FORECASTING MODELING AND SIMULATION ANALYSIS OF A POWER SYSTEM IN CHINA, BASED ON A CLASS OF SEMI-PARAMETRIC REGRESSION APPROACH

Xiaojia Wang[1*] , Zhiqiang Chen[2] & Shanlin Yang[3]

[1,2,3] China Key Laboratory of Process Optimization and Intelligent Decision-Making, School of Management, HeFei University of Technology, Hefei, Anhui, China
[1] tonysun800@sina.com; [2] hfut211@163.com; [3] hgdysl@gmail.com

ABSTRACT

Forecasting electricity consumption is one of the most important challenges in electricity system planning. This paper presents an improved semi-parametric regression model using the Student distribution function of residual to replace the nonparametric component of the traditional semi-parametric model, thus eliminating the effects of the residual disturbance term. Compared with general linear models, the models make statistical inferences and can automatically regulate the boundary effect, which gives the forecast result a higher accuracy. A case study using data from China is presented to demonstrate the effectiveness of the approach.

OPSOMMING

Die vooruitskatting van elektrisiteitverbruik is een van die belangrikste uitdagings in elektrisiteitstelselbeplanning. Dié artikel bevat 'n verbeterde, semi-parametriese regressie-model, wat gebruik maak van die Studentverdelingsfunksie van residuee om die nie-para-metriese komponent van die tradisionele semi-parametriese model te vervang, en sodoende die effekte van die residuversteuringsterm uit te skakel. In vergelyking met algemene lineêre modelle, kan die model statistiese afleidings maak en outomaties die grenseffek reguleer, wat lei tot groter akuraatheid van die vooruitskatting. 'n Gevallestudie wat gebruik maak van data van China demonstreer die effektiwiteit van die benadering.

[1,2,3] The author was enrolled for a PhD (Engineering Management) degree in the School of Management, Hefei University of Technology.

* Corresponding author

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

1. INTRODUCTION

In recent decades, the total consumption of electricity in China has undergone a sustained and significant increase. According to official figures [1], during the period from 1980 to 2009, the annual rates of variation ranged from 2.97% (1980/81) to 6.79% (2008/09), while the electricity consumption in 2009, at 3.6595×10^{12} kwh, was 1117% higher than in 1980. After the United States, China has the second largest electricity consumption rate in the world. Despite the relationship between the growth of electricity demand and certain social, economic, and policy factors, the pace of change between them is sometimes not consistent. There have been widespread electricity shortages throughout the last 60 years of Chinese history. During the 1997 Asian financial crisis, China experienced an electricity surplus for the first time. However, electricity shortages again appeared in 2002 and worsened in 2004. In 2004, 24 provinces had power shortages, and the total gap in China was 31 GW. With the rapid growth of China's economy since 2005, there has been an electricity shortage in China almost every year.

The disharmony between electricity demand and these factors in China suggests an important task. According to the factors already obtained, the demand drivers and the fundamental pillars in building a forecasting model [2-4] both need to be determined. Furthermore, methods of predicting electricity consumption precisely, effectively, and practically also need to be created. A proper solution of these problems will help to accelerate the future development of China, and also help one to understand the power operating environment, since inaccurate consumption forecasting will increase the operating costs of utility companies.

Since there is no consensus about the best approach to forecasting electricity consumption, various methods have been developed in recent years. Generally speaking, from the classification analysis of the predictive behaviour itself, the methodology for electricity consumption forecasting can be divided into three categories: numerical approximation class processing methods, statistical regression class processing methods, and intelligent optimisation class processing methods.

First, numerical approximation class processing methods (NACPM) rely solely on the variation of the data itself to find the information supporting predictive behaviour; they do not consider the effects of the other factors. Based on this view, many scholars have drawn a number of useful results. Wang et al. [5] investigated a dynamic GM (1,1) model based on the cubic spline function interpolation principle to forecast the electricity consumption of China. The authors used piecewise polynomial interpolation thought processing electricity consumption data to analyse the electricity consumption trends to make predictions. In references [6]-[7], Wang et al. use Gauss orthogonalisation theory to improve the grey prediction model, and, in constructing the grey combinative interpolation model to forecast the electricity consumption of China, they achieved good prediction results. In addition,

Wang also introduced Markov Chain theory to the grey combinative interpolation model, and constructed the Markov grey orthogonalisation model for electricity consumption prediction [8]-[9], which also obtained good prediction accuracy.

Second, statistical regression class processing methods (SRCPM) often consider the synergy of multiple factors that affect predictor variables to measure predictive behaviour. Statistical regression class methods are widely used for the electricity consumption forecasting problem. For example, Ching Lai [10] 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. [11] studied the influence of economic variables on the annual electricity consumption in northern Cyprus between 1988 and 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. [12] have estimated the long-term electricity load by applying system dynamics, which construct the model according to an analysis of historical electricity consumption. This method discovered the significant influence of uncertain factors such as economy and policy. Narayan and Prasad [13] studied any causal effects between electricity consumption and real GDP for 30 OECD countries, using the bootstrapped causality testing approach to show electricity consumption affecting the real GDP in Australia, Iceland, Italy, the Slovak Republic, the Czech Republic, Korea, Portugal, and the UK. They found that electricity conservation policies will negatively impact real GDP in the eight countries mentioned above and, for the remaining 22 countries of the OECD, the electricity conversation policies will not affect real GDP. Nikolopoulos et al. [14] compared multiple linear regression (MLR) with the artificial neural network, nearest neighbour analysis, and human judgment; the application results showed that the MLR was less accurate than other methods as a result of its inability to handle complex non-linearities in the relationship between the dependent variable and the cues, as well as its tendency to misaddress the in-sample data. Abdel-Aal et al. [15] applied an abductory induction mechanism (AIM) 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 about 5-6% over the year. Yan [16] also presented residential consumption models using climatic variables for Hong Kong.

Third, 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 programmes. Research on this method (IOCPM) provides new and useful ideas for predicting behaviour itself. Nasr et al. [17] presented an Artificial Neural Networks (ANN) approach to electrical energy consumption forecasting in Lebanon. Four ANN models are presented and implemented in the research: 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. Niu et al. [18] used a particle swarm optimisation (PSO) algorithm to predict the electricity load in China. The PSO algorithm was adopted to solve the disturbance vector α , as it has the virtue of optimum-seeking. Metaxiotis [19] provides an overview of the studies examining Artificial Intelligence (AI) technologies, as well as their current use in the field of short-term electrical load forecasting. Santos [20] has also used the ANN algorithm to make load forecasts, and with this method, the possibility of including weather-related variables in the input vector has also been analysed.

Based on the analysis of the literature above, one may refer to the researchers' experience of how they chose the factors for electricity consumption forecasting in the correlation field. Considering the actual situation of China's national conditions, after comprehensive data analysis and filter processing of electricity consumption data, we chose the following

five factors that best reflect the truth of China's electricity consumption data during the period 1980 to 2009:

(1) Gross domestic product (GDP)
(2) Total import and export volume (TEIV)
(3) Investment in fixed assets (IFA)
(4) Industrial added value (IAV)
(5) Disposable income (DI)

The factors chosen above are the main reasons for complexity and periodic shape. They reflect the status of China's current development in accordance with its national conditions.

The remainder of the paper is organised as follows: Section 2 introduces an overview of electricity consumption in China. Section 3 discusses the methodology and the data of the study, and provides an accurate model for electricity consumption forecasting. Case analysis and results comparisons are used in Section 4, leading to the conclusion in Section 5.

2. OVERVIEW OF ELETRICITY CONSUMPTION IN CHINA

With the rapid development of China's economy, total electricity energy consumption increases sharply. The changes in electricity supply and demand in China since 1980 can be described in three stages. During the first stage, from 1978 to 1996, electricity consumption grew steadily by 7% per annum. Stage two stretched from 1997 to 2000 and, with the influence of the Asian financial crisis, electricity consumption grew slowly. However, from 2001 electricity consumption increased rapidly by 15% per annum, keeping pace with China's economic development.

Electricity consumption, economic growth, and environmental constraints interact in a dynamic way. Continued growth in electricity consumption and the enhancement of environmental factors have led the transformation of economic development in China by promoting industrial structure reforms and improving electricity availability. At present, China's power structure is mainly dominated by thermal power. Coal consumption for energy accounts for 50% of the total national coal consumption. The rapid growth of electricity consumption led to the rapid growth of coal consumption, which increased environmental pollution. It is therefore necessary to improve the accuracy with which electricity consumption is predicted to obtain a more accurate understanding of future environmental pollution. This predictive ability can provide policymakers with more accurate information for the development of relevant policies, and eventually achieve the goal of having reduced the 2020 carbon dioxide emissions per unit of GDP by about 40-45% when compared with 2005.

3. METHODOLOGY AND DATA

3.1 Data sets

China is at a critical stage of its economic development, which is China's first priority. Therefore we select the mainly economic factors that affect electricity consumption because, in this context, indicators of economic performance can better reflect the trends and levels of electricity consumption.

So, how can one discover suitable economic indicators? One knows that the volatility of GDP continuously influences the trend of electricity consumption; therefore, the GDP value can be seen as one indicator. At the same time, the 'troika' of total import and export volume (TEIV), investment in fixed assets (IFA), and disposable income (DI) can also accurately describe the trends of China's economic growth. Thus these three indicators can be included in the indicator system, since they are representative and rational.

Furthermore, industrial production is an important component of economic production in China's current industrial structure. Industrial electricity consumption accounts for a large proportion of total electricity consumption – generally 70% or more. On the one hand, industrial production creates huge economic benefits; but on the other, it consumes a large amount of electricity resources. Therefore the industrial added value (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. Not only do these indicators reflect the true background of China's power consumption, but their inclusion also enhances the integrity of the index selection.

For the period 1980-2009, the annual figures for electricity consumption were obtained by the National Bureau of Statistics of China in its survey called 60 Years of New China Statistical Data Compilation.

The annual data for the GDP, TEIV, IFA, IAV, and DI for the same period were also taken by the 60 Years of New China Statistical Data Compilation.

The historical data of electricity consumption are reported in Figure 1, and the independent variables (i.e., GDP, TEIV, IFA, IAV, and DI) are presented in Figure 2.

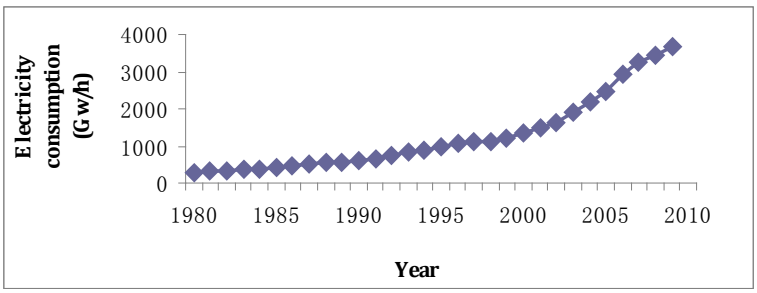

Figure 1: Historical data for electricity consumption in China

In Figure1, electricity consumption shows a trend of substantial linear growth. In 1997 a marked decrease in the electricity consumption was detected. The growth rate of electricity consumption at 2.79% (1997/98) was much lower than the average growth rate of 9.06%, probably due to the Asian financial crisis of that period.

Generally, the trends of the other five factors are consistent with electricity consumption, but each of the factors also presents differing characteristics. From Figure 2 (a), the GDP trend of China maintains a sustained, significant increase; an inflection point only appears in 1997, and the period of inflection terminated in 2004, explained by the Asian financial crisis. Interestingly, the level of GDP growth has been slightly greater than the growth rate for electricity consumption, while the other four factors (TEIV, IFA, IAV, DI) are more closely associated with electricity consumption than with GDP. In Figure 2, the trend of TEIV, IFA, IAV, and DI only keeps pace with electricity consumption from 1993, after which their rates start a modest rise, slightly higher than consumption. It is worth noting that the TEIV value had a clear downward trend in 2008 (dropping by 16.27% on link relative ratio), owing to the global financial crisis; but it did not appear to impact electricity in China significantly. From Figure 2, one can garner information on two aspects. First, these characteristics show that since 1993, the growth rate of China's economy has increased quickly, and the economic structure is increasingly diversified. Second, there was a strong relationship between the aforementioned factors and electricity consumption. The trend comparison chart of electricity consumption and the factors are presented in Figure 3.

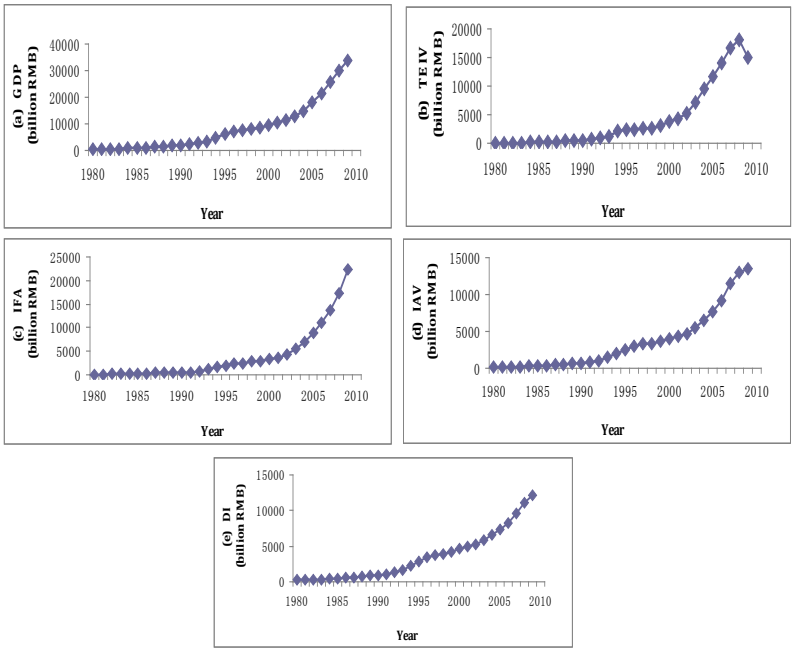

Figure 2: Historical data for the variables: (a) GDP; (b) TEIV; (c) IFA; (d) IAV; (e) DI

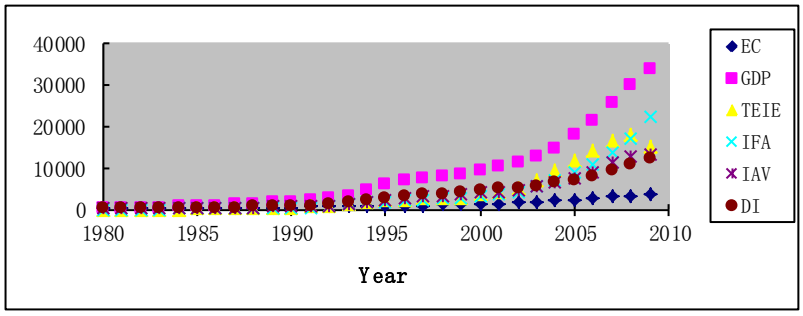

Figure 3: Trend comparison of electricity consumption and the five factors

3.2 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, standardised methods for data standardisation are used for two reasons. Initially they eliminate the variation of the difference of each variable when making dimensionless processing. Second, they consider the distribution of original data, which is what is required to establish the semi-parametric forecasting model. The calculation method is as follows:

$$\varphi = \frac{x - \mu}{\sigma}$$

where X is raw data to be standardised, $\mu = E[x]$ is the mean value, and $\sigma = \sqrt{Var(x)}$ is the standard deviation of the raw data.

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

3.3 Build the semi-parametric prediction model

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

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. The first is a linear relation; the second is a certain interference factor in which the relation to observation values is completely unknown, causing it to fall under the error item without any reason. In this case, too much information will be lost if the non-parametric model is used (though it has greater flexibility); 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 $Y_i = (y_{i1}, \cdots, y_{id})$ are observations, or historical electricity consumption, and $X_i = (x_{i1}, \cdots, x_{ip})^T$ are explanatory variables, or indicators. The errors $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{id})$ are assumed to be independent and identically distributed (denoted as iid). $\Omega = (\beta_1, \cdots, \beta_d)$ is the $p \times d$ matrix of unknown parameters, and $g(\xi) = (g_1(\xi), \cdots, g_d(\xi))$ is the $1 \times d$ vector of unknown functions. In this paper, the distribution function of Student residuals replaces the unknown function $g(\xi)$. For simplicity, let

$$Y = (y_1, \cdots, y_d) = (Y_1^T, \cdots, Y_n^T)^T \quad ; \quad X = (x_1, \cdots, x_p) = (X_1, \cdots, X_n)^T$$

$$G = (g_1, \cdots, g_d) = (g(\xi_1)^T, \cdots, g(\xi_n)^T)^T \quad ; \quad \varepsilon = (\varepsilon_1, \cdots, \varepsilon_d) = (\varepsilon_1^T, \cdots, \varepsilon_n^T)^T$$

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 [21]). Because it not only contains the parameter weight (which describes the known composition of function relation in observation values) but also contains the non-parameter weight (which exclusively shows the model deviation that is unknown in function relation), the model can generalise and describe numerous actual problems, bringing it closer to reality.

In this sub-section, the prediction principle diagram based on the semi-parametric multiple regression model is provided to analyse the forecasting process, which has multiple impact factors. Subsequently the specific steps on how to build the improved semi-parametric prediction model are given.

3.3(a) Prediction principle diagram

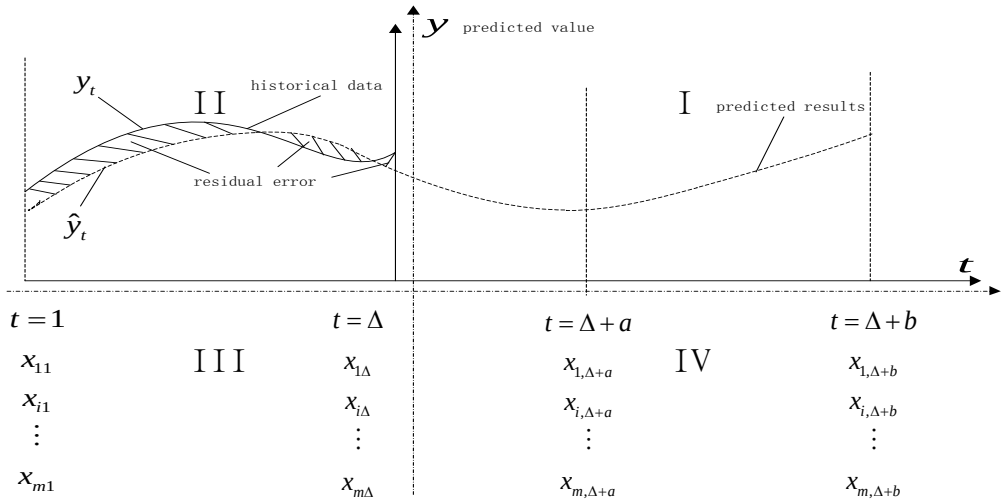

Figure4: Prediction principle diagram of semi-parametric model with multiple impact factors

In Figure4, the tagging below the horizontal axis are the factors for each time period. Assuming that one has collected M kinds of factors associated with the predictive object y, one denotes $X=[x_1,x_2,\cdots,x_m]$. Supposing $X_t=[x_{1t},x_{2t},\cdots,x_{mt}]$ at the historical time period t $(1\le t\le\Delta)$ that the amount of value to be predicted is y_t, we need to predict the future at the time period $t\in[\Delta+a,\Delta+b]$ under the law of historical development. In Figure 4, if the time axis is the horizontal axis, and if one considers the vertical line with the current time point as the vertical axis, then Figure 4 can be regarded as a two-dimensional coordinate system with the time point 'present' as the coordinate origin. Thus Figure 4 can be divided into four quadrants, I to IV. Therefore, from Figure 4 one finds that the implication of semi-parametric regression forecasting is as follows. First, use the data of quadrants II and III to precede a historical fitting operation and derive the forecasting model. Next, use the data of quadrant IV as the input of the forecasting model, thus obtaining the forecasting result of quadrant I.

3.3(b) Modelling steps

Step 1: By establishing the multiple linear regression method and solving the parameter part $Y=X\Omega$, obtain $\hat{Y}$, the estimated value of Y;

Step 2: List the fitting residuals, calculate the standardised residuals and Student residual, make a distribution test on the Student residual, and draw the Q-Q plot, observing whether it satisfies the normal distribution. The specific process is as follows:

(1) Calculate the Student residual r_i

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

Where $\hat{\varepsilon}_i$ is the residual vector and $\hat{\varepsilon}_i \sim N(0, \sigma^2(I-H))$, $H = X(X^T X)^{-1} X^T$, lever quantity h_{ii} is the i-th element on the leading diagonal of H , MSE is the mean-square error.

(2) Normal Q-Q plot test for Student residual

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

2.2 calculate

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

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

2.3 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, the means dissatisfy.

Similarly, if the random variable r_i satisfies the following probability distribution law, one can also conclude that the Student residual satisfies the normal distribution.

Table 1: 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, μ and σ are, respectively, defined as sample mean and sample standard deviation operated by r_i ;

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

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

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

Step 5: Build the semi-parametric forecasting model

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

4. CASE STUDY

The main goal of this study is to predict electricity consumption in China using the semi-parametric regression model. We first present an empirical illustration of China's annual electricity consumption forecasting to examine the performance of the semi-parametric regression approach. Because the reforms of 1978 significantly altered the economic development mode of China, one usually takes 1980 as the time division point. Thus we use the annual electricity consumption data after 1980 in this paper, using the 1980-2005 data for model building and the 2006-2010 data as testing data.

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

$$MAPE(\%) = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}(i) - y(i)|}{y(i)} \tag{6}$$

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}(i) - y(i)| \tag{7}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{8}$$

where $\hat{y}_i$ and y_i represent the forecast 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 factors data from 1980 to 2009. Using the method mentioned in section 2, we give the following standardised data in Table 2.

Table 2: Standardised data, 1980-2009

Year	EC	GDP	TEIV	IFA	IAV	DI
1980	-0.9557	-0.8622	-0.7518	-0.6867	-0.8405	-0.9568
1981	-0.9467	-0.8585	-0.7488	-0.6858	-0.8392	-0.9483
1982	-0.9283	-0.8539	-0.7482	-0.6809	-0.8363	-0.9400
1983	-0.9043	-0.847	-0.7466	-0.6773	-0.8309	-0.9304
1984	-0.8784	-0.8337	-0.7404	-0.6701	-0.8204	-0.9155
1985	-0.8443	-0.8143	-0.7247	-0.6573	-0.8036	-0.8878
1986	-0.8131	-0.7643	-0.7212	-0.6211	-0.7727	-0.8852
1987	-0.7656	-0.7475	-0.7118	-0.6106	-0.7516	-0.8729
1988	-0.7200	-0.7192	-0.7061	-0.6042	-0.7416	-0.8462
1989	-0.6795	-0.7008	-0.6907	-0.6092	-0.7255	-0.8007
1990	-0.6431	-0.6849	-0.6778	-0.6076	-0.7071	-0.7836
1991	-0.5858	-0.6777	-0.6309	-0.6024	-0.6857	-0.7110
1992	-0.5098	-0.6226	-0.5965	-0.5577	-0.6299	-0.6445
1993	-0.4351	-0.5326	-0.5575	-0.4679	-0.5307	-0.5446
1994	-0.3504	-0.3948	-0.3920	-0.3965	-0.3962	-0.3857
1995	-0.2662	-0.2600	-0.3353	-0.343	-0.2571	-0.1949
1996	-0.1976	-0.1488	-0.3238	-0.2909	-0.1428	-0.0316
1997	-0.1507	-0.0653	-0.2723	-0.2544	-0.0545	0.0551

Year	EC	GDP	TEIV	IFA	IAV	DI
1998	-0.1198	-0.0072	-0.2745	-0.1921	-0.0267	0.1220
1999	-0.0451	0.0493	-0.2191	-0.1661	0.0202	0.1999
2000	0.0926	0.1514	-0.0488	-0.1110	0.1262	0.3147
2001	0.2145	0.2632	0.0041	-0.0337	0.2164	0.4188
2002	0.3852	0.3775	0.1711	0.0794	0.3143	0.5242
2003	0.6363	0.5433	0.5181	0.2964	0.5052	0.6581
2004	0.9239	0.7876	0.9732	0.5646	0.7661	0.8787
2005	1.2173	1.0507	1.3616	0.8937	1.0717	1.0959
2006	1.6862	1.4051	1.7985	1.2754	1.4295	1.3711
2007	1.9959	1.9350	2.2666	1.7669	2.0325	1.7650
2008	2.1773	2.3082	2.5060	2.4056	2.4195	2.2021
2009	2.4105	2.6796	1.9743	3.3367	2.5461	2.5094

Next, we establish the multiple linear regression model, which uses the standardised data from Table 2 to calculate the fitted values $\hat{y}_i$, residual $\hat{\varepsilon}_i$ and Student residual r_i .

Table 3: Residual value, 1980-2009

Year	y_i	$\hat{y}_i$	$\hat{\varepsilon}_i$	r_i
1980	-0.9557	-0.8705	-0.0852	-1.5463
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

Next, we test the distribution of the Student residuals by means of 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 eliminate the local disturbance of the forecast process.

Using the method given in section 2.3 (b) for data normality inspection, one can draw a Q-Q scatter diagram for Figure 5. One can see from Figure 5 that the scatter points are approximately in a straight line, which means that the Student residuals satisfy the normal distribution.

Similarly, we can also verify the above result by using the frequency inspection in Table 1. By frequency analysis of the Student residuals in Table 3, 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).

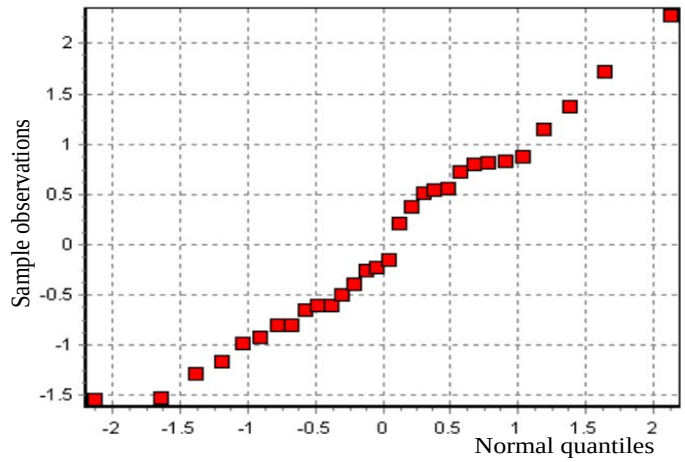

Figure 5: Q-Q scatter diagram

After verifying the distribution of the Student residuals, the non-parametric part G of the forecasting model is calculated. From system (3), it follows 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

$$G = (\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r_i-\mu)^2}{2\sigma^2}})^T, \quad i = (1, 2, \cdots, 30) \tag{9}$$

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

Let μ and σ into system (9). Next, one obtains the value

$$G = (\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r_i-\mu)^2}{2\sigma^2}})^T, \quad Y_i - \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r_i-\mu)^2}{2\sigma^2}}, \quad (i = 1, 2, \cdots, 30) \tag{10}$$

And, taking the results of system (10) into system (4), we can estimate the parameters $\Omega = (\hat{\beta}_1, \cdots, \beta_5)$ of the linear part of the semi-parametric prediction model (4).

$$\Omega = (\hat{\beta}_1, \cdots, \beta_5) = (0.3191, -0.4758, -0.5478, -0.3885, 2.0488)$$

and $\beta_0 = 0.0518$.

Therefore, the semi-parametric prediction model is

$$\tilde{Y}_t = 0.0518 + 0.3191GDP_t - 0.4758TEIV_t - 0.5478IFA_t - 0.3885IAV_t + 2.0488DI_t + \varepsilon \quad (11)$$

Here

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

Thus, we can use system (11) to make 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 forecast values as well as the relative errors (RE) for the three methods.

Table 4: Observed and forecast electricity consumption[a] in China, 1980-2010, for three different approaches

Year	Observed value y_i	Observed value $\tilde{y}_i$	GM(1,1)		ANN		SPRM[b]	
			FV	RE(%)	FV	RE(%)	FV	RE(%)
Model building Stage: 1980-2005								
1980	-0.9557	-1.0938	-0.9557	0.00	-1.0549	-15.61	-1.1222	-2.59
1981	-0.9467	-1.0867	-1.0184	-7.57	-1.0333	-21.06	-1.1073	-1.89
1982	-0.9283	-1.1006	-0.9968	-7.38	-1.0059	-24.81	-1.0939	0.61
1983	-0.9043	-1.1194	-0.9731	-7.61	-0.9811	-25.24	-1.0780	3.69
1984	-0.8784	-1.1185	-0.9468	-7.79	-0.9427	-29.47	-1.0557	5.61
1985	-0.8443	-1.1072	-0.9178	-8.71	-0.9063	-29.92	-1.0163	8.21
1986	-0.8131	-1.1056	-0.8858	-8.94	-0.8729	-30.31	-1.0221	7.55
1987	-0.7656	-1.0685	-0.8505	-11.08	-0.8435	-25.28	-1.0099	5.48
1988	-0.7200	-1.0067	-0.8115	-12.71	-0.7972	-25.51	-0.9563	5.00
1989	-0.6795	-0.9350	-0.7684	-13.08	-0.7365	-30.02	-0.8680	7.16
1990	-0.6431	-0.9076	-0.7208	-12.08	-0.6767	-34.61	-0.8421	7.21
1991	-0.5858	-0.8081	-0.6684	-14.12	-0.6091	-35.88	-0.7506	7.11
1992	-0.5098	-0.6525	-0.6104	-19.73	-0.5423	-33.35	-0.6640	-1.77
1993	-0.4351	-0.5117	-0.5464	-25.58	-0.4880	-29.58	-0.5412	-5.76
1994	-0.3504	-0.6219	-0.4760	-35.84	-0.4502	-28.48	-0.5526	11.14
1995	-0.2662	-0.5533	-0.3981	-49.54	-0.4253	-59.76	-0.5058	8.591
1996	-0.1976	-0.4606	-0.3121	-57.94	-0.3666	-85.52	-0.4351	5.54
1997	-0.1507	-0.4404	-0.2171	-44.06	-0.2587	-71.66	-0.4612	-4.73
1998	-0.1198	-0.4021	-0.1122	6.34	-0.1633	-36.31	-0.3638	9.52
1999	-0.0451	-0.3192	0.0036	107.98	-0.0300	33.48	-0.2871	10.05
2000	0.0926	-0.1699	0.1315	-42.01	0.1672	-80.56	-0.1621	4.59
2001	0.2145	-0.0429	0.2726	-27.08	0.3932	-83.31	-0.0428	0.27
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	0.2995	11.44
2004	0.9239	0.7331	0.7907	14.41	1.2448	-34.73	0.7631	-4.09
2005	1.2173	0.9229	1.0005	17.80	1.3831	-13.62	1.0229	-10.84
Testing Stage: 2006-2010								
2006	1.6862	1.4963	1.2323	26.92	1.3314	24.54	1.5525	-3.75
2007	1.9959	1.7342	1.4879	25.45	1.7004	29.81	1.6409	5.37
2008	2.1773	1.9529	1.7701	18.70	1.8656	-6.07	1.8434	5.61
2009	2.4105	2.1126	2.0817	13.64	2.1001	-8.42	2.2896	-8.37
2010	2.5456	2.2477	2.4257	4.71	2.5276	4.97	2.1771	3.14

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

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. Figure 6 shows the model percentage error distributions for the SPRM prediction approach. In this figure, calibrations 1 to 26 correspond to the model building stage, and calibrations 27 to 31 correspond to the testing stage.

Table 5: Comparative analysis of forecasting errors

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	-3.52	0.0421	0.0025
Testing Stage: 2006-2010			
GM(1,1)	17.88	0.3636	0.1505
ANN	12.75	0.2581	0.0814
SPRM	5.25	0.1013	0.0120

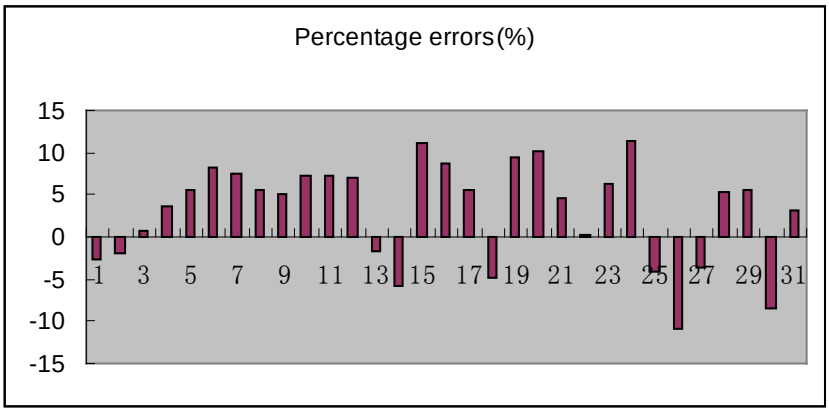

Figure 6: Percentage errors for the SPRM approach

5. CONCLUSION

The major contribution of this paper is to propose 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 use it as the non-parametric part of this semi-parametric regression model, thereby eliminating the local disturbance of the forecast process and effectively reducing the prediction error or other system errors. The forecast results demonstrate that the model performs remarkably well, and also demonstrate the effectiveness and reliability of the approach.

ACKNOWLEDGEMENTS

The authors 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.

REFERENCES

[1] **The National Bureau of Statistics.** 2009. *National Bureau of Statistics of China. 60 Years of New China.* China Statistics Press.

[2] **Harris, J.L. & Lon-Mu, L.** 1993. Dynamic structural analysis and forecasting of residential electricity consumption. *International Journal of Forecasting,* 9, pp. 437-455.

[3] **Jannuzzi, G. & Schipper, L.** 1991. The structure of electricity demand in the Brazilian household sector. *Energy Policy,* 19(9), pp. 879-891.

[4] **Ranjan, M. & Jain, V.K.** 1999. Modeling of electricity energy consumption in Delhi. *Energy,* 24(4), pp. 351-361.

[5] **Wang, X.J., Yang, S.L. et al.** 2010. Dynamic GM (1,1) model based on cubic spline for electricity consumption prediction in smart grid. *China Communications,* 7(4), pp. 83-88.

[6] **Wang, X.J. & Yang, S.L.** 2010. Electricity demand forecasting based on threepoint Gaussian quadrature and its application in smart grid. *The 6th International Conference on Wireless Communications Networking and Mobile Computing.*

[7] **Wang, X.J., Shen, J.X. & Yang, S.L.** 2010. Application research on Gaussian orthogonal interpolation method for electricity consumption forecasting of smart grid. *Power System Protection and Control,* 38(21), pp. 141-145,151.

[8] **Wang, X.J., Yang, S.L. & Ding, J.** 2010. Simulation of orthogonalization prediction based on grey Markov chain for electricity consumption. *Journal of System Simulation,* 22(10), pp. 2253-2256.

[9] **Wang, X.J., Yang, S.L. & Wang, H.J.** 2010. Application research on grey orthogonal prediction model based on Markov chain for electricity consumption forecasting of smart grid. *The 12th Chinese Annual Conference on Management Science.*

[10] **Ching-Lai, H., Marnont, A., Simon, W. & Shanti, M.** 2005. Analyzing the impact of weather variables on monthly electricity demand. *IEEE Trans,* 20, pp. 2078-2085.

[11] **Egelioglu, F., Mohamad, A.A. & Guven, H.** 2001. Economic variables and electricity consumption in Northern Cyprus. *Energy,* 26, pp. 355-362.

[12] **Wei, L., Wu, J. & Liu, Y.** 2000. Long-term electricity load forecasting based on system dynamics. *Automation of Electric Power Systems,* 20, pp. 24:47.

[13] **Narayan, P. K. & Prasad, A.** 2008. Electricity consumption - real GDP causality nexus: Evidence from a bootstrapped causality test for 30 OECD countries. *Energy Policy,* 36, pp. 910-918.

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

[15] **Abdel-Aal, R.E., Al-Garni, A.Z. & Al-Nassar, Y.N.** 1997. Modeling and forecasting monthly electric energy consumption in eastern Saudi Arabia using abductive networks. *Energy,* 22(9), pp. 911-921.

[16] **Yan, Y.Y.** 1998. Climate and residential electricity consumption in Hong Kong. *Energy,* 23(1), pp. 17-20.

[17] **Nasr, G.E., Badr, E.A. & Younes, M.R.** 2002. Neural networks in forecasting electricity energy consumption: Univariate and multivariate approaches. *International Journal of Energy Research,* 26, pp. 67-78.

[18] **Niu, D.X., Zhao, L., Zhang, B. & Wang, H.F.** 2007. The application of particle swarm optimization based grey model to power load forecasting. *Chinese Journal of Management Science,*15(1), pp. 69-73.

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

[20] **Santos, P.J., Martins, A.G., Pires, A.J., Martins, J.F. & Mendes, R.V.** 2006. Short term load forecast using trend information and process reconstruction. *International Journal of Energy Research,* 30, pp. 811-822.

[21] **Engle, R.F., Granger, C.W.J., Rice, J. & Weiss, A.** 1986. Semi-parametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association,* 81, pp. 310-320.

[22] **Taylor, J.W.** 2006. Density forecasting for the efficient balancing of the generation and consumption of electricity. *International Journal of Forecasting,* 22, pp. 707-724.

[23] **Taylor, J.W., de Menezes, L.M.M. & McSharry, P.E.** 2006. A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting,* 22, pp. 1-16.

[24] **Akay, D. & Atak, M.** 2007. Grey prediction with rolling mechanism for electricity demand forecasting of Turkey. *Energy,* 32, pp. 1670-1675.

[25] **Zhou, P., Ang, B.W. & Poh, K.L.** 2006. A trigonometric grey prediction approach to forecasting electricity demand. *Energy,* 31, pp. 2839-2847.