be modelled by decomposing the usage data into periodic components, and including them in regression models. After the seasonal variation is removed, the remaining serial correlation in the partial residuals can be removed by using relatively simple models such as autoregressive models (Engle et al. 1986, Smith 2000). Another main source of short-term variation in electricity usage are the weather-related factors, such as temperature, humidity, wind speed, etc. Among these factors, temperature has the largest impact on electricity usage. Wang, Liu & Hong (2016) studies the 'recency effect' of temperature using a big data approach and finds that including lags of temperature in the model significantly improves the forecasting performance. Temperature effect on electricity usage is often nonlinear, especially in areas where electricity is used for both heating and cooling (Al-Zayer & Al-Ibrahim 1996). Engle et al. (1986) uses an asymmetric V-shaped function to approximate this nonlinear relationship, the tip of the V-shape is the transition point between the needs for heating and cooling. One widely adopted approach is to transform temperature to degree-days or degree-hours based on two transition points and use piecewise regression to model the temperature effect (Gupta 1985, Al-Zayer & Al-Ibrahim 1996). However, under the similar overall pattern, the temperature effect on electricity usage can still be quite different at different times of the day, in different days of the week, and between holidays and non-holidays As a result, some researchers model the temperature effect differently for different times of the day and types of the day (Engle et al. 1993, Ramanathan et al. 1997). However, with this approach, the difficulty of identifying nonlinear parametric functions is compounded by the large number of models that must be identified for different hour/daytype subsets.

Nonparametric smoothing methods can be used to avoid the difficulty and subjectivity in identifying (and justifying) parametric nonlinear models. Examples of nonparametric modeling of electricity usage include artificial neural network (ANN) (Ho, Hsu & Yang 1992, Peng, Hubele & Karady 1992), smoothing splines (Engle et al. 1986, Harvey & Koopman 1993), or methods based on more general basis functions (Smith 2000). Liu et al. (2006) models the temperature effect by each hour of non-holiday weekdays and weekends/holidays using Locally Weighted Scatterplot Smoothing (LOWESS) (Cleveland 1979). Xie et al. (2019) proposes a nonparametric Bayesian framework for short-term wind power forecast.

In recent years partly due to the great progress in computer technology and the availability of large amount of data, machining learning models have been developed to analyze complex, nonlinear relationships. They have been very successful in areas such as image processing, pattern recognition, and language processing. Machine learning models have been successful in load forecasting due to the nonlinear nature of the problem. The feed-forward neural networks (FNNs) are among the earlier applications of neural networks in load forecasting (Hippert, Pedreira & Souza 2001, Lee, Cha & Park 1992, Park et al. 1991, Drezga & Rahman 1999). Recurrent neural networks (RNN) are able to maintain the time order of the data therefore suitable for time series forecasting, for example, Elman (1990) proposed the Elman recurrent neural network (ERNN). However, RNNs including ERNNs suffer from vanishing or exploding gradients (Bengio, Simard & Frasconi 1994) and do not utilise information in long-term memories very well. RNN with long short-term memory (LSTM) approach is developed to overcome this problem by using a gated mechanism to control the flow of gradient (Hochreiter & Schmidhuber 1997). By controlling the gate values, part of the old state is preserved in the flow (in contrast, the old state is replaced at each time step in ERNN), therefore LSTM can maintain long-term memory. LSTM has been successfully applied in load forecasting (e.g., Kong et al. 2019, Bouktif et al. 2018). The gated recurrent unit (GRU) proposed by Cho et al. (2014) combines the 'forget' gate and the 'input' gate of the LSTM and offers a computationally more efficient alternative to LSTM. GRUs have seen some applications in load forecasting (Wang et al. 2018). Bianchi et al. (2017) provides a comparative review of RNNs (ERNN, LSTM, GRU, and Ecoh-state Network) in short-term load forecasting. Convolutional neural networks (CNNs) (Lecun et al. 1998) are originally developed for image processing but they also provide an efficient solution in time series forecasting (Borovykh, Bohte & Oosterlee 2017). CNNs have shown promising performance in load forecasting (Kuo & Huang 2018, Amarasinghe, Marino & Manic 2017). One of the recent developments of CNNs is the temporal convolutional network (TCN). TCNs are autoregressive in nature and are able to process arbitrary length of the input and output sequences by using causal convolutions and residual connections. TCNs are relatively new in load forecasting but the applications have begun to emerge and show promising results

(Dorado Rueda, Durán Suárez & del Real Torres 2021, Gasparin, Lukovic & Alippi 2022). A comparative review of deep learning models in short-term load forecasting can be found in Gasparin, Lukovic & Alippi (2022).

In this paper we develop an accurate and computationally efficient semi-parametric method for short-term electricity forecasting. The nonlinear relationship between load and temperature is modelled by each hour of the day and each (non-holiday) day of the week plus holidays, resulting in 192 subsets. In each subset, the temperature effect is modelled independently using polynomial splines, hence is allowed to be different at different hours of different days. By varying the smoothing parameters, polynomial splines can adapt to various different nonlinear relationships. Polynomial splines have an explicit functional form globally, so the estimated model can be easily interpreted and understood. Once the smoothing parameters are determined, the model can be estimated as a standard least squares regression globally, therefore it is computationally more efficient than local polynomial smoothers, which require estimation to be done at many focal points. In nonparametric smoothing with time series data, estimation efficiency can be improved significantly if the serial correlation is properly modelled and removed, in fact, it is shown that for local polynomial smoothers, the optimal rate of convergence of iid data can be achieved (Xiao et al. 2003, Su & Ullah 2006, Liu, Chen & Yao 2010). In this paper, we remove the serial correlation by modelling the innovation process as an Autoregressive Integrated Moving Average (ARIMA) process (Box & Jenkins 1976). A modified backfitting algorithm (Hastie & Tibshirani 1990) is used to estimate the model iteratively between the spline component and ARIMA component. Model selection is embedded in this iterative estimation procedure. The approach described in this paper can be applied in other applications sharing similar features, such as other utilities (natural gas, water, etc.), telecommunications (call volume, data usage, etc.). The idea considered in this paper is similar to Kohn, Ansley & Wong (1992) in which smoothing splines are used to model the transfer function the noise is assumed to be a stationary ARMA process, in this paper, polynomial splines are used and the noise is allowed to be a seasonal ARIMA process.

This paper is organized as follows. Section 2 describes the data. The motivation of using polynomial splines to model the temperature effect is given in Section 3, a brief description of polynomial splines, and the model selection procedure is also included in this section. Section 4 discusses the removal of serial correlation using an ARIMA model. The proposed two-component semi-parametric model is introduced in Section 5. The performance of the proposed model is summarised in Section 6, including a comparison of the forecasting performance with several other models. A summary is provided in Section 7.

2. The data

Electricity usage data are highly correlated, and the effect of temperature on electricity usage is usually highly nonlinear. The model considered in this paper is developed with such applications in mind. The data set used in this paper contains four years of hourly records of electricity load and several meteorological variables (temperature, humility, and wind speed) between 1998 and 2001 from a major electric company in the United States. The electricity usage is the total system load including industrial, residential and commercial usage, measured in megawatts (MW). The first three years of data are the same as used in Liu et al. (2006), one additional year of data (2001) are included in this study. The first three years of data (1998-2000) are used to build the model, and the last year of data (2001) are reserved for post-sample forecasting. The time series plot of of the hourly electricity usage in Figure 1 shows typical seasonal variation in hourly electricity usage. The temperature (°F) in the data set is the weighted average of the hourly recordings at four local weather stations in the service area, the weights reflect the electricity usages in the areas where the weather stations are located. Among the meteorological variables, temperature is the only one found significant. The hourly temperature also shows strong seasonality (Figure 1).

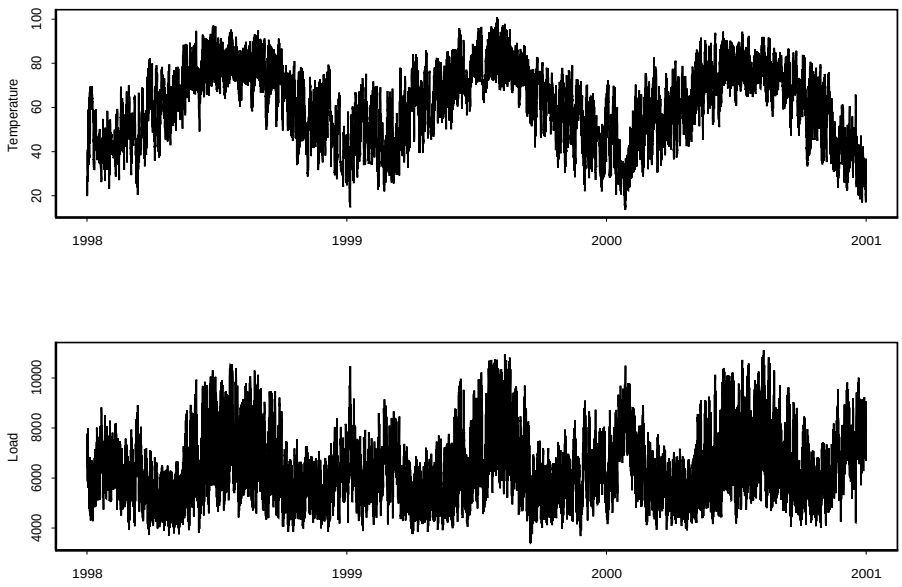

Figure 1. Hourly electricity usage and temperature.

Figure 2 is the scatter plot of hourly electricity usage versus temperature. In the service area, electricity is used for both heating in winter and cooling in summer, therefore

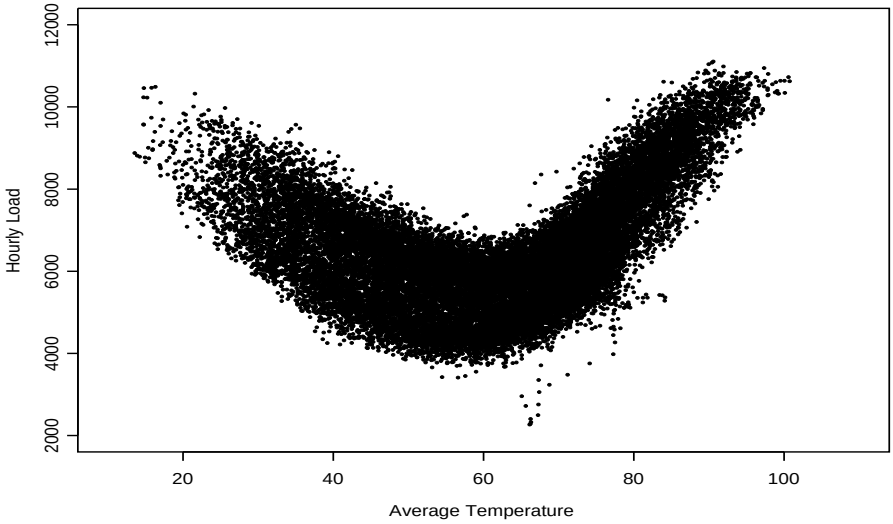

Figure 2. Usage vs temperature.

the relationship between temperature and electricity usage shows an asymmetric U-shaped pattern. This figure also shows large variation in electricity usage at the same temperature, which can be attributed to the usage difference in different hours and days. Figures 1 and 2 show a few observations with abnormally low electricity usage, as discussed in Liu et al. (2006), these outliers are the result of a hurricane in summer 1999, which caused large-scale service interruptions. These outliers are adjusted in a similar way as in Liu et al. (2006), and the outlier-adjusted data are used in the analyses.

3. Modelling the varying temperature effect

Electricity demand is driven by human behaviour, so hourly electricity usage usually displays periodic patterns. The periodic patterns can be modelled by seasonal ARIMA models. However, heating and cooling are among the major demands of electricity, so temperature explains a large portion of the variation in electricity demand, an ARIMA model without considering the temperature effect might not be adequate. Individual's response to temperature largely depends on his/her daily activities, therefore the temperature effect is expected to be different at different times of the day, in different days or the week, and between holidays and non-holidays. The difference in usage pattern between different hours and days itself may vary, which makes the temperature effect more complex. To see this, scatter plots of electricity usage versus temperature are made for each hour of the day, a few

are shown below in Figures 3 and 4, where non-holiday weekdays are coloured in green, weekends and holidays are coloured in red. In each figure, the mean curves for each day are estimated using LOWESS and superimposed to accentuate the different usage patterns. The scatter plots show that

- The usage levels are different at different times of the day. Usage in active hours are higher than in inactive hours on average, as shown in Figure 3, which shows the usage pattern in 8am and 8pm on the same scales.
- Difference exists in usage level between different days, which is evident from the different mean curves in Figures 3 and 4. For the same hour, usage tends to be higher in non-holiday weekdays than in weekends and holidays. This difference is more pronounced in active hours, as seen from the two-band pattern in Figure 3.
- The shape of some of the scatter plots are quite different, which indicates that the temperature-usage relationship is potentially of different functional forms in different hours and days of the week.
- The difference in the temperature-usage relationship between different days varies by hours of the day, and is generally more pronounced in active hours than in inactive hours. In general, Mondays Tuesdays, Wednesdays and Thursdays are similar, as their mean curves are very close. Sunday, Saturday and Holidays as a group are different from the non-holiday weekdays, but they are less similar to each other than are the weekdays, as their curves are usually more distant from each other. Friday pattern is interesting in that it is similar to Monday-Thursday in the early hours of the day, but after evening its pattern becomes closer to that of the weekends and holidays (Figure 3).
- The difference in the temperature-usage relationship between different hours varies by day of the week. Generally, during inactive hours, the hour-to-hour usage change in non-holiday weekdays is similar to that in weekends and holidays, but more different during active hours. For example, as shown in Figure 4, the increase of electricity usage from 5am to 6am is higher in non-holiday weekdays than in weekends and holidays.

The above observations indicate that the temperature-usage relationship may be of different functional forms in each hour and day of the week, therefore it seems necessary to subset the data by the hours of the day and days of the week with holidays as a separate group and study the temperature effect in each of the resulting 192 subsets. The holidays include New Year's Day, Presidents' Day, Martin Luther King Day, Good Friday, Memorial Day, Independence Day, Labor Day, Veteran's Day, Thanksgiving and Christmas. Hence we consider the following model

$$Y_{ij} = f_i(X_{ij}) + e_{ij} \ , \tag{1}$$

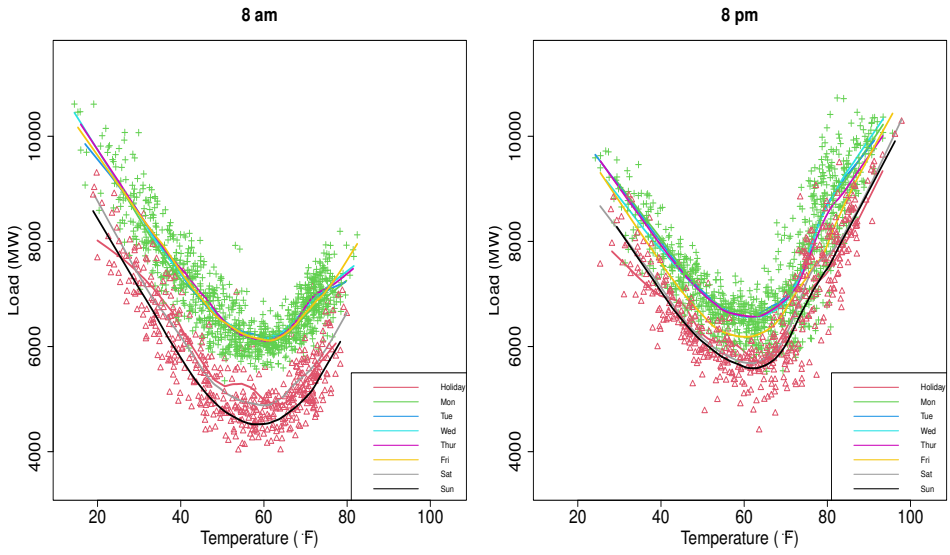

Figure 3. Usage pattern in 8am and 8pm.

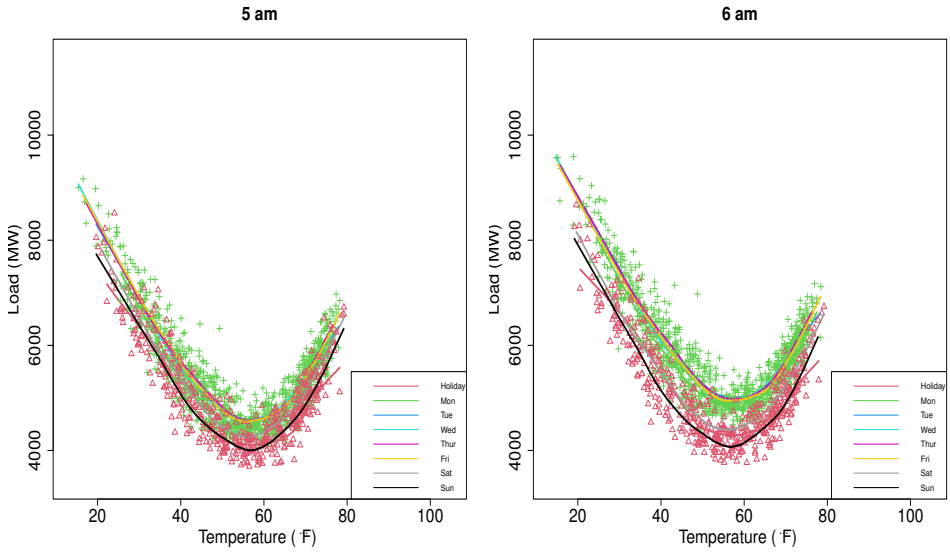

Figure 4. Usage pattern in 5am and 6am.

where $i = 1, 2, \cdots, 192$, $j = 1, \cdots, n_i$. Identifying and estimating so many potentially different nonlinear relationships can be challenging, as a solution, we use polynomial splines to model $f_i(\cdot)$ due to their data-driven nature and computational efficiency.

3.1. Polynomial splines

Polynomial splines are piecewise polynomials defined on disjoint partitions of the support of X, with the pieces joining smoothly at a set of interior points (the *knots*). Precisely, a polynomial spline of degree $m \geq 0$ defined on an interval $[a, b]$ with knot sequence $\boldsymbol{\lambda} = \{\lambda_0, \lambda_1, \cdots, \lambda_{k+1}\}$ $(a = \lambda_0 < \lambda_1 < \cdots < \lambda_{k+1} = b)$ is a function consisting of pieces of polynomials of degree m on each of the intervals $[\lambda_l, \lambda_{l+1}),\ l = 0, \cdots, k-1$, and $[\lambda_k, \lambda_{k+1}]$. Given knot sequence $\boldsymbol{\lambda}$ and degree m, the collection of spline functions form a function space spanned by a set of *basis functions*. Commonly used basis functions include the *truncated power basis* consisting of functions $\{1, x, \cdots, x^m, (x-\lambda_1)_+^m, \cdots, (x-\lambda_k)_+^m\}$, where $(x)_+^m = x^m$ if $x > 0$, 0 otherwise, and the *B-spline* basis. The truncated power basis are easy to construct, and when used in regression as independent variables, the estimated coefficients usually have practical meanings and can be interpreted in a meaningful way. The truncated power basis functions are based on the powers of X, therefore are correlated by construction, and multicolinearity can be a problem in regression analysis. The difference in the scale of the basis functions can also cause numerical instability in computer solutions. On the other hand, B-spline basis function values are always between zero and one, with many zeros, therefore are more stable numerically in computer solutions. B-spline basis are orthogonal, therefore when used as independent variables, the model will be free of multicolinearity. The theoretical properties of B-splines can be found in texts such as de Boor (2001) and Schumaker (1981). B-splines are used in this paper, but the results do not depend on the choice of basis functions. By varying the smoothing parameters (the number and location of the knots, and the degree of the polynomial), polynomial splines can adapt to different types of nonlinear relationships, hence are very flexible and useful in modeling nonlinear relationships (Huang & Shen 2004). The properties of polynomial splines are well documented (e.g., Stone 1994, Huang 2003, Zhou, Shen & Wolfe 1998). Once the knots and degree of the polynomial are known, the model becomes a piecewise polynomial regression model that can be estimated using standard least square method globally, therefore polynomial spline models are very computationally efficient. However, the smoothing parameters must be selected carefully.

3.2. Spline model selection

In this paper we use a method similar as in Eubank (1999) to select the smoothing parameters. For the model selection criterion, we experimented with the Akaike Information Criteria (AIC), AIC with finite sample correction (AICc, Sugiura 1978), and the Bayesian Information Criteria (BIC). Our experience is that AIC tends to select more knots and over fit the model. AICc has larger penalty on the number of parameters than AIC, but for the typical

sample sizes and number of parameters in this paper, the penalty is not big enough and it also tends to over fit the model. BIC tends to select more parsimonious models due to its large penalty term. As a result, BIC is used as the model selection criterion. With given number of knots (k) and the degree of the polynomial (m), the location of the knots are selected from a sequence of 'candidate' knots placed on the percentile points of X. Polynomial splines models with all possible combinations of the candidate knots are estimated, the set of knots that minimises BIC is chosen as the optimal location of the knots for the given k and m. This process is repeated for different k and m and the model with the smallest BIC value is selected. In practice, m rarely exceed 3 (cubic splines), so typically only degrees 0-3 need to be considered. The number of interior knots depends on the application and should be chosen from a limited range, such as $k \leq 3$, as the number of evaluations required is of order $O(n_c^k)$ (where n_c is the number of the 'candidate' knots) and increases very fast with k. If k must be greater than 3, then one may use the theoretically optimal number of knots $n^{1/(2p+1)}$ (Huang 2003) as a guide (where n is the sample size and p is the number of continuous derivatives f has), and place the knots equally spaced or on the percentile points of X. This algorithm is described below:

- For given m and k, consider n_c potential knots placed on the percentile points of X,
  - Let $\boldsymbol{\lambda}_l = \lambda_{l1} < \lambda_{l2} < \cdots < \lambda_{lk}$ be the l-th combination of k different candidate knots ($l = 1, \cdots, \binom{n_c}{k}$).
  - Estimate model l: $\boldsymbol{Y} = \boldsymbol{X}_l \boldsymbol{\beta}_l + \boldsymbol{\epsilon}$, where $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)^\top$, $\boldsymbol{X}_l$ is the matrix consisting of the $m + k + 1$ polynomial spline basis constructed using degree m and $\boldsymbol{\lambda}_l$.
  - Identify the location of the k interior knots by minimising

$$BIC = n \ln\left(\frac{RSS_l}{n}\right) + (m + k + 1)\ln(n),$$

    where RSS_l is the residual sum of squares for model l.

- Repeat the above process for all combinations of k and m, select the combination of k, m and $\boldsymbol{\lambda}$ that minimises BIC.

This model selection method is data-driven and selects the models automatically for each subset, therefore helps alleviate the problem of identifying many potentially different nonlinear models.

3.3. Application of the model selection method

Before applying the above model selection method on the data, electricity usage Y_t is tested using the Dickey--Fuller test (Dickey & Fuller 1981), the test result indicates a

unit root. As a result, the first difference $W_t = \nabla Y_t = Y_t - Y_{t-1}$ is used as the dependent variable in subsequent analyses. For each subset, the 'candidate' knots are placed on every 5th percentile point of temperature. We consider polynomial degrees $m = 0, 1, 2, 3$. In our analysis, we found that almost all subseries favor fewer than 2 knots, in the very few cases where more than two knots are selected, the BIC values are only slightly lower than that of two knots. As a result, to expedite the computation, the maximum number of interior knots is set at 2, when $k = 0$, the resulting model is a linear regression model. The results show the degrees of the polynomial are 1 or 2 for most of the subseries. The number and location of the knots and degrees of the polynomial selected in the first iteration for selected days are given in Table 1, where a '0' for knot 1 indicates that the model has no interior knots, which is a linear regression model. A '0' for knot 2 means that there is no need for the second knot, depending on if knot 1 is zero or not, the resulting model is either a linear model or a model with one interior knot. Note the model selection shown in Table1 is done before the removal of the serial correlation, so in iterative estimation, the selected smoothing parameters may converge to different values after the serial correlation is removed. Model (1) is estimated with the selected smoothing parameters, the within-sample residual root means square error (RMSE) is 104.72.

4. Serial correlation in the residuals

As noted by some researchers (e.g., Xiao et al. 2003, Su & Ullah 2006, Liu, Chen & Yao 2010), for time series data, the serial correlation must be properly modelled and removed to improve the efficiency of nonparametric smoothing. The autocorrelation function (ACF) and partial autocorrelation function (PACF) of the partial residuals $\hat{e}_t$ of the polynomial splines model (1) in the first iteration show persistent autocorrelation and a strong seasonality of 24 (Figure 5). As observed by Liu et al. (2006), there is also a weekly seasonality (lag 168) in the data.

Based on the above observations, the following multiplicative ARIMA model with double seasonality is needed to remove the serial correlation in the noise:

$$\Phi_{p_1}(B)\Phi_{p_2}(B^{24})\Phi_{p_3}(B^{168})\nabla^{d_1}\nabla_{24}^{d_2}\nabla_{168}^{d_3}e_t = \Theta_{q_1}(B)\Theta_{q_2}(B^{24})\Theta_{q_3}(B^{168})\varepsilon_t \qquad (2)$$

The orders of the model $(p_1, d_1, q_1, p_2, d_2, q_2, p_3, d_3, q_3)$ can be selected by examining the sample ACF, PACF, IACF, etc. (Box & Jenkins 1976). Given that these orders are often rather small (usually ≤ 2), they can be selected by an exhaustive search from selected combinations of $(p_1, d_1, q_1, p_2, d_2, q_2, p_3, d_3, q_3)$ to minimise BIC. One of the advantages of this approach is that it is automatic so it is suitable if the model must be selected repeatedly.

Table 1. Degree and knots selected in iteration 1 for selected days.

	Wednesdays			Thursdays			Fridays		
Hour	Knot1	Knot2	Degree	Knot1	Knot2	Degree	Knot1	Knot2	Degree
1	61.21	0	1	48.80	0	1	70.12	71.76	1
2	44.02	64.86	1	58.78	0	1	60.31	72.17	1
3	63.51	75.58	2	0	0	2	43.68	0	1
4	0	0	2	52.48	0	1	30.79	0	1
5	38.91	0	1	46.47	0	1	30.85	0	1
6	35.36	0	1	32.95	0	1	29.13	0	1
7	38.47	0	1	39.32	0	1	36.73	0	1
8	26.50	0	1	0	0	1	0	0	1
9	46.87	0	1	0	0	2	34.84	58.41	1
10	0	0	2	0	0	2	34.19	55.60	1
11	58.00	0	1	0	0	2	62.05	0	1
12	61.39	0	1	60.63	0	1	70.11	76.80	1
13	86.42	89.27	3	62.56	82.58	1	70.42	81.55	1
14	67.89	85.18	1	64.34	87.66	1	86.76	90.52	3
15	65.58	84.30	1	82.08	0	2	49.95	76.30	2
16	62.19	0	1	66.03	87.96	1	66.86	82.05	1
17	66.39	0	1	69.38	0	1	46.22	79.20	2
18	76.66	0	1	48.02	74.49	1	77.54	0	1
19	61.19	78.63	1	46.07	68.06	2	70.73	79.30	1
20	64.22	82.20	1	68.07	80.90	1	71.40	74.50	1
21	61.37	63.74	1	65.65	68.92	1	74.12	0	1
22	0	0	1	0	0	1	68.71	0	1
23	60.26	0	1	43.11	71.52	2	0	0	2
24	30.96	37.40	2	62.05	0	1	52.23	0	1

More specifically, the following components of model (2) are considered in the identification of the ARIMA model in this example:

$$
\begin{aligned}
\nabla_{s_i}^{d_i} &= (1-B^{s_i})^{d_i}, \text{ where } s_1=1, s_2=24, s_3=168, d_1=d_2=d_3=1,\\
\Phi_1(B) &= 1-\phi_{11}B-\phi_{12}B^2-\phi_{13}B^3,\\
\Phi_{24}(B) &= 1-\phi_{21}B^{23}-\phi_{22}B^{24}-\phi_{23}B^{25},\\
\Phi_{168}(B) &= 1-\phi_{31}B^{167}-\phi_{32}B^{168}-\phi_{33}B^{169},\\
\Theta_1(B) &= 1-\theta_{11}B-\theta_{12}B^2-\theta_{13}B^3,\\
\Theta_{24}(B) &= 1-\theta_{21}B^{23}-\theta_{22}B^{24}-\theta_{23}B^{25},\\
\Theta_{168}(B) &= 1-\theta_{31}B^{167}-\theta_{32}B^{168}-\theta_{33}B^{169}.
\end{aligned}
$$

Note that within each multiplicative component, not all the combinations of the terms need to be considered, which further simplifies the model selection. For example, for $\Phi_{24}(B)$,

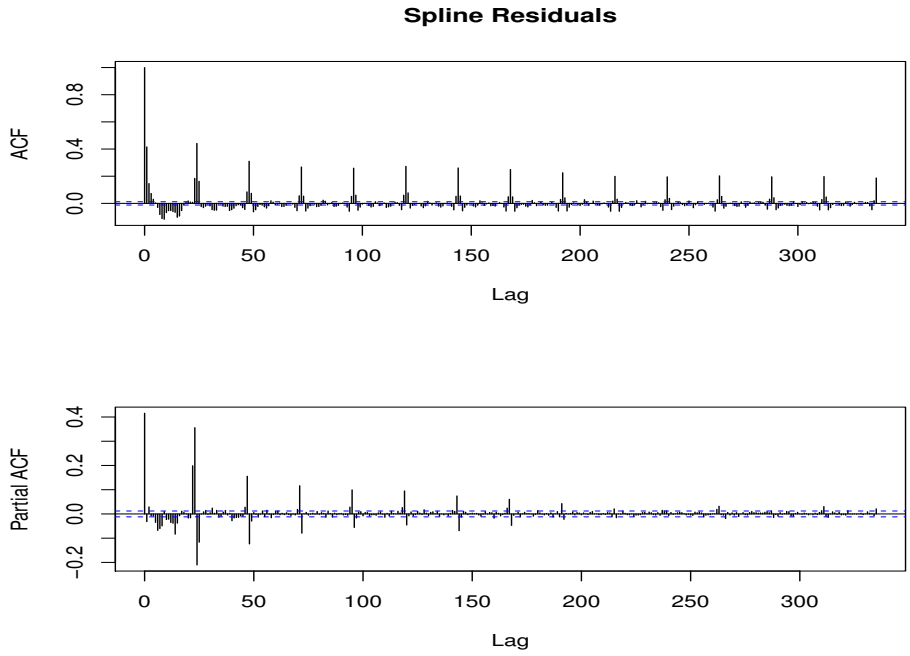

Figure 5. Spline residual ACF and PACF.

the following sub-models are considered: the 'null', $1-\phi_{22}B^{24}$, $1-\phi_{21}B^{23}-\phi_{22}B^{24}$, $1-\phi_{22}B^{24}-\phi_{23}B^{25}$, and $1-\phi_{21}B^{23}-\phi_{22}B^{24}-\phi_{23}B^{25}$.

The above algorithm is used to identify the following ARIMA model, which agrees with the selection made by using the traditional Box--Jenkins approach,

$$(1-B^{24})e_t = \frac{(1-\theta_1 B^{24})(1-\theta_2 B^{168})}{(1-\phi_1 B-\phi_2 B^2)(1-\phi_3 B^{23}-\phi_4 B^{24}-\phi_5 B^{25})}\,\varepsilon_t\ , \tag{3}$$

where e_t is the residual from the nonparametric model (1), ε_t is i.i.d. $\mathrm{N}(0,\sigma^2)$. By introducing this time series model, the within-sample RMSE is reduced from 104.72 to 80.68.

5. A two-component model and iterative estimation

Based on the above observations, we combine models 1 and 3 to form the following two-component model:

$$W_{ij} = f_i(X_{ij}) + e_{ij}\ , \tag{4}$$

$$(1-B^{24})e_t = \frac{(1-\theta_1 B^{24})(1-\theta_2 B^{168})}{(1-\phi_1 B-\phi_2 B^2)(1-\phi_3 B^{23}-\phi_4 B^{24}-\phi_5 B^{25})}\,\varepsilon_t\ , \tag{5}$$

where $i = 1, \ldots, 192$, W_{ij} is the jth first-differenced electricity usage in subset i, and X_{ij} is the corresponding temperature, ε_t is i.i.d. $\mathrm{N}(0, \sigma^2)$. After model (4) is estimated, the time series is restored to its original time order for the ARIMA model, therefore time index t is used in 5. The following estimation procedure is used to estimate the two components iteratively.

5.1. Iterative model selection and estimation

Note the spline models (4) are selected in the presence of strong serial correlation, and the selection of the smoothing parameters may be biased. In addition, the identification of the ARMA model will not be possible without having an accurate estimate of f, as the noise e_t is not observable. This dilemma can be solved by selecting the models iteratively between the splines (4) and the ARIMA models (5). Such iterative model selection and estimation method is very useful when the underlying relationship is potentially different in many different subsets, such as in this example.

Under a strong mixing condition for $\{e_t\}$ and mild conditions on the smoothness of $f(\cdot)$ and the number of interior knots, the polynomial spline estimator of f is asymptotically consistent (Huang 2003, Wang & Yang 2007). Consequently, the estimated residuals $\hat{e}_t$ are also consistent estimates of e_t, as a result, the estimated ARMA parameters obtained using $\hat{e}_t$ yields consistent estimates of the ARMA parameters. Let $\phi(B)$ and $\theta(B)$ be polynomials in the back shift operator B, with coefficients satisfying the stationary and invertible conditions, then model (5) can be written as

$$\psi(B)e_t = \varepsilon_t \ ,$$

where

$$\Psi(B) = \frac{(1 - B^{24})(1 - \phi_1 B - \phi_2 B^2)(1 - \phi_3 B^{23} - \phi_4 B^{24} - \phi_5 B^{25})}{(1 - \theta_1 B^{24})(1 - \theta_2 B^{168})} = \sum_{i=0}^{\infty} \psi_i B^i$$

is a polynomial of the backshift operator B, Equation (5) can then be written as

$$e_t = -\sum_{i=1}^{t-1} \psi_i e_{t-i} + \varepsilon_t \ .$$

Then (4) becomes

$$W_t - \sum_{i=1}^{t-1} \psi_i [W_{t-i} - f(X_{t-i})] = f(X_t) + \varepsilon_t,$$

where ε_t are independent $\mathrm{N}(0, \sigma^2)$. Let $\eta_t = \sum_{i=1}^{t-1} \psi_i [W_{t-i} - f(X_{t-i})]$. As mentioned above, the initial estimates of $f(\cdot)$ and η_t are consistent and are used to replace the unknown parameters in the following iterative estimation procedure based on the backfiting algorithm by Hastie & Tibshirani (1990):

- Let W_t denote the first-differenced electricity demand at t. Let $Z_t = W_t - \hat{\eta}_t$, set $\hat{\eta}_t = 0$ to initialize.
- Do the following until convergence:
  - Split Z_t into subseries Z_{ij} by hour and day, where $i = 1, \cdots, 192$.
  - Estimate $Z_{ij} = f_i(X_{ij}) + e_{ij}$ using polynomial splines, obtain $\widehat{f}_i(\cdot)$.
  - Combine $\widehat{f}_i(X_{ij})$ into one series $\widehat{f}(X_t)$ in the original time order, let $\widehat{e}_t = W_t - \widehat{f}(X_t)$.
  - Select the model and estimate

$$(1 - B^{24})\widehat{e}_t = \frac{(1 - \theta_1 B^{24})(1 - \theta_2 B^{168})}{(1 - \phi_1 B - \phi_2 B^2)(1 - \phi_3 B^{23} - \phi_4 B^{24} - \phi_5 B^{25})}\, \varepsilon_t,$$

    obtain $\widehat{\varepsilon}_t$.
  - Obtain the linear projection $\widehat{\eta}_t = \widehat{e}_t - \widehat{\varepsilon}_t$.
  - Set $Z_t = W_t - \widehat{\eta}_t$.
- After convergence, the estimated value of W_{ij} is $\widehat{W}_{ij} = \widehat{f}_i(X_{ij}) + \widehat{\eta}_{ij}$ and the estimate of ε_t is $\widehat{\varepsilon}_t = W_t - \widehat{W}_t$.

6. Estimation and forecasting results

The algorithm described above is run iteratively, the within-sample residual root mean square errors (RMSE) for both the nonparametric component (4) and the ARIMA component (5) are continuously monitored, as well as the selected smoothing parameters and the estimated coefficients. All these values stabilise after 25 iterations. The RMSEs of the spline component and the ARIMA component are defined as

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

respectively. Figure 6 shows the evolution of within-sample RMSEs of the splines and ARIMA components over the iterations, which shows that the algorithm converges before iteration 25, as a result, the algorithm is terminated at 30 iterations. At convergence, the within-sample RMSE is further reduced to from 80.68 to 75.6.

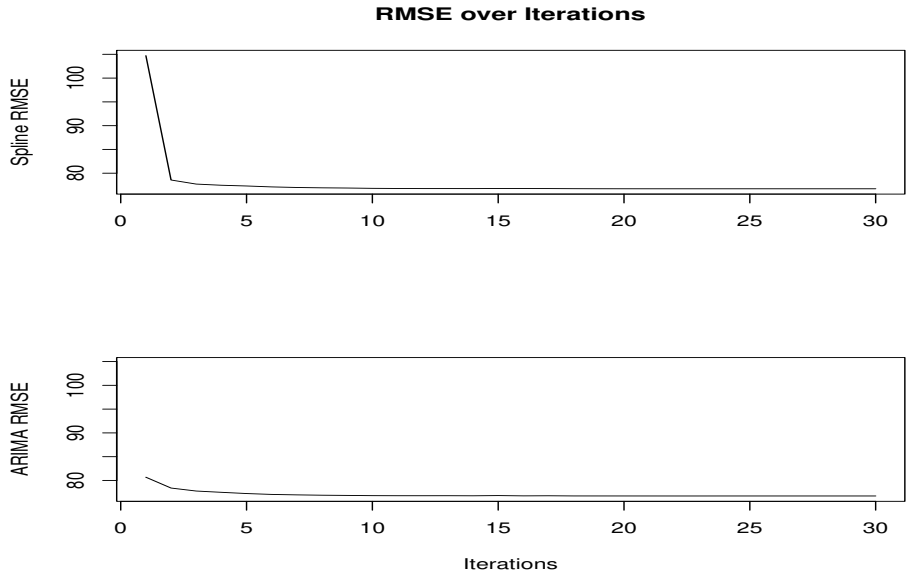

Figure 6. The within-sample RMSE over iterations.

The smoothing parameters (degree of the polynomial, number and location of the knots) of the spline models and the orders of the ARIMA model are selected in each iteration. Although the preliminary estimates are asymptotically consistent, in finite samples, under strong correlation, the smoothing parameters selected may change after the serial correlation is removed. Some of the selected smoothing parameters at iteration 30, after the algorithm converges, are list in Table 2. A comparison with the smoothing parameters selected in the first iteration before the removal of serial correlation (Table 1) shows that some of the parameters changed during the iterations. The interior knots selected in the first 30 iterations are plotted in Figure 7 for each of the 192 subsets, which shows that the locations of most of the knots change in the first a few iterations, but stabilizes before iteration 25 as the algorithm converges. The ARIMA model is also reselected in each iteration, but we find that the orders selected in model (3) remain unchanged throughout the iterations.

The hourly estimated curves for Sundays are shown in Figures 8 and 9 below. From these figures, one can see that the change of hourly electricity usage from the previous hour W_t is of different level at different times. Beginning from 1 am of a day, W_t increases by the hour, until the first peak is reached at 8 or 9 am, then it starts to decrease until noon or early afternoon, at which it starts to increase again until another peak is reached in early evening, after which it decreases again, until another cycle begins around midnight. This is consistent with the typical two-peak pattern in hourly electricity usage. The temperature effect is also different at different times and days. In addition to the differences in the shape of the estimated

Table 2. Knots and degrees at iteration 30 for selected days.

	Wednesdays			Thursdays			Fridays		
Hour	Knot1	Knot2	Degree	Knot1	Knot2	Degree	Knot1	Knot2	Degree
1	0	0	1	0	0	1	72.92	0	1
2	67.71	0	1	0	0	1	74.67	0	2
3	0	0	1	0	0	1	0	0	1
4	0	0	1	0	0	1	0	0	1
5	0	0	1	0	0	1	0	0	1
6	40.80	57.12	1	48.87	0	2	37.08	61.14	1
7	38.47	0	1	35.93	0	1	34.17	0	1
8	0	0	1	30.51	31.67	1	0	0	1
9	46.87	0	1	43.24	0	1	34.84	0	1
10	46.40	0	1	36.08	0	1	34.19	0	1
11	39.72	50.84	1	0	0	2	61.09	0	1
12	59.10	0	1	41.83	0	2	63.87	0	1
13	86.42	89.27	3	67.85	83.94	1	70.42	81.55	1
14	67.89	85.18	1	64.34	87.66	1	86.76	90.52	3
15	65.58	82.10	1	57.28	85.61	1	74.40	79.89	1
16	0	0	1	45.32	51.23	3	66.86	82.05	1
17	92.01	0	1	0	0	1	46.22	52.91	1
18	83.73	0	1	43.38	48.02	3	77.54	0	1
19	63.52	79.91	1	41.06	0	3	43.02	73.84	2
20	64.22	0	1	68.07	80.90	1	61.23	82.42	1
21	66.60	0	1	68.92	0	1	77.11	78.21	1
22	33.53	69.25	1	69.94	81.04	1	82.00	0	2
23	63.78	0	1	39.74	49.35	1	0	0	2
24	54.31	76.07	1	57.96	0	1	49.01	0	1

curves, in the early hours of a day, W_t is negatively associated with temperature, there is a transition point at 8 or 9 am at which the association changes to positive. The association stays positive until late afternoon or early evening, then it becomes negative again. The above observations from Figures 8 and 9 are representative of Weekends and Holidays. For non-holiday weekdays similar patterns are observed with some differences, for example, in weekdays W_t reaches its peak at 6 or 7 am, about one hour earlier than in weekends/Holidays, and the association between W_t and temperature transition from negative to positive also about one hour earlier, at 7 am or 8 am. The hourly estimated curves for other days are not included because the patterns shown are very similar.

At convergence, the estimated ARMA coefficients of model 5 are shown in Table 3. The sample ACF and PACF of the final residuals $\hat{\varepsilon}_t$ indicate that the residual series is roughly a white noise process (Figure 10).

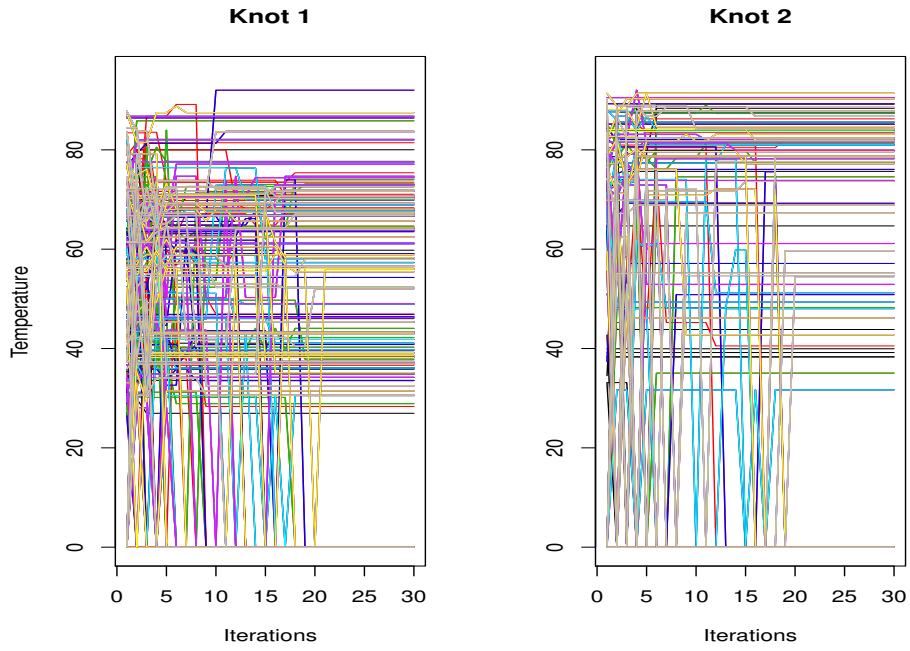

Figure 7. The selected interior knots over iterations.

Table 3. The estimated ARMA coefficients of (5).

Coefficient	Value	Std.Err	t-value	Type	Order	p-value
ϕ_1	.433	.006	70.67	AR	1	$< .0001$
ϕ_2	.110	.006	17.96	AR	2	$< .0001$
ϕ_3	.083	.006	13.63	AR	23	$< .0001$
ϕ_4	.128	.008	15.96	AR	24	$< .0001$
ϕ_5	.047	.006	7.63	AR	25	$< .0001$
θ_1	.811	.005	167.6	MA	24	$< .0001$
θ_2	-.021	.006	-3.37	MA	168	0.0007

The computation is carried out using SAS 9.4 on a PC with an 11th generation Intel Core i7-11770k 8-core processor, the total computational time for 30 iterations is 18 minutes and 24 seconds.

6.1. Forecasting performance

To evaluate the forecasting performance, a post-sample (one-step ahead) forecast using data of year 2001 is done with the converged two-component model. Actual temperatures are used in the forecast. The post-sample forecast RMSE is 85.22. As an additional

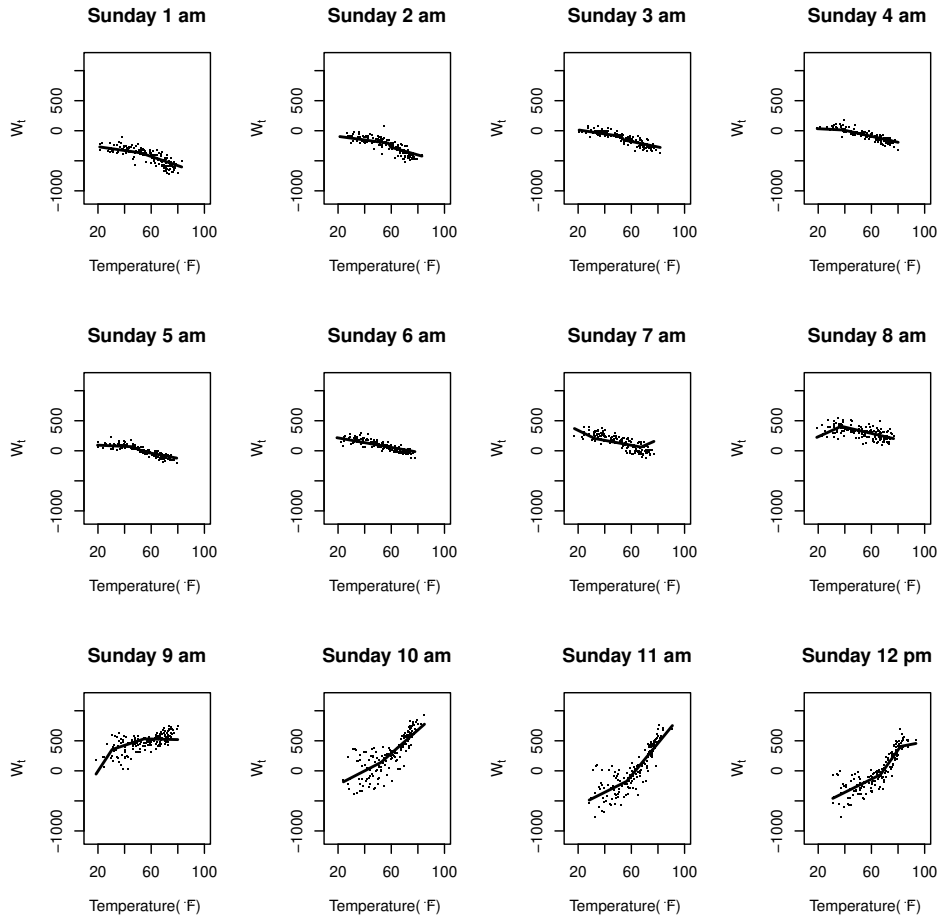

Figure 8. $\widehat{f}(X_{ij})$: Sundays 1am to 12pm.

accuracy measure, the Mean Absolute Percentage Error (MAPE) is also calculated. MAPE $= 1/n \sum_{t=1}^{n}(|Y_t - \widehat{Y}_t|/Y_t)$, where Y_t is the actual observation and $\widehat{Y}_t$ is the forecast at time t. The within- and post-sample MAPEs are 0.82% and 0.886%, respectively. The forecast errors are plotted in Figure 11 below, other than a few outliers, it shows no particular systemic pattern.

To help assess the uncertainty involved with the forecasts, the frequency distribution and summary statistics of the forecast errors are provided in Tables 4 and 5 below. Where in Table 5, Q_1 and Q_3 are the first and third quartile, respectively, and $\text{MOE}_{p\%}$ denotes the margin of error of a $p\%$ prediction interval.

To provide some visual examples of the forecasting performance, the post-sample forecasts for two weeks in winter (January 1 to January 14, 2001) and two weeks in summer

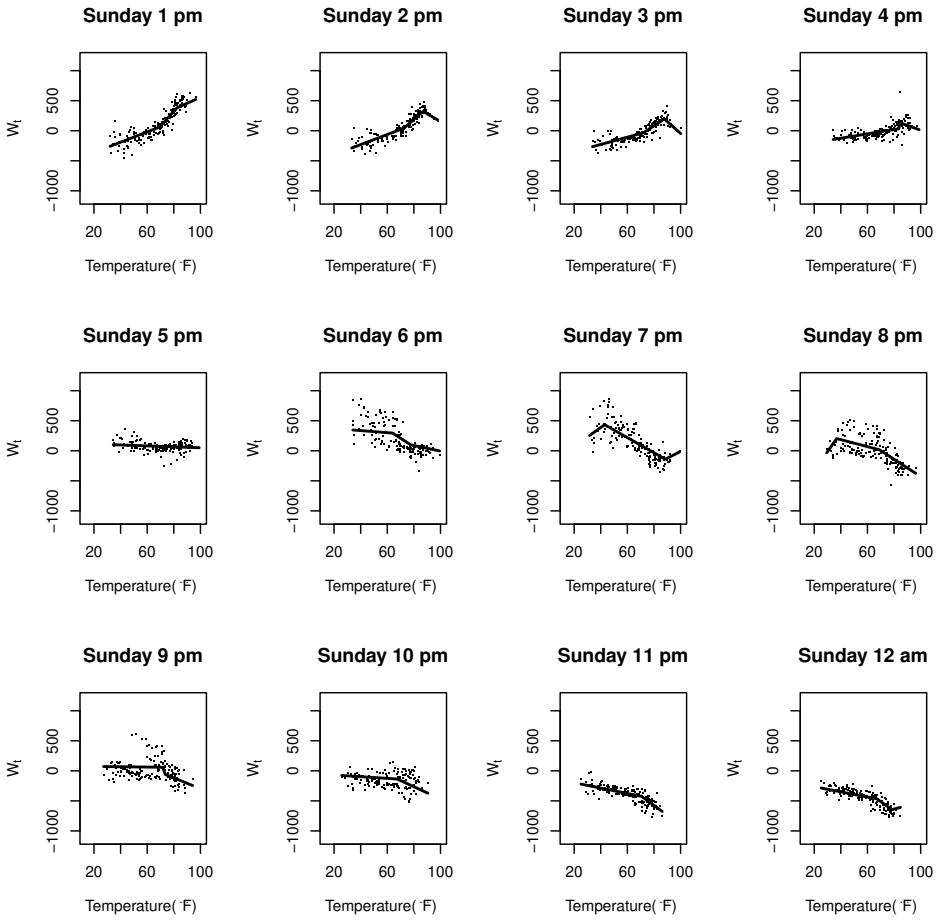

Figure 9. $\widehat{f}(X_{ij})$: Sundays 1pm to 12am.

Table 4. The frequency distribution of the forecast errors.

Classes	Frequency	%
$[-900, -500)$	7	0.08
$[-500, -300)$	41	0.47
$[-300, -100)$	717	8.19
$[-100, 100)$	7238	82.64
$[100, 300)$	718	8.20
$[300, 500)$	35	0.40
$[500, 1100)$	3	0.03
Total	8759	100

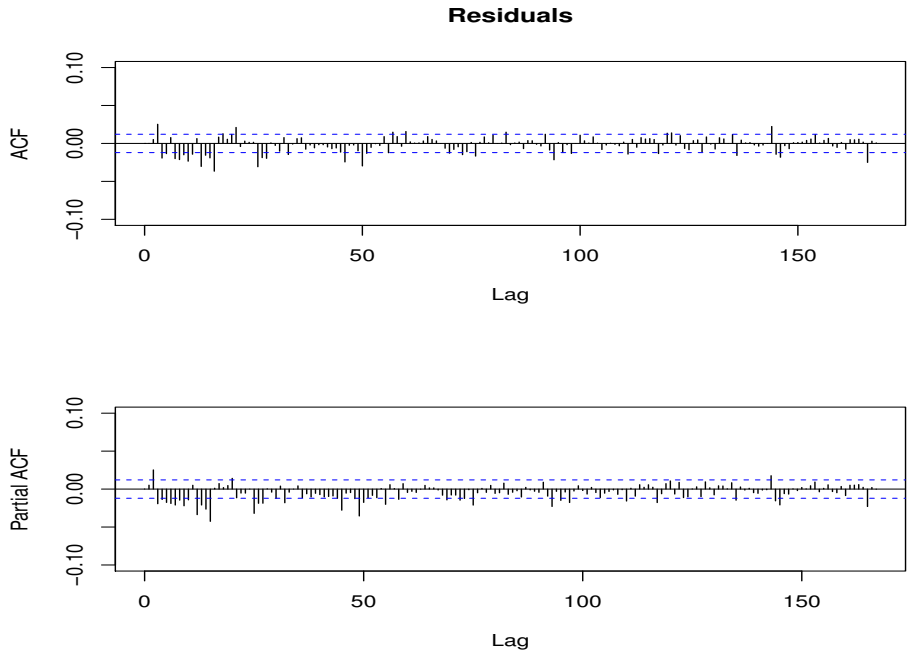

Figure 10. The ACF and PACF of the final residuals.

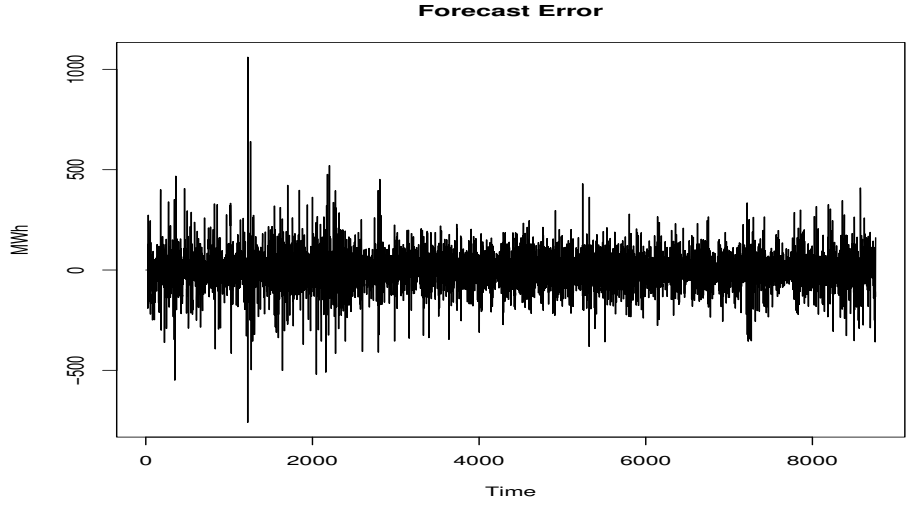

Figure 11. Forecast errors.

429 (August 1 to August 14, 2001) are plotted in Figure 12, where the actual observations are

Table 5. Summary statistics of the forecast errors.

Min	Q_1	Median	Q_3	Max
-823.388	-40.405	0.613	42.350	1061.160
Mean	stdev	$\text{MOE}_{90\%}$	$\text{MOE}_{95\%}$	$\text{MOE}_{99\%}$
-0.066	85.972	141.40	168.48	221.43

represented by the black solid line, and the forecasts are represented by the red dashed line. Also shown in the figure are the forecast errors in the same time periods.

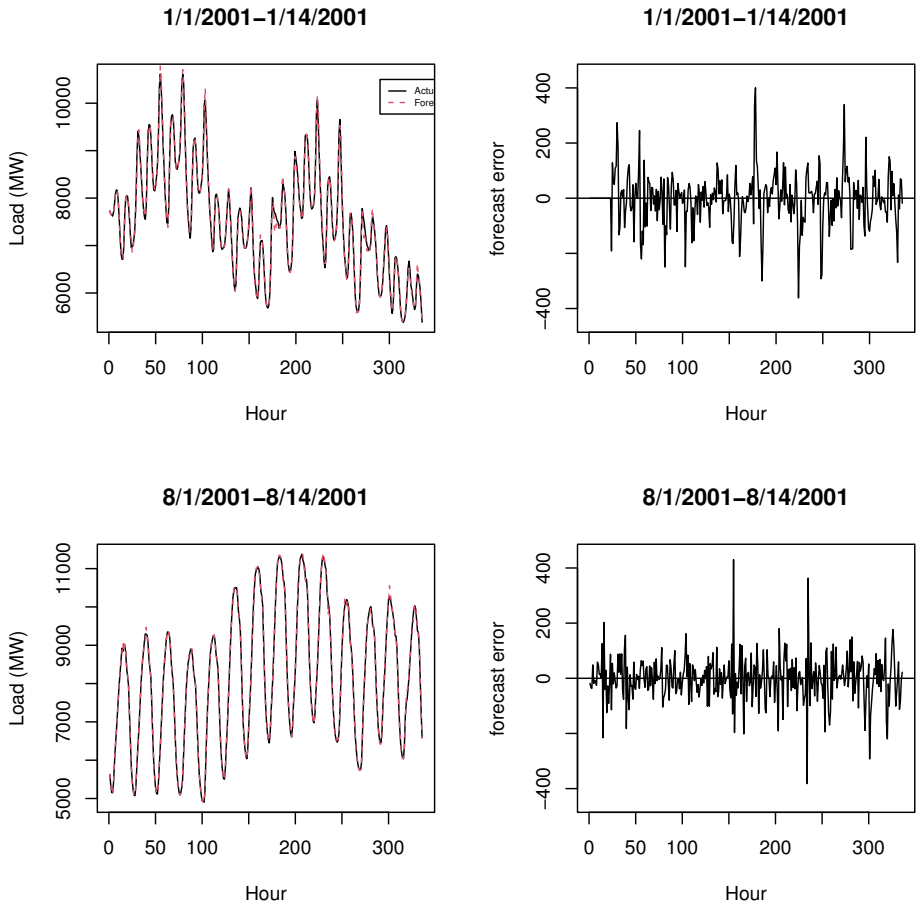

Figure 12. Segments of forecasts and forecast errors.

6.2. Model comparison

In Liu et al. (2006), the local polynomial-based nonparametric transfer function (henceforth LPTF) model performs favorably to the EGRV model by Engle et al. (1993) in post-sample forecasting. The EGRV model is reported to have outperformed several well-accepted models in a competitive forecasting experiment hosted by the Puget Sound Power and Light Company (Ramanathan et al. 1997), therefore the LPTF is used as a benchmark to evaluate the performance of the Polynomial Spline-based nonparametric Transfer Function model considered in this paper (henceforth PSTF).

Due to their ability of modelling complex nonlinear relationships, deep learning models have gained popularity in load forecasting in recently years. Among various deep-learning architectures, the long short-term memory (LSTM) is designed to overcome the vanishing gradient problem and is able to to utilise information in long-term memory of time series data. LSTM-based models have been widely used to forecast short-term load forecasting and have been shown to perform very well. For example, Gasparin, Lukovic & Alippi (2022) shows that the LSTM-based model with exogenous variables outperform several other deep-learning models in forecasting aggregate load; Bouktif et al. (2018) shows that the LSTM-based model outperforms several other models including ridge regression, nearest neighbour, random forest, gradient boosting, neural network, and extra trees. As a result, we develop an LSTM-based deep learning model using the same data set and compare the performance. The first two years of the data are used as the training set, the third year of data are used to validate the training results, and the fourth year of data are reserved as the test set. The numeric variables (load and temperature) are standardized, and the categorical variables (day of the week and hours of the day) are coded as dummy variables before training. The hyper parameters (the number of hidden layers, the number of hidden nodes, batch sizes, number of epoches, etc.) are selected via gird searches using validation RMSE as the criterion. Bouktif et al. (2018) uses the Genetic Algorithm (Goldberg 1989) to find that six hidden layers perform the best in a LSTM model. We select the number of hidden layers through grid search and find the optimal number to be three for this data set. The lags of load contain important information about future electricity demand. The optimal number of lags is found to be 36 via a grid search from a range of 1 to 60 with an increment of 3, as a result, the first 36 lags of load are included in the inputs. The number of neurons in the LSTM layer (32) is selected from 20 to 60 with an increment of 4. The batch size is set to be 150. The maximum number of epoches is set at 400, but during in training we found that all instances converged before epoch 400. In the model selection process, all exogenous variables (temperature, day of the week, and hours of the day) are found important because including them significantly improves the performance of the model. The optimizer used is the adaptive moment estimation (ADAM).

The identified model is trained until convergence and the trained model is applied on the test set for one-step-ahead forecasts. The process is repeated 200 iterations, the average of the training and testing MAPE and RMSE are reported in Table 7, with the standard deviations in the parenthesis.

The computation is done on the same PC mentioned before with Anaconda. Multiple iterations are run with different hyper parameters to select the model, the total computation time is 4.9 hours. The computational time depends on the complexity of the model and ranges from 36.9 to 241.3 seconds each iteration, with an average of 89.5 seconds and a standard deviation of 44.3 seconds. We notice that the process can be sped up significantly by focusing on more promising ranges of the parameters. For example, in selecting the number of lags, given the strong 24-lag serial correlation in the data, focus can be placed on lags 24--48, and sparser grids may be used on short and very long lags. The number of epochs (400) can also be safely reduced, because most of the iterations converge before 200 epochs.

As a baseline, the following Linear Transfer Function (LTF) model (6) is identified using the Box-Jenkins approach (Box & Jenkins 1976), and used to perform the same post-sample forecast. The estimated coefficients of the LTF model are listed in Table 6.

$$\nabla_{24}\nabla_{168}Y_t = \frac{\omega_0}{1-\delta_1 B}X_t + \frac{(1-\theta_1 B^{24})(1-\theta_2 B^{168})}{(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)(1-\phi_4 B^{23}-\phi_5 B^{24}-\phi_6 B^{25})}\,a_t \tag{6}$$

Table 6. Estimated coefficients of the linear transfer function model (6).

Coefficient	Value	Std.Err	t-value	p-value	Type	Order
ϕ_1	1.489	.006	241.30	$<.0001$	AR	1
ϕ_2	-.437	.011	-40.34	$<.0001$	AR	2
ϕ_3	-.066	.006	-10.69	$<.0001$	AR	3
ϕ_4	.121	.006	19.89	$<.0001$	AR	23
ϕ_5	.179	.009	20.12	$<.0001$	AR	24
ϕ_6	.074	.006	12.19	$<.0001$	AR	25
θ_1	.764	.004	127.71	$<.0001$	MA	24
θ_2	.837	.003	240.33	$<.0001$	MA	168
ω_0	.058	.045	1.27	.205	Num.	0
δ_1	-.853	.172	-4.95	$<.0001$	Denom.	1

All the above models are applied on the same data set and the within- and post-sample RMSE and MAPE are compared, the results are summarised in Table 7.

The results in Table 7 shows that the performance of the LTF model is limited because only linear temperature effect is considered. The PSTF and LPTF models model the temperature effects nonlinearly by different times and days, and perform significantly better

Table 7. Performance comparison.

	Within-sample		Post-sample	
Model	RMSE	MAPE	RMSE	MAPE
LTF	90.81	1.005%	109.90	1.101%
LPTF	89.41	1.006%	96.42	1.051%
PSTF	75.64	0.820%	85.22	0.886%
LSTM	75.20(4.412)	0.871(.070)%	84.12(4.553)	0.934(.069)%

than the LTF which models the temperature effect linearly. The PSTF model proposed in this paper performs better than the LPTF model, with a 15.4% reduction of within-sample RMSE and an 11.5% reduction in post-sample forecasting RMSE. This improvement in performance, however, does not necessarily indicate that the SPTF model is superior to the LPTF model in methodology, but rather reflects some technical differences in the details. For example, in Liu et al. (2006), days are categorized into two groups: workdays and non-workdays. In this study we find that difference in electricity usage pattern still exists among the days within the work/nonwork day groups, especially between Saturday, Sunday and Holidays (e.g., Figure 4), so the temperature effects are modelled by each day of the week and holidays. We believe that the more detailed subsetting in this paper is part of the reason that the SPTF model performs better. In fact, when the same workday/non-workday classification is used in the SPTF model, the difference in performances between the two models is smaller. Another possible reason for the improvement is the early introduction of the double-seasonal ARIMA model with seasonality of 24 and 168 (model 5), which is used from the beginning to the end of the iterations, while in Liu et al. (2006), an ARIMA model with seasonality 24 is used in the iterative estimation, and a double-seasonal ARIMA model similar to model 5 with seasonality of 168 is only introduced at the end of the iteration to account for the weekly seasonality remained in the data. We understand this is for computational expediency, but believe that the weekly seasonality, although not as prominent as the daily seasonality (lag 24), must be accounted for during the iterative estimation so that the models converge to the 'true' solutions. Without these technical differences, we believe that the SPTF model and the LPTF model should performs similarly. However, the SPTF model only takes a fraction of the time required for the LPTF model to estimate and forecast, which is an important advantage in short-term forecasting. The LSTM model, due to its abilities of modelling nonlinear relationships and learning from past information, also performs very well. Its performance is comparable with the SPTF model proposed in this paper, with a slightly lower post-sample RMSE and a slightly higher post-sample MAPE.

7. Summary and discussion

In this paper a semi-parametric model for forecasting nonlinear time series data is developed. The model consists of a nonparametric component and a parametric ARIMA component. The nonparametric component models the transfer function using polynomial splines, which is data-driven and computationally efficient. The model components are identified automatically, and the estimation is carried out iteratively using a modified backfitting procedure. Because the transfer function f is modelled using a nonparametric smoother, this model is very flexible and can be used to model highly nonlinear relationship of unknown functional forms. By modeling the noise e_t explicitly as an ARIMA process, the serial correlation is removed and the transfer function can be estimated efficiently. In addition, the correlation structure represented by the estimated ARIMA parameters can be used to improve forecasting performance. This modeling procedure is used to forecast hourly electricity usage and found to be accurate and computationally efficient. The method developed in this paper can be applied in a variety of applications sharing similar features of electricity usage, such as other utilities (water, natural gas, etc.), telecommunications, etc. The proposed approach can be extended to higher dimensional regression problems via additive models. Heteroscedasticity is a common feature in forecasting, the methodology considered in this paper can be extended to accommodate heteroscedastic noise, the research on this topic is currently ongoing.

Data availability statement:

The data used in this paper is available upon reasonable request.

References

AL-ZAYER, J. & AL-IBRAHIM, A. (1996). Modelling the impact of temperature on electricity consumption in the eastern province of Saudi Arabia. *Journal of Forecasting* **15**, 97–106.

AMARASINGHE, K., MARINO, D.L. & MANIC, M. (2017). Deep neural networks for energy load forecasting. In *2017 IEEE 26th International Symposium on Industrial Electronics (ISIE)*. pp. 1483–1488.

ANEIROS-PÉREZ, G. & VIEU, P. (2008). Nonparametric time series prediction: A semi-functional partial linear modeling. *Journal of Multivariate Analysis* **99**, 834–857.

AUESTAD, B. & TJÖSTHEIM, D. (1990). Identification of nonlinear time series: First order characterization and order estimation. *Biometrika* **77**, 669–687.

BENGIO, Y., SIMARD, P. & FRASCONI, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* **5**, 157–166.

BIANCHI, F., MAIORINO, E., M.C., K., RIZZI, A. & JENSSEN, R. (2017). An overview and comparative analysis of recurrent neural networks for short term load forecasting. *CoRR* .

BOROVYKH, A., BOHTE, S. & OOSTERLEE, C.W. (2017). Conditional time series forecasting with convolutional neural networks.

BOUKTIF, S., FIAZ, A., OUNI, A. & SERHANI, M.A. (2018). Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies* **11**.

BOX, G. & JENKINS, G. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day, 1st edn.

CAI, Z., FAN, J. & YAO, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association* **95**, 941–956.

CARROLL, R., FAN, J., GIJBELS, I. & WAND, M. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association* **92**, 477–489.

CHEN, D. & BUNN, D. (2014). The forecasting performance of a finite mixture regime-switching model for daily electricity prices. *Journal of Forecasting* **33**, 364–375.

CHEN, R. & TSAY, R. (1993a). Functional-coefficient autoregressive models. *Journal of the American Statistical Association* **88**, 298–308.

CHEN, R. & TSAY, R. (1993b). Nonlinear additive ARX models. *Journal of the American Statistical Association* **88**, 955–967.

CHEN, R. & TSAY, R. (1996). Nonlinear transfer functions. *Journal of Nonparametric Statistics* **66**, 193–204.

CHO, K., VAN MERRIENBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H. & BENGIO, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation.

CLEVELAND, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the Americian Statisical Association* **74**, 829–836.

DE BOOR, C. (2001). *A Practical Guide to Splines*. Springer-Verlag: New York.

DICKEY, D. & FULLER, W. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econmetrica* **49**, 1057–1072.

DORADO RUEDA, F., DURÁN SUÁREZ, J. & DEL REAL TORRES, A. (2021). Short-term load forecasting using encoder-decoder wavenet: Application to the french grid. *Energies* **14**.

DOUC, R., MOULINES, E. & STOFFER, D. (2014). *Nonlinear Time Series Theory, Methods and Applications with R Examples*. Chapman and Hall/CRC.

DREZGA, I. & RAHMAN, S. (1999). Short-term load forecasting with local ann predictors. *IEEE Transactions on Power Systems* **14**, 844–850.

ELMAN, J.L. (1990). Finding structure in time. *Cognitive Science* **14**, 179–211.

ENGLE, F., GRANGER, C., RAMANATHAN, R. & VAHID-ARRAGHI, F. (1993). Probabilistic methods in forecasting hourly loads. *Electric Power Research Institute, EPRI TR-101902, Palo Alto, California* .

ENGLE, F., GRANGER, C., RICE, J. & WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* **81**, 310–320.

EUBANK, R.L. (1999). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker: New York.

FAN, J. & GILBELS, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall: Suffolk.

FAN, J. & YAO, Q. (2003). *Nonlinear Time Series.* Springer: New York.

FAN, J., YAO, Q. & CAI, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society, Series B* **65**, 57–80.

GASPARIN, A., LUKOVIC, S. & ALIPPI, C. (2022). Deep learning for time series forecasting: The electric load case. *CAAI Transactions on Intelligence Technology* **7**, 1–25.

GOLDBERG, D.E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* USA: Addison-Wesley Longman Publishing Co., Inc.

GUPTA, P. (1985). Adaptive short-term forecasting of hourly loads using weather information. In *Comparative Models for Electrical Load Forecasting*, eds. D. Bunn & E. Farmer. John Wiley & Sons: New York.

HÄRDLE, W., HALL, P. & ICHIMURA, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics* **21**, 157–178.

HÄRDLE, W., LIANG, H. & GAO, J. (2000). *Partially Linear Models.* Physica-Verlag,Heidelberg.

HART, J. & VIEU, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *The Annuals of Statistics* **18**, 873–890.

HARVEY, A. & KOOPMAN, S. (1993). Forecasting hourly electricity demand using time-varying splines. *Journal of the American Statistical Association* **88**, 1228–1236.

HASTIE, T. & TIBSHIRANI, R. (1990). *Generalized Additive Models.* Chapman and Hall: London.

HIPPERT, H., PEDREIRA, C. & SOUZA, R. (2001). Neural networks for short-term load forecasting: a review and evaluation. *IEEE Transactions on Power Systems* **16**, 44–55.

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.

HOCHREITER, S. & SCHMIDHUBER, J. (1997). Long Short-Term Memory. *Neural Computation* **9**, 1735–1780.

HONG, T. & FAN, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting* **32**, 914–938.

HONG, T., PINSON, P., FAN, S., ZAREIPOUR, H., TROCCOLI, A. & HYNDMAN, R.J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting* **32**, 896–913.

HONG, T., XIE, J. & BLACK, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting* **35**, 1389–1399.

HUANG, J. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics* **31**, 1600–1635.

HUANG, J. & SHEN, H. (2004). Functional coefficient regression models for nonlinear time series models: a polynomial spline approach. *Scandinavian Journal of Statistics* **31**, 515–534.

KEZUNOVIC, M., PINSON, P., OBRADOVIC, Z., GRIJALVA, S., HONG, T. & BESSA, R. (2020). Big data analytics for future electricity grids. *Electric Power Systems Research* **189**, 106788.

KOHN, R., ANSLEY, C.F. & WONG, C.M. (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika* **79**, 335–346.

KONG, W., DONG, Z., JIA, Y., HILL, D.J., XU, Y. & ZHANG, Y. (2019). Short-term residential load forecasting based on lstm recurrent neural network. *IEEE Transactions on Smart Grid* **10**, 841–851.

KUO, P.H. & HUANG, C.J. (2018). A high precision artificial neural networks model for short-term energy load forecasting. *Energies* **11**.

LAM, C. & YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics* **40**, 694 – 726.

LECUN, Y., BOTTOU, L., BENGIO, Y. & HAFFNER, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324.

LEE, K., CHA, Y. & PARK, J. (1992). Short-term load forecasting using an artificial neural network. *IEEE Transactions on Power Systems* **7**, 124–132.

LEWIS, P. & STEVENS, J. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (mars). *Journal of the American Statistical Association* **86**, 864–877.

LIU, J., CHEN, R., LIU, L.M. & HARRIS, J. (2006). A semi-parametric time series approach in modelling hourly electricity loads. *Journal of Forecasting* .

LIU, J., CHEN, R. & YAO, Q. (2010). Nonparametric transfer function models. *Journal of Econometrics* **157(1)**, 151–164.

LIU, X., CAI, Z. & CHEN, R. (2015). Functional coefficient seasonal time series models with an application of Hawaii tourism data. *Computational Statistics* **30**, 719–744.

LIU, X. & CHEN, R. (2016). Regime-switching factor models for high-dimensional time series. *Statistica Sinica* **26**, 1427–1451.

LIU, X. & CHEN, R. (2020). Threshold factor models for high-dimensional time series. *Journal of Econometrics* **216**, 53–70.

MASRY, E. (1996a). Multivariate local polynomial regression for time series: Uniform consistency and rates. *Journal of Time Series Analysis* **17**, 571–599.

MASRY, E. (1996b). Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and Their Applications* **65**, 81–101.

NEWEY, W. & STOKER, T. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica* **61**, 1199–1223.

PARK, D., EL-SHARKAWI, M., MARKS, R., ATLAS, L. & DAMBORG, M. (1991). Electric load forecasting using an artificial neural network. *IEEE Transactions on Power Systems* **6**, 442–449.

PENG, T., HUBELE, N. & KARADY, G. (1992). Advancement in the application of neural networks for short-term load forecasting. *IEEE Transactions on Power Systems* **7**, 250–257.

POSKITT, D. (1989). A method for the estimation and identification of transfer function models. *Journal of the Royal Statistical Society* **B51**, 29–46.

RAMANATHAN, R., ENGLE, R., GRANGER, C., VAHID-ARAGHI, F. & BRACE, C. (1997). Short-run forecasts of electricity loads and peaks. *International Journal of Forecasting* **13**, 161–174.

ROBINSON, P. (1983). Nonparametric estimators for time series. *Journal of Time Series Analysis* **4**, 185–207.

SCHUMAKER, L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

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

SMITH, M., WONG, C. & KOHN, R. (1998). Additive nonparametric regression with autocorrelated errors. *Journal of the Royal Statistical Society Series B* **60**, 311–331.

STONE, C.J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22**, 118–184.

SU, L. & ULLAH, A. (2006). More efficient estimation in nonparametric regression with nonparametric autocorrelated errors. *Econometric Theory* **22**, 98–126.

SUGIURA, N. (1978). Further analysts of the data by akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods* **7**, 13–26.

TIAO, G. & BOX, G. (1981). Modelling multiple time series with applications. *Journal of the American Statistical Association* **76**, 802–816.

TONG, H. (1983). *Threshold Models in Nonlinear Time Series Analysis (Lecture Notes in Statistics 21)*. New York: Springer-Verlag.

TSAY, R. & CHEN, R. (2018). *Nonlinear Time Series Analysis*. Hoboken, NJ, USA: John Wiley & Sons.

WANG, L. & YANG, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Annals of Statistics* **35(6)**, 2474–2503.

WANG, L. & YANG, L. (2009). Spline estimation of single-index models. *Statistica Sinica* **19**, 765–783.

WANG, P., LIU, B. & HONG, T. (2016). Electric load forecasting with recency effect: A big data approach. *International Journal of Forecasting* **32**, 585–597.

WANG, Y., LIU, M., BAO, Z. & ZHANG, S. (2018). Short-term load forecasting with multi-source data using gated recurrent unit neural networks. *Energies* **11**.

XIA, Y. & LI, W. (1999). On single-index coefficient regression models. *Journal of the American Statistical Association* **94**, 1275–1285.

XIAO, Z., LINTON, O., CARROLL, R. & MAMMEN, E. (2003). Model efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association* **98**, 980–992.

XIE, W., ZHANG, P., CHEN, R. & ZHOU, Z. (2019). A nonparametric bayesian framework for short-term wind power probabilistic forecast. *IEEE Transactions on Power Systems* **34**, 371–379.

ZHOU, S., SHEN, X. & WOLFE, D. (1998). Local asympototics for regression splines and confidence regions. *The Annals of Statistics* **26**, 1760–1782.