Xiaojia Wang[1], Bengong Yu[1], Zhiqiang Chen[1], Liqiang Hou[1]

National Key Laboratory of Process Optimization and Intelligent Decision-Making,
HeFei University of Technology, Hefei, Anhui, China (1)

Forecasting Modeling and Analysis of Power Engineering in China Based on Semi-Parametric Regression Model

Abstract: *Electricity consumption forecasting is considered one of the most important tasks in energy planning, and it has great significance on management decision-making for power generation organizations and power policy adjustments for governments. In this paper, we present a new semi-parametric regression model for consumption forecasting in electrical power systems. We have used the distribution function of student residuals to replace the nonparametric component of the traditional semi-parametric model, thus eliminating the effects of the residual disturbance term according to the change trend of the consumption data themselves. Then, we use differential element theory set information aggregation intervals to create a dynamic weight distribution and improve the forecasting accuracy of the prediction models. Compared with general linear models, our models make statistical inferences and can automatically regulate the boundary effect, which gives the forecast result a higher accuracy. To present a case study, we use the historical data of electricity consumption and related influential factors in China from 1981 to 2010. The simulation results show that both in the model building stage and in the testing stage for this particular case, the SPRM prediction approach proposed in this paper outperforms the other two contrast models, the MAPE of SPRM is 3.21%, much lower than the other two values 3.84% and 13.07%.*

Streszczenie. *W artykule opisano model regresji semiparametrycznej do przewidywania zużycia energii elektrycznej w systemach elektroenergetycznych. W celu eliminacji wywołujących zakłócenia, nieparametrycznych składowych w tradycyjnym modelu semiparametrycznym, zastosowano rozkład studenta. Wykorzystano także metodę różnicową w ustalaniu interwałów zbierania danych, analizowanych przy przewidywaniu. Działanie i skuteczność modelu zweryfikowano z wykorzystaniem prawdziwych danych z lat 1981 do 2010. (**Analiza i modelowanie przewidujące w elektroenergetyki w Chinach – model regresji semiparametrycznej**).*

Keywords: Electricity consumption; Forecasting; Semi-parametric regression; Information aggregation intervals; Dynamic weight distribution

Słowa kluczowe: zużycie energii elektrycznej, przewidywanie, regresja semiparametryczna, interwały zbierania informacji, dynamiczny rozkład wag.

1 Introduction

Electricity consumption forecasting is an important and integral component in the operation of any electric utility whose accuracy directly influences a power system's security, profitability and quality. According to the difference of the prediction mechanism, for electricity consumption prediction, the literature always considers two main problems: mid-long-term consumption forecasts (M-LTCFs, i.e., the annual consumption forecast) for system planning and short-term consumption forecasting (STCF, i.e., the monthly consumption forecast) for maintenance programs. However, the accuracy of consumption forecasting has a significant effect on power system planning and operation, which means that scientific analysis and precise forecasting are even more important. The focus of the present work is on M-LTCFs of annual electricity consumption because it is one of the most important factors for the government to consider when adjusting their power policies.

Adequate electricity production requires that each relevant department of the power system be able to forecast its consumption accurately. However, predicting electricity consumption is Complex because there are many influential factors that characterise and directly or indirectly affect the underlying forecasting process of annual electricity consumption. Most of these factors are uncertain or uncontrollable. These factors can be grouped into two categories: realistic factors (RFs) and Simulation factors (SFs). RFs, such as climate factors, social activities or economic indicators, can affect the real expected value of electricity consumption; at this time node, electricity consumption data are not yet gathered, as this kind of RF acts before the predicted behaviour. SFs, such as reasonable selections of prediction models, can affect the presumable predictive value of electricity consumption; at this time node, actual consumption data have been generated, as this kind of SF acts in the process of prediction.

The aim of the present paper is to analyse and forecast annual electricity consumption in China utilising a semi-parametric regression approach. To integrate the consumption data more accurately and effectively on the basis of the semi-parametric regression approach mentioned above, this paper uses a differential element method to set consumption information aggregation intervals and make dynamic weight distributions.

The remainder of the paper is organised as follows: Section 2 offers a literature review to introduce the different techniques used on the analysis and forecasting of electricity consumption, Section 3 introduces an overview of electricity consumption in China, and Section 4 discusses the methodology and the data of the study, and it provides an accurate model for electricity consumption forecasting. We present a case analysis and result comparisons in Section 5, and, finally, we present our conclusions.

2 Literature Review

Although consumption forecasting is difficult to implement, research on consumption forecasting has attracted wide attention, as the need for and relevance of forecasting electricity consumption has become a much-discussed issue in recent decades. There are many scholars engaged in efforts to improve the accuracy of forecasting methods. Generally speaking, from the classification analysis of the predictive behaviour itself, the methodology for electricity consumption forecasting can be divided into three categories: a).numerical approximation class processing methods; b).statistical regression class processing methods and c).intelligent optimisation class processing methods.

First, numerical approximation class processing methods (NACPM) rely on variations of the data themselves to find information supporting predictive behaviour; these methods do not consider the effects of the other factors. Based on this method, many scholars have drawn a number of useful results. Vincenzo Bianco et al. [1] analysed and forecasted non-residential electricity consumption in Romania by utilising a grey prediction model and a Holt-

Winters model. The author compared the forecast results of the two prediction models and checked the reliability of the predictions. Wang et al. [2] investigated a dynamic GM (1,1) model based on the cubic spline function interpolation principle to forecast electricity consumption in China. The author used piecewise polynomial interpolation thought processing annual electricity consumption data to analyse the electricity consumption trends to make mid-long-term predictions. In reference [3], Diyar Akay et al. proposed a kind of grey prediction model to predict the Turkey's total and industrial electricity consumption, the author used rolling mechanism improved the traditional grey model and obtained high prediction accuracy. In references [4], Wang et al. used Gauss orthogonalisation theory to improve the grey prediction model, and, in constructing the grey combinative interpolation model to forecast electricity consumption in China, they achieved good prediction results. Wang also introduced the Markov Chain theory to the grey combinative interpolation model and constructed the Markov grey orthogonalisation model ([5]) for electricity consumption prediction, which also obtained good prediction accuracy.

Next, statistical regression class processing methods (SRCPM) often consider the synergy of multiple factors affecting predictor variables to measure predictive behaviour. Statistical regression class methods are widely adopted for the electricity consumption forecasting problem. For example, Roula Inglesi [6] analysed the relationship between electricity demand and income, price and population in South Africa and forecasted the electricity demand by creating a model using the Engle-Granger methodology for cointegration and error correction. Ching Lai [7] investigated the impact of weather variables on monthly electricity demand in England and Wales. A multiple regression model was developed to forecast monthly electricity demand based on weather variables, gross domestic product and population growth. Egelioglu et al. [8] studied the influence of economic variables on the annual electricity consumption in northern Cyprus during the period 1988-1997. Through multiple regression analysis, it was found that the number of customers, the price of electricity and the number of tourists correlated with the annual electricity consumption. Wei et al. [9] estimated the long-term electricity load by applying system dynamics, which constructs the model according to an analysis of the historical electricity consumption. This method revealed the great influence of uncertain factors, such as economics and policies. Narayan and Prasad [10] studied the causal effects between electricity consumption and real GDP for 30 OECD countries using the bootstrapped causality testing approach to show how electricity consumption affects the real GDP in Australia, Iceland, Italy, the Slovak Republic, the Czech Republic, Korea, Portugal and the UK. The implication was that electricity conservation policies would negatively impact the real GDP in these 8 countries mentioned, but not the remaining 22 countries. Nikolopoulos et al. [11] compared multiple linear regression (MLR) with an artificial neural network, a nearest neighbour analysis and human judgment; the application results showed that the MLR was less accurate than the other methods because of its inability to handle complex non-linearity in the relationship between the dependent variable and the cues and because of its tendency to misaddress the in-sample data. Abdel-Aal et al. [12] applied an AIM (abductor induction mechanism) model to the domestic consumption in the eastern province of Saudi Arabia in terms of key weather parameters, demographics and economic indicators. It was found that an AIM model, which uses only the mean relative humidity and air temperature, gave an average forecasting error of approximately 5-6% over the year. Yan [13] also presented residential consumption models using climatic variables for Hong Kong.

Finally, intelligent optimisation class processing methods (IOCPM) simulate or reveal some natural phenomena to obtain optimisation methods that adapt to the environment and thus solve the combination forecasting problems that are difficult for traditional forecasting techniques to address by presenting a series of practical programs. Research on IOCPM has provided new and useful ideas for the predicting behaviour itself. M.R. AlRashidi et al. [14] presented a particle swarm optimisation (PSO) algorithm to forecast the long-term electric load in Kuwait. The PSO algorithm was employed to minimise the error associated with the estimated model parameters, and it improved the accuracy of prediction. Nasr et al. [15] presented an artificial neural networks (ANN) approach to electrical energy consumption forecasting in Lebanon. They presented and implemented four ANN models: a univariate model based on past consumption values; a multivariate model based on energy consumption forecasting time series and degree days; a multivariate model based on energy consumption forecasting total imports; and a model combining energy consumption forecasting, degree days and total imports. Metaxiotis [16] provided an overview of studies examining artificial intelligence (AI) technologies and their current use in the field of short-term electrical load forecasting. Santos [17] also used the ANN algorithm to make load forecasts; in this method, the possibility of including weather-related variables in the input vector was also analysed.

3 Analysis and Determination of Influencing Factors on China's Electricity Consumption

Since the reform and opening up, the total consumption of electricity in China has undergone a sustained and significant increase, as shown in Figure 1, where the average growth rate was 9.06% from 1981 to 2010. In the process of growth, however, electricity consumption will inevitably be affected by a number of related factors. Thus, it is necessary to analyse the influential factors to extract the trend of electricity consumption in China and to determine the key factors. Through the literature review and practical experience of China's power sector, we can see that the electricity consumption of China is mainly affected by the regional distribution of electricity, climate, and other relevant policies and economic factors. The following section will analyse and describe the factors of electricity consumption in China specifically.

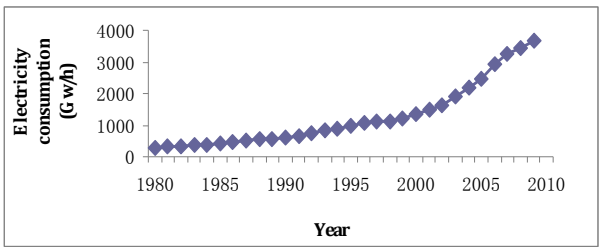

Fig.1. Historical data for electricity consumption in China

3.1 Regional Distribution of Electricity

The vast territory of China covers more than 60 degrees longitude and nearly 50 degrees latitude, and it crosses five time zones. Electricity consumption is obviously affected by different regional factors. The electricity resources of China present a reverse distribution. The developed economic regions in the middle and east of China have an electricity demand that far outweighs electricity supply; thus, the electric power supply gap is enormous. However, in the less

developed areas in western of China, electric power resources are more abundant, so the electricity supply outweighs electricity demand, and the phenomenon of electric energy waste is serious. We can see, then, that the regional distribution on the influence of China's electricity consumption is objective; however, it is also an unchangeable situation because the influence of the regional distribution on power consumption has generally been a relatively stable situation. For annual electricity consumption forecasting, the regional distribution can be regarded as constant factors.

3.2 Climatic Factors

China covers three climatic zones: the equatorial belt, the tropical belt and the temperate belt. The geographical location distribution of China means the climate has the characteristics of four distinct seasons. Figure 2 shows temperature fluctuations within 2°C in China in recent decades. The slope of the temperature over time is relatively gentle, and it presents cyclical fluctuation characteristics. Thus, there is no obvious correlation between the growth rate of China's annual electricity consumption and the volatility of China's annual average temperature.

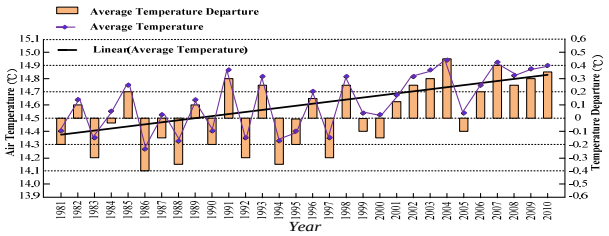

Fig.2. Historical average temperature in China

3.3 Policies and economic

As a developing country, China has implemented an export-oriented economic policy for a long time, so its energy dependence is strong. Usually, turmoil in the financial market will cause volatility in the energy market. Conversely, the instability of the energy market exacerbates the risk of the financial market. Electrical energy plays an important role in energy markets, and therefore, electrical energy is closely connected with economic development. China is currently in a critical stage because economic development is its first priority. Therefore, economic factors can be understood as the key to the development of electric energy.

4 Methodology and Data

4.1 Datasets

Indicators of economic performance can better reflect the trends and levels of electricity consumption. The question, then, is how to find suitable economic indicators. We know that the volatility of the GDP continuously influences the trend of electricity consumption; therefore, the GDP value can be seen as one of the indicators of the index system. At the same time, the "Troika" of the TEIV, IFA and DI can also accurately describe the trends of China's economic growth. Thus, these three indicators can be included in the indicator system because they are representative and rational.

Furthermore, industrial production is an important component of economic production in the current industrial structure of China. Industrial electricity consumption accounts for a large proportion of total electricity consumption, generally 70% or more. On the one hand, industrial production creates great economic benefits; on the other hand, it consumes a large amount of electricity resources. Therefore, the IAV indicator, which reflects the growth trend of industrial production, can also be included in the indicator system.

In summary, we use the indicators GDP, TEIV, IFA, IAV and DI to construct the index system.

These indicators not only reflect the true background of China's power consumption, but their inclusion also enhances the integrity of the index selection. A trend comparison chart of electricity consumption and the five factors is presented in Fig. 3.

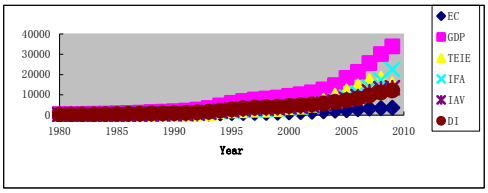

Fig.3. Trend comparison chart of electricity consumption and the five factors

By analysing the data of electricity consumption and the index system, we found that the electricity consumption and the five factors maintained the same linear growth trend. Through the uniform distribution test, we also found that the data of electricity consumption and influential factors over the past 30 years satisfy the uniform distribution and can be divided according to China's five-year development plan, that is, the change trends of electricity consumption and economic growth present a stage of synchronisation. Test results are listed in the following Table 1.

Table 1. Results of uniform distribution test

Six-Sample Kolmogorov-Smirnov Test							
Five year plan		1981~1985	1986~1990	1991~1995	1996~2000	2001~2005	2006~2010
N		5	5	5	5	5	5
Uniform Parameters	Minimum	-0.1	-0.03	-0.01	-0.1	-0.03	-0.25
	Maximum	-0.05	0.04	0.09	-0.04	0.19	0.3
Most Extreme Differences	Absolute	0.2	0.2	0.385	0.394	0.238	0.217
	Positive	0.2	0.2	0.2	0.2	0.2	0.217
	Negative	-0.2	-0.2	0.385	0.394	-0.238	0.2
Kolmogorov-Smirnov Z		0.447	0.447	0.86	0.882	0.532	0.485
Asymp. Sig. (2-tailed)		0.988	0.988	0.45	0.418	0.94	0.973

4.2 Data Standardisation

Many researchers have noted the importance of standardising variables for multivariate analysis. Otherwise, variables measured at different scales do not contribute equally to the analysis. For example, in boundary detection, a variable that ranges between 0 and 100 will outweigh a variable that ranges between 0 and 1. In effect, using these variables without standardisation gives the variable with the larger range a weight of 100 in the analysis. Transforming the data to comparable scales can prevent this problem. Typical data standardisation procedures equalise the range and/or data variability.

The methodology for data standardisation can be divided into three categories: extreme value methods, standardised methods and mean value methods. In this paper, we use a standardised method for two reasons: first, it eliminates the variation of the difference of each variable during dimensionless processing; second, it considers the distribution of the original data, which is what we need to establish the semi-parametric forecasting model. The calculation method is as follows:

where is the raw data to be standardised, is the mean value and is the standard deviation of the raw data.

After standardisation, all variables have the same weight during analysis. In addition, we may decide to weight the data based on our knowledge of the relative importance of the variables.

4.3 Building the Semi-parametric Prediction Model

In the course of electricity consumption data processing, many researchers use the parametric model because its construction is simple and is easy to process. Furthermore, under a majority of situations (for instance, kinds of static problems of conventional historical consumption data), the use of this model remains in accordance with objective fact, and it can satisfy practical needs because a majority of system errors are compensated for and rectified and can be expressed in the parameter model before data processing. However, under some situations (for instance, some dynamic forecast issues of consumption), as the observed values include system errors that cannot be rectified and are not parametric, there are non-ignored differences between the parametric model and objective practicality.

Conversely, scientific integrity is questioned when the system errors attempt to eliminate or compensate for themselves as a harmful composition. In fact, the system errors contain considerable information that influences the observed values. Therefore, if they can be identified and withdrawn correctly, not only can the accuracy of the parameter estimate be increased, but data can be provided for the study of the other subjects.

In addition, the factor of impacting observed values can be divided into two parts: a linear relation and a certain interference factor in which the relationship to observation values is completely unknown, causing it to fall under an error item for no reason. In this case, too much information will be lost if the non-parametric model (though it has greater flexibility) is used; thus, the imitated result is unacceptable if the linear model is adopted.

Given the above problems, other data forecast processing models need to be considered, such as the semi-parametric model:

$$Y_i = X_i^T \beta + g(\xi_i) + \varepsilon_i \qquad (i = 1, 2, \cdots, n) \tag{1}$$

where is observations, or historical electricity consumption, and is explanatory variables, or indicators. The error is assumed to be independent and identically distributed (iid.). is the matrix of unknown parameters, and is the vector of unknown functions. In this paper, we use the distribution function of student residuals to replace the unknown function . For simplicity, let

$$Y = (y_1, \cdots, y_d) = (Y_1^T, \cdots, Y_n^T)^T;$$

$$X = (x_1, \cdots, x_p) = (X_1, \cdots, X_n)^T;$$

$$G = (g_1, \cdots, g_d) = (g(\xi_1)^T, \cdots, g(\xi_n)^T)^T;$$

$$\varepsilon = (\varepsilon_1, \cdots, \varepsilon_d) = (\varepsilon_1^T, \cdots, \varepsilon_n^T)^T$$

The matrix form of the model (1) is

$$Y = X\Omega + G + \varepsilon \tag{2}$$

This is an important type of statistical model developed in the 1980s (Engle [18], 1986). Because it not only contains the parameter weight (which describes the known composition of the function relation in the observation values) but also contains the non-parameter weight (which exclusively shows the model deviation is unknown in the function relation), the model can generalise and describe numerous actual problems, bringing it closer to reality. As a result, the model is extensively developed, and its research is increasingly mature.

In this sub-section, we first provide the prediction principle diagram based on the semi-parametric multiple regression model, and we analyse the forecasting process, which has multiple impact factors. Next, we give the specific steps on how to build the improved semi-parametric prediction model.

4.3(a). Prediction principle diagram

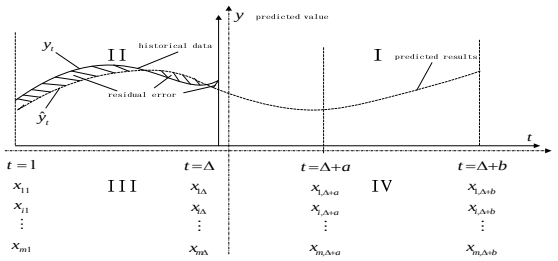

Fig.4. Prediction principle diagram of semi-parametric model with multiple impact factors

In Fig. 4, the tagging below the horizontal axis is the factor for each time period. Assuming that we have collected M kinds of factors associated with predictive object , we denote . Supposing at the historical time period where the value to be predicted is , we need to predict the future at the time period under the law of historical development. If we consider the time axis as the horizontal axis and consider the vertical line with the current time point as the vertical axis, then Fig. 4 can be regarded as a two-dimensional coordinate system with the time point "present" as the coordinate origin. Thus, Fig. 4 can be divided into four quadrants, from Ⅰ to Ⅳ. Therefore, from Fig. 4, we can obtain that the implication of semi-parametric regression forecasting is as follows. We first use the data of quadrants Ⅱ and Ⅲ to proceed a historical fitting operation and derive the forecasting model. Next, we conduct the data of quadrant Ⅳ as the input of the forecasting model; thus, we can obtain the forecasting result of quadrant Ⅰ.

4.3(b). Modelling steps

Step 1: By establishing the multiple linear regression method and solving the parameter part , we obtain , which is the estimated value of .

Step 2: List the fitting residuals, we calculated the standardised residuals and student residual, made a distribution test on the student residual, drew the Q-Q plot, and observed whether it satisfied the normal distribution. The specific process is as follows r_i:

(1) Calculate the student residual

$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{MSE \cdot (1 - h_{ii})}}, \qquad i = 1, 2, \cdots, n$$

where $\hat{\varepsilon}_i$ is the residual vector and $\hat{\varepsilon}_i \sim N(0, \sigma^2(I - H))$, $H = X(X^T X)^{-1} X^T$. The lever quantity h_{ii} is the i-th element

on the leading diagonal of H, and MSE is the mean-square error.

(2) Normal Q-Q plot test for the student residual

(a). obtain the student residual r_i in ascending order $r_{(1)}, r_{(2)}, \cdots, r_{(n)}$;

(b). calculate

$$q_{(i)} = \Phi^{-1}[\frac{i-0.375}{n+0.25}], \quad i = 1,2,\cdots,n$$

where $\Phi^{-1}(x)$ is the inverse function of the standard normal distribution function, and constants 0.375 and 0.25 are corrections.

(c). Use points $(q_{(i)}, r_{(i)})$ $(i = 1,2,\cdots,n)$ in the Cartesian coordinate system to draw a scatter diagram. Observe the points $(q_{(i)}, r_{(i)})$ $(i = 1,2,\cdots,n)$; if they are roughly in a straight line, then the student residual satisfies the normal distribution, if not, it is not satisfied.

Similarly, if the random variable r_i satisfies the following probability distribution law, we can also conclude that the student residual satisfies the normal distribution. (Table 2)

Table 2. The frequency inspection of student residual normality

$r_i \sim N(0,1)$	(-1,1)	(-1.5,1.5)	(-2,2)
P	0.68	0.87	0.95

Step 3: If the student residual satisfies the normal distribution, select the appropriate residual fitting function, replace the unknown function G, and eliminate the local disturbance caused by the residual. Generally, if the student residual satisfies the normal distribution, we select the Gaussian function, that is,

$$g(\xi_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r_i-\mu)^2}{2\sigma^2}} \tag{3}$$

where r_i is the student residual, and μ and σ are, respectively, defined as the sample mean and sample standard deviation operated by r_i;

Step 4: Let $g(\xi_i)$ into system (1); conduct transposition processing; obtain the improved semi-parametric model

$$Y_i - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r_i-\mu)^2}{2\sigma^2}} = X_i^T \beta + \varepsilon_i \quad (i = 1,2,\cdots,n) \tag{4}$$

Solving system (4), estimate the parameter $\hat{\beta}_i$;

Step 5: Build the semi-parametric forecasting model

$$Y_{t+1} = X_{t+1}\Omega + G_{t+1} + \varepsilon \quad t = 0,1,2,\cdots,n \tag{5}$$

4.4 Forecasting Results Correction Based on Interval Information Aggregation

Using system (5), we can easily calculate the predictive value of electricity consumption at any time node. To further improve the prediction accuracy and to make the forecasting process more in line with the characteristics of the annual electricity consumption in China, this paper uses differential element theory to set the information aggregation intervals. We use an improved semi-parametric forecasting model to calculate the interval predictive function, and we define the rational weight to the interval predictive function through dynamic weight distribution and correct the forecasting results. Finally, a more accurate and meaningful electricity consumption prediction will be given.

Next, we will provide the specific realisation steps of the above idea.

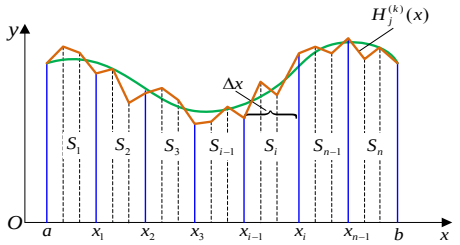

Fig.5. Information aggregation intervals set using differential element theory

(1) Segmentation

Subdividing S into n strips $S_1, S_2, \cdots, S_n$ of equal time intervals (see figure 5), the width of the time interval $[a,b]$ is $b-a$, so the width of each of the n strips is

$$\Delta x = \frac{b-a}{n}$$

These strips divide the interval $[a,b]$ into n subintervals

$$[x_0, x_1], \ [x_1, x_2], \ [x_2, x_3], \cdots, [x_{n-1}, x_n]$$

where $x_0 = a$ and $x_n = b$. The right endpoints of the subintervals are

$$x_1 = a + \Delta x,$$
$$x_2 = a + 2\Delta x,$$
$$x_3 = a + 3\Delta x,$$
$$\vdots$$

(2) Summation

The area of the region S that lies under the graph of the function $H_i^{(j)}(x)$ is the area sum of the infinite differential curve trapezoid

$$S = \sum_{j=1}^{p}\sum_{k=1}^{q} \left| \int_0^1 H_j^{(k)}(x)dx \right| \tag{6}$$

where p, q are constants and $p, q = 1,2,\cdots,n$.

(3) Dynamic Weight Distribution

From formula (6), we can obtain the areas of each of the n strips S_i, $(i = 1,2,\cdots,n)$. Then, the weight can be defined as

$$\varpi_i = \frac{S_i}{S} = \frac{\sum_{j=1}^{p} \left| \int_0^1 H_j^{(k)}(x)dx \right|}{\sum_{j=1}^{p}\sum_{k=1}^{q} \left| \int_0^1 H_j^{(k)}(x)dx \right|} \tag{7}$$

where k is a constant, and $k = 1,2,\cdots,n$.

5 Case Study

The main goal of this study is to predict electricity consumption in China using the improved semi-parametric regression model. We first present an empirical illustration of China's annual electricity consumption forecasting to examine the performance of our semi-parametric regression approach. Because the reform of 1978 significantly altered the economic development of China, we take 1981, the opening year of the "sixth five-year plan", as the time division point. Therefore, we use the annual electricity consumption data after 1981 in our paper: the 1981-2005 data (from the sixth five-year plan to the tenth five-year plan) for model building and the 2006-2010 (the eleventh five-year plan) data as testing data [21].

Improving the prediction accuracy is one of the main tasks in establishing a prediction model. However, in any

type of forecasting method, it is essential to determine the prediction error; therefore, how to control the prediction error and thus provide feedback to the forecasting technique is an important task. In this paper, we give three statistical measures to evaluate the prediction accuracy of our approach: the mean absolute error (MAE), mean absolute deviation (MAD) and the mean squared error (MSE). The MAE was used to measure the forecasting accuracy of the method; it usually expresses accuracy as a percentage and can also be written the as the mean absolute percentage error (MAPE). MAD and MSE are two measures of the average errors. The three measures are defined as follows:

(8) $$MAPE(\%) = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}(i) - y(i)|}{y(i)}$$

(9) $$MAD = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}(i) - y(i)|$$

(10) $$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

where $\hat{y}_i$ and y_i represent the forecasted and observed values, respectively.

When the semi-parametric regression forecasting approach is used to model and predict China's annual electricity consumption, we first standardise the electricity consumption data and the impact factor data from 1981 to 2010. Using the method described in section 4, we can easily obtain the standardised data.

Next, we establish the multiple linear regression model. We first calculate the fitted values $\hat{y}_i$, the residual $\hat{\varepsilon}_i$ and the student residual r_i as follows (Table 3).

Table.3. Residual value table for 1981-2009

Year	y_i	$\hat{y}_i$	$\hat{\varepsilon}_i$	r_i
1981	-0.9467	-0.8621	-0.0846	-1.5323
1982	-0.9283	-0.8566	-0.0717	-1.2968
1983	-0.9043	-0.8495	-0.0548	-0.9901
1984	-0.8784	-0.8341	-0.0443	-0.7999
1985	-0.8443	-0.8106	-0.0337	-0.6058
1986	-0.8131	-0.7984	-0.0147	-0.2639
1987	-0.7656	-0.7776	0.0120	0.2146
1988	-0.7200	-0.7508	0.0308	0.5510
1989	-0.6795	-0.7286	0.0491	0.8747
1990	-0.6431	-0.6875	0.0444	0.7945
1991	-0.5858	-0.6495	0.0637	1.1400
1992	-0.5098	-0.6063	0.0965	1.7152
1993	-0.4351	-0.5637	0.1286	2.2790
1994	-0.3504	-0.3915	0.0411	0.7251
1995	-0.2662	-0.2966	0.0304	0.5450
1996	-0.1976	-0.2401	0.0425	0.8075
1997	-0.1507	-0.1771	0.0264	0.5095
1998	-0.1198	-0.0982	-0.0216	-0.4034
1999	-0.0451	-0.0185	-0.0266	-0.4956
2000	0.0926	0.1255	-0.0329	-0.6100
2001	0.2145	0.2486	-0.0341	-0.6554
2002	0.3852	0.4261	-0.0409	-0.8103
2003	0.6363	0.6449	-0.0086	-0.1611
2004	0.9239	0.9810	-0.0571	-1.1653
2005	1.2173	1.2283	-0.0110	-0.2329
2006	1.6862	1.6205	0.0657	1.3746
2007	1.9959	1.9798	0.0161	0.8203
2008	2.1773	2.2063	-0.0290	-0.9220
2009	2.4105	2.4071	0.0034	0.3676
2010	2.5456	2.6079	0.0358	1.2896

Next, we test the distribution of the student residuals by the normal Q-Q plot test. If the student residuals satisfy the normal distribution, we select an appropriate function to replace the unknown function G, and we eliminate the local disturbance of the forecast process.

Using the method given in section 4.3 (b) for data normality inspection, we can draw a Q-Q scatter diagram for Fig. 6. We can see from Fig. 6 that the scatter points are approximately in a straight line, which means the student residuals satisfy the normal distribution.

Similarly, we can also verify the above result with the frequency inspection in Table 2. By frequency analysis of the student residuals in Table 4, we can see that 73.3% ($22/30 = 0.733 \approx 0.68$) of the $r_i(i = 1,2,\cdots,30)$ falls within the interval (-1,1), 86.6% ($26/30 = 0.867 \approx 0.87$) falls within the interval (-1.5,1.5) and 96.6% ($29/30 = 0.967 \approx 0.95$) falls within the interval (-2,2).

After we verified the distribution of the student residuals, we calculated the non-parametric part G of the forecasting model. From system (3), we know that

$$g(\xi_i) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(r_i-\mu)^2}{2\sigma^2}}$$

and $G = (g_1,\cdots,g_d) = (g(\xi_1)^T,\cdots,g(\xi_n)^T)^T$, so we have

(11) $$G = (\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(r_i-\mu)^2}{2\sigma^2}})^T, \quad i = (1,2,\cdots,30)$$

here, with a sample mean of $\mu = 0.007583$ and $\sigma = 0.986877$.

Let μ and σ into system (11). Next, we can easily obtain the value

(12) $$Y_i - \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(r_i-\mu)^2}{2\sigma^2}}, \quad (i = 1,2,\cdots,30)$$

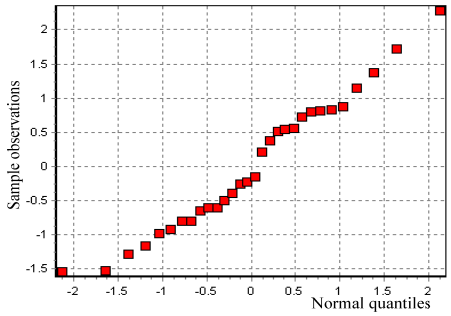

Fig.6. Q-Q scatter diagram

Next, we set the information aggregation intervals. The setting standard of the width of information aggregation interval was in accordance with China's five-year development plan. From the starting year of the data sample to the terminal year, we set a total of six intervals, corresponding to the sixth five-year plan (from 1981 to 2010). This kind of division satisfies the characteristics of China's stage of development, makes certain the objectivity of the factor selection in each interval, and it benefits the analysis of the causes of electricity consumption while improving the accuracy of the electricity consumption forecast. The expressions are as follows:

(13) $$Y_t = f(P, X_t, t)$$

Here, X is a vector composed of related factors, P is a parameter vector of the prediction model, Y is the value to be predicted, t is the time number, and $t_1 = [1981,1985]$, $t_2 = [1986,1990]$, $t_3 = [1991,1995]$, $t_4 = [1996,2000]$, $t_5 = [2001,2005]$, $t_6 = [2006,2010]$.

The prediction functions of each information aggregation interval are as follows:

$$Y_1 = f(P, X_1, t_1) = 0.47632825 - 0.01217365GDP + 0.50841970TEIV + 1.55220177IAV$$

$$Y_2 = f(P, X_2, t_2) = -0.62129011 + 0.48820799GDP + 0.39332524IFA - 0.90085495IAV$$

$$Y_3 = f(P, X_3, t_3) = -0.04189416 + 0.17379825GDP + 0.78314703IAV + 0.12151406DI$$

$$Y_4 = f(P, X_4, t_4) = 2.23773325 + 3.35757076IFA + 6.96624885IAV - 6.83689285DI$$

$$Y_5 = f(P, X_5, t_5) = 0.26536488 + 0.14856407GDP + 0.89119562TEIV - 0.38975257IAV$$

$$Y_6 = f(P, X_6, t_6) = 0.06661806 + 0.86728505GDP - 0.25848989IFA + 0.34781584DI$$

Next, we determine the rational weight distribution of the prediction functions of each information aggregation interval. To standardise the operation process, we set the integral interval as $[0,1]$. Then, when $t \in t_1$,

$$S_1 = \sum_{j=1}^{4} \left| \int_0^1 H_j^{(1)}(x)dx \right|$$

where $H^{(1)}(x)$ is the growth function of electricity consumption $\tilde{y}_i$, and,

$$H_1^{(1)}(x) = -0.0555x - 1.0868;$$

$$H_2^{(1)}(x) = -0.0751(x - 0.25) - 1.1006;$$

$$H_3^{(1)}(x) = 0.0033(x - 0.5) - 1.1194;$$

$$H_4^{(1)}(x) = 0.0452(x - 0.75) - 1.1186.$$

Then, we can obtain

$$S_1 = \sum_{j=1}^{4} \left| \int_0^1 H_j^{(1)}(x)dx \right| = 1.1089$$

Similarly, when $t \in t_i \quad (i = 2,3,4,5,6)$, we can obtain S_2, S_3, S_4, S_5, S_6.

$$S_2 = \sum_{j=1}^{4} \left| \int_0^1 H_j^{(2)}(x)dx \right| = 1.0042;$$

$$S_3 = \sum_{j=1}^{4} \left| \int_0^1 H_j^{(3)}(x)dx \right| = 0.6167;$$

$$S_4 = \sum_{j=1}^{4} \left| \int_0^1 H_j^{(4)}(x)dx \right| = 0.3694;$$

$$S_5 = \sum_{j=1}^{4} \left| \int_0^1 H_j^{(5)}(x)dx \right| = 0.4385;$$

$$S_6 = \sum_{j=1}^{4} \left| \int_0^1 H_j^{(6)}(x)dx \right| = 2.0321$$

Thus, from formula (7), we can calculate the value of weight $\omega_i \ (i = 1,2,\cdots,6)$.

Therefore, we present the final semi-parametric prediction model with variable weight

(14)
$$\tilde{Y}_t = \omega_1 f(P, X_1, t_1) + \omega_2 f(P, X_2, t_2) + \omega_3 f(P, X_3, t_3) + \omega_4 f(P, X_4, t_4) + \omega_5 f(P, X_5, t_5) + \omega_6 f(P, X_6, t_6)$$
$$= 0.17172886 + 0.43291718GDP + 0.17136345TEIV + 0.19924826IFA + 0.66450236IAV - 0.31295116DI$$

where

$$\tilde{Y}_t = Y_t - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r_i - \mu)^2}{2\sigma^2}}$$

Thus, we can use system (14) to conduct an electricity consumption forecasting study. We also apply the GM (1, 1) and ANN models for comparison purposes. In the case of the GM (1,1) model, the resulting model is

$$x(t+1) = 18882.9496e^{-0.099067t} + 15876.6496, t = 1,2,3,\cdots.$$

Table 4 shows the forecasted values as well as the relative errors (REs) for the three methods.

Measures of the corresponding forecasting errors are shown in Table 5. Both in the model building stage and in the testing stage for this particular case, the SPRM prediction approach outperforms the GM (1,1) and ANN models. Fig. 7 shows the model percentage error distributions for the SPRM prediction approach. In this figure, calibrations 1 to 25 correspond to the model building stage, and calibrations 26 to 30 correspond to the testing stage.

Table 4. Observed and forecasted electricity consumption[a] in China, 1981-2010, for three different approaches.

Year	Observed value y_i	Observed value $\tilde{y}_i$	GM(1,1) FV	GM(1,1) RE(%)	ANN FV	ANN RE(%)	SPRM[b] FV	SPRM[b] RE(%)
Model building Stage: 1980-2005								
1981	-0.9467	-1.0867	-1.0184	-7.57	-1.0333	-21.06	-1.1075	1.91
1982	-0.9283	-1.1006	-0.9968	-7.38	-1.0059	-24.81	-1.0942	-0.58
1983	-0.9043	-1.1194	-0.9731	-7.61	-0.9811	-25.24	-1.0783	-3.67
1984	-0.8784	-1.1185	-0.9468	-7.79	-0.9427	-29.47	-1.0564	-5.55

1985	-0.8443	-1.1072	-0.9178	-8.71	-0.9063	-29.92	-1.0136	-8.45
1986	-0.8131	-1.1056	-0.8858	-8.94	-0.8729	-30.31	-1.0241	-7.37
1987	-0.7656	-1.0685	-0.8505	-11.08	-0.8435	-25.28	-1.0101	-5.46
1988	-0.7200	-1.0067	-0.8115	-12.71	-0.7972	-25.51	-0.9663	-4.01
1989	-0.6795	-0.9350	-0.7684	-13.08	-0.7365	-30.02	-0.8587	-8.16
1990	-0.6431	-0.9076	-0.7208	-12.08	-0.6767	-34.61	-0.8313	-8.40
1991	-0.5858	-0.8081	-0.6684	-14.12	-0.6091	-35.88	-0.7604	-5.90
1992	-0.5098	-0.6525	-0.6104	-19.73	-0.5423	-33.35	-0.6856	5.07
1993	-0.4351	-0.5117	-0.5464	-25.58	-0.4880	-29.58	-0.5521	7.89
1994	-0.3504	-0.6219	-0.4760	-35.84	-0.4502	-28.48	-0.5524	-11.17
1995	-0.2662	-0.5533	-0.3981	-49.54	-0.4253	-59.76	-0.5158	-6.77
1996	-0.1976	-0.4606	-0.3121	-57.94	-0.3666	-85.52	-0.4452	-3.34
1997	-0.1507	-0.4404	-0.2171	-44.06	-0.2587	-71.66	-0.4701	6.74
1998	-0.1198	-0.4021	-0.1122	6.34	-0.1633	-36.31	-0.3728	-7.28
1999	-0.0451	-0.3192	0.0036	107.98	-0.0300	33.48	-0.2802	-12.21
2000	0.0926	-0.1699	0.1315	-42.01	0.1672	-80.56	-0.1627	-4.23
2001	0.2145	-0.0429	0.2726	-27.08	0.3932	-83.31	-0.0446	3.96
2002	0.3852	0.1463	0.4285	-11.24	0.6253	-62.33	0.1372	-6.22
2003	0.6363	0.3382	0.6006	5.61	0.9947	-56.32	0.3282	-2.95
2004	0.9239	0.7331	0.7907	14.41	1.2448	-34.73	0.7531	2.72
2005	1.2173	0.9229	1.0005	17.80	1.3831	-13.62	0.9781	5.98
Testing Stage: 2006-2010								
2006	1.6862	1.4963	1.2323	26.92	1.3314	24.54	1.4658	-2.03
2007	1.9959	1.7342	1.4879	25.45	1.7004	29.81	1.7409	0.38
2008	2.1773	1.9529	1.7701	18.70	1.8656	-6.07	2.0194	3.40
2009	2.4105	2.1126	2.0817	13.64	2.1001	-8.42	2.1904	3.68
2010	2.5456	2.2477	2.4257	4.71	2.5276	4.97	2.3469	4.41

Remarks: [a]The electricity consumption values are standardised data;
[b]The proposed semi-parametric regression model in this paper.
FV: forecasted value.

Table 5. Comparative analysis of forecasting erro

Models	MAPE(%)	MAD	MSE
			Model building Stage: 1980-2005
GM(1,1)	-13.07	0.0801	0.0083
ANN	-3.84	0.1082	0.0189
SPRM	-3.21	0.0401	0.0023
			Testing Stage: 2006-2010
GM(1,1)	17.88	0.3636	0.1505
ANN	12.75	0.2581	0.0814
SPRM	2.79	0.0094	0.0043

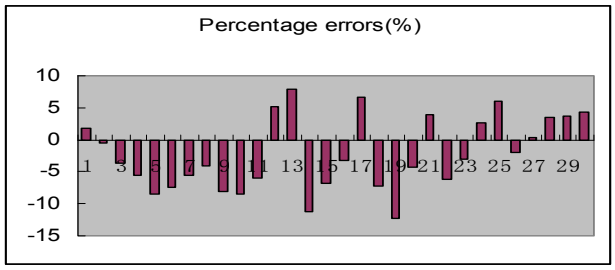

Fig. 7. Percentage errors for the SPRM approach.

Nomenclature

The notations used throughout the paper are stated below:

$\hat{\alpha}$	Estimator of the parameter α
A^T	Transpose of A
$x(t)$	Value of influence factors at time t
y	Electricity consumption function
$N(\mu,\sigma^2)$	Normal distribution function
EC	Electricity consumption
GDP	Gross domestic product
TEIV	Total import and export volume
IFA	Investment in fixed assets
IAV	Industrial added value
DI	Disposable income
M-LTCF	Mid-long-term consumption forecasts
STCF	Short-term consumption forecasting

6 Conclusion

Mid-long term electricity consumption forecasting of a power system is a complicated task because the consumption is affected directly or indirectly by various factors primarily associated with the regional distribution, climate and economy. For a long time, many prediction methods have been proposed in an attempt to improve the forecast accuracy, but no one forecasting method is applicable to all situations.

The major contribution of this paper is the proposal of a new statistical methodology to forecast electricity consumption. The proposed semi-parametric regression models, which are an integration of parametric and nonparametric regression models, capture the complex cooperative relationship between electricity consumption and its drivers. By analysing the distribution characteristics of the student residuals, we introduce a corresponding distribution function, and we use it as the non-parametric part of this semi-parametric regression model. In addition, we use differential element theory to calculate the value of the weight assigned to each piecewise prediction function, thereby eliminating the local disturbance of the forecast process and effectively reducing the prediction error or other system errors. The forecasting results demonstrate that the model performs remarkably well, and they demonstrate the effectiveness and reliability of our approach.

Some areas for possible future improvement include the following:

•Accuracy is an important index to measure the ability of forecasting methods, and irregular data will greatly affect the prediction accuracy; therefore, it is necessary to establish a more comprehensive data pre-processing mechanism.

•Weight determination methods of low computational complexity and strong applicability will be sought. This paper uses differential element thought to solve the model weight, which can better reflect the predictions, but the solution process is complicated. We will explore a new weight determination method to improve the efficiency and applicability of the forecasting model.

Acknowledgements

The authors would like to thank the China Key Laboratory of Process Optimization and Intelligent Decision-Making for their valuable comments and feedback regarding this research study. This paper was supported by the National Natural Science Foundation of China Grant No.71101041 and No.71071045, National 863 project Grant No. 2011AA05A116, Foundation of Higher School Outstanding Talents Grant No. 2012SQRL009 and National Innovative Experiment Program No.111035954.

Finally, we are grateful to the many editors who gave their attention to this paper.

REFERENCES

[1] Vincenzo Bianco, Oronzio Manca, Sergio Nardini, Alina A. Minea, Analysis and forecasting of nonresidential electricity consumption in Romania, Applied Energy. 87(2010), 3584-3590.

[2] Wang XJ, Yang SL, Ding J, Wang HJ. Dynamic GM (1,1) model based on cubic spline for electricity consumption prediction in smart grid,China Communications. 7(2010), No.4, 83-88.

[3] Diyar Akay, Mehmet Atak. Grey prediction with rolling mechanism for electricity demand forecasting of Turkey, Energy. 32(2007), 1670-1675.

[4] Wang XJ, Shen JX, Yang SL. Application research on Gaussian orthogonal interpolation method for electricity consumption forecasting of smart grid, Power System Protection and Control.38(2010),No.21,141-145,151.

[5] Wang XJ, Yang SL etc. Simulation of Orthogonalization Prediction Based on Grey Markov Chain for Electricity Consumption,Journal of System Simulation. 22(2010), No.10, 2253-2256.

[6] Roula Inglesi. Aggregate electricity demand in South Africa: Conditional forecasts to 2030, Applied Energy. 87(2010),197-204.

[7] Ching-Lai H, Marnont A, Simon W, Shanti M. Analyzing the impact of weather variables on monthly electricity deman, IEEE Trans. 20(2005), 2078-2085.

[8] Egelioglu F, Mohamad AA, Guven H. Economic variables and electricity consumption in Northern Cyprus, Energy. 26(2001), 355-362.

[9] Wei L, Wu J, Liu Y. Long-term electricity load forecasting based on system dynamics, Automation of Electric Power Systems. 24(2000), No. 24, 47.

[10] Paresh Kumar Narayan, Arti Prasad. Electricity consumption-real GDP causality nexus: Evidence from a bootstrapped causality test for 30OECD countries, Energy Policy. 36(2008), 910-918.

[11] K. Nikolopoulos, P. Goodwin, A. Patelis, V. Assimakopoulos. Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches, European Journal of Operational Research. 180(2007), 354-368.

[12] Abdel-Aal RE, Al-Garni AZ, Al-Nassar YN. Modeling and forecasting monthly electric energy consumption in eastern Saudi Arabia using abductive networks, Energy. 22(1997), 911-921.

[13] Yan YY. Climate and residential electricity consumption in Hong Kong, Energy. 23(1998), 17-20.

[14] M.R. AlRashidi, K.M. EL-Naggar. Long term electricity load frecasting based on particle swarm optimization, Applied Energy. 87(2010), 320-326.

[15] Nasr GE, Badr EA, Younes MR. Neural networks in forecasting electricity energy consumption: Univariate and multivariate approaches. International Journal of Energy Research. 26 (2002), 67-78.

[16] Metaxiotis K, Kagiannas A, Askounis D, Psarras J. Artificial intelligence in short term electricity load forecasting: a state of the art survey for the researcher, Energy Conversion and Management. 44(2003), 1525-1534.

[17] Santos PJ, Martins AG, Pires AJ, Martins JF, Mendes RV. Short term load forecast using trend information and process reconstruction, International Journal of Energy Research. 30(2006), 811-822.

[18] Robert F. Engle, C.W.J. Granger, John Rice, Andrew Weiss. Semi-parametric estimates of the relation between weather and electricity sales, Journal of the American Statistical Association. 81(1986), 310-320.

[19] National Bureau of Statistics of China, 60 Years of New China, China Statistics Press. 2009.

[20] J.F.Liu, Z.L.Deng, Self-Tuning Weighted Measurement Fusion Kalman Filter for ARMA Signals with Colored Noise, Applied Mathematics & Information Sciences, 6(2012), 1-7.

[21] Kumar Krishen, Applications of Space Technologies to Commercial Sector, Advances in Industrial Engineering and Management, 1(2012), No.1, 1-9