Electrical Energy Consumption Forecasting Based on Cointegration and a Support Vector Machine in China

ZHANG XING-PING, GU RUI
School of Business Administration,
North China Electric Power University
No.2 Beinong Road, Zhuxinzhuang, Dewai, Beijing 102206
CHINA
Zhangxingping302@163.com

Abstract: - By undertaking a cointegration analysis with annual data over the period 1985~2005 in China, the estimation results show that there is cointegration relationship between electrical energy consumption and economic growth taking into account industry structure changes and technical efficiency. The model shows that three explanatory variables, the GDP per capita, heavy industry share and efficiency improvement are the crucial factors which influence the electric energy consumption. The three explanatory variables and the actual electrical energy consumption are input into a support vector machine(SVM), a Gaussian radial basis function is taken as the kernel function and electrical energy consumptions from 1994~2006 are forecasted. The forecast results prove that the multivariable SVM is valid in forecasting electrical energy consumption in China.

Key-Words: Cointegration analysis; Electrical energy consumption; Johansen cointegration test; Multivariate time series; Support vector machine ; Unit root test

1 Introduction

Electrical consumption forecasting is the basis for electric energy planning. Many scholars [1~5] have applied econometrics to study electricity demand and its main determining factors is usually analyzed correctly in theory, but it is greatly affected by fluctuations in the sample data. A lot of non-linear programming and combinational forecasting methods such as fuzzy logic methods are applied widely in electric load forecasting. But results produced by fuzzy logic methods are quite difficult to express and set up, and the parameters are not easy to modulate [6, 7]. A new machine learning technique called support vector machines (SVM) is not only helpful for solving problems involving small sample, devilish learning, high dimension and local minima, but also strong generalizability. So SVM was widely applied in electric load forecasting, and some research results [8~12] indicate that SVM has distinct advantages in electric load forecasting. SVM is seldom used in forecasting the electrical energy consumption, and when it is, actual electrical energy consumption is taken as the only input variable of the SVM, while the major factors which impact electrical consumption are not considered [13].

In this paper, the three variables, GDP per capita, heavy industry share and efficiency improvement, are taken as the explanatory variables, and the electrical energy consumption is taken as the explained variable. An equilibrium relationship between the explanatory and the explained variables is analyzed by the cointegration analysis. Taking these crucial factors and actual electricity consumption as the input variables of a SVM, and selecting the rational kernel function of the SVM, the electrical energy consumption is forecasted.

2 Multivariate Cointegration Analysis of Electrical energy Consumption

2.1 Cointegration Theory

Cointegration theory seeks to determine whether there is a stationary relationship among nonstationary economic variables, and whether there is a long-term equilibrium relationship among them. It avoids the disadvantages of unreliable regression results generated by spurious regression, and it can differentiate long-term stationary relationships from short-term dynamic relationships among variables. Before cointegration analysis came along, the combination of variables had to be stationary. The variable autoregression model, which includes *g* variables and *k* lags, is expressed as:

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} + \mu_t \quad (1)$$

Project supported by Education Ministry of China (06JC790014)
The authors are grateful to Robert Blohm rb112@columbia.edu for revising the English and making a few mathematical corrections.

Supposed all y_t are I(1); then a suitable transformation of equation (1) is made, and the error correction model is obtained as:

$$\Delta y_t = \mathbf{\Pi} y_{t-k} + \sum_{i=1}^{k-1} \mathbf{\Gamma}_i \Delta y_{t-i} + \mu_t \qquad (2)$$

where $\mathbf{\Pi} = \sum_{j=1}^{k} \beta_j - \mathbf{I}_g$, $\mathbf{I}_g$ is the *g*-step unit matrix, and $\mathbf{\Gamma}_i = \sum_{j=1}^{i} \beta_j - \mathbf{I}_g$.

Matrix $\mathbf{\Pi}$ is the coefficient matrix which reflects the long-term relationships of the variables. When the variables are in a long-term equilibrium state, the difference in the first variables of equation (2) is the zero vector, and E(μ_t)=0; so $\mathbf{\Pi} y_{t-k} = 0$ when the variables are in a long-term equilibrium state, and this can be judged by calculating the rank and the eigenvalues of matrix $\mathbf{\Pi}$.

When all the endogenous variables are I(1), and when all the variables of $\mathbf{\Pi} y_{t-k}$ are I(0), the stochastic error term is a stationary process. If $0 < Rank(\mathbf{\Pi}) = m < g$, there are matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and $\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^{\mathrm{T}}$, So equation (2) is transformed into equation (3).

$$\Delta y_t = \boldsymbol{\alpha} \boldsymbol{\beta}^{\mathrm{T}} y_{t-k} + \sum_{i=1}^{k-1} \mathbf{\Gamma}_i \Delta y_{t-i} + \mu_t \qquad (3)$$

Each row of the matrix $\boldsymbol{\beta}^{\mathrm{T}} y_{t-k}$ is a stationary combined variable, that is, each row is a linear combined form which enables the variables $y_{1,t-1}, y_{2,t-1}, \cdots, y_{g,t-1}$ to be cointegrated.

2.2 Explained and Explanatory Variables

Lots of documents show that GDP plays the most important role in determining electricity consumption in China. Thus there is a positive correlation between electrical consumption and GDP. Taken into account the population, the GDP per capita is taken as an explanatory variable.

In China, the share of industrial electricity consumption is rising from 71.75% in 2000 to 74.89% in 2006. Most of the electrical energy is consumed by the heavy industry, in 2006 for example, electrical energy consumed by heavy industrial took up 60.26% of all electrical energy consumption, and 79.71% of all industrial electrical energy consumption. The breakdown of electrical energy consumption has been changing in China; electrical energy consumption by the light industry increased 1.87% and by heavy industry decreased 0.14% in 2006. So the heavy industry share or the ratio of heavy industry production value to gross industry production reflects changing industrial structure. So the heavy industry share is a key factor which influences the electrical energy consumption, and is taken as an explanatory variable.

As the science and technology level has steadily increased since 1997, the comprehensive social and technology level index increased by 1.5% in 2006 to 47.11%. Consequently, efficiency improvement plays an important role in electrical energy consumption; so the ratio of increase in industrial value to industrial electricity consumption is used to reflect efficiency improvement.

So electricity consumption (*Q*) is chosen as the explained variable, and GDP per capita (*PCGDP*), heavy industry share (*HIS*), and efficiency improvement (*EI*) are chosen as the explanatory variables. The sample space is from 1985 to 2005. The impact of inflation is removed, and the samples are shown in table 1.

Table 1 Sample Data from 1985 to 2005

Year	Q (*10 million KW.h*)	PCGDP (*yuan*)	HIS	EI
1985	4705.9	858	52.6	1.178
1986	5096.4	963	52.4	1.231
1987	5514.3	1112	51.8	1.311
1988	5956.0	1366	50.7	1.509
1989	6390.8	1519	51.1	1.566
1990	6895.7	1644	50.6	1.584
1991	7399.1	1893	51.6	1.750
1992	7991.0	2311	53.4	2.017
1993	8590.4	2998	53.5	2.547
1994	9260.4	4044	53.7	3.214
1995	10023.4	5046	52.7	3.744
1996	10764.3	5846	51.9	4.206
1997	11284.5	6420	51.0	4.472
1998	11598.5	6796	50.7	4.640
1999	12305.2	7159	50.8	4.646
2000	13471.4	7858	50.2	4.719
2001	14633.5	8622	50.6	4.740
2002	16311.5	9398	50.9	4.570
2003	19031.6	10542	64.5	4.492
2004	21971.4	12336	66.5	4.547
2005	24940.4	14040	69.0	4.682

2.3 Cointegration Analysis

Because the economic variables in a time series are usually nonstationary, and there is neither randomness nor a definite tendency, the sample data should be transformed by taking the natural log so as to reduce vibration, and by taking the difference so as to eliminate instability and heteroscedasticity. Before cointegration analysis, the Augment Dickey-Fuller (ADF) test was applied to test whether a data series is stationary. The null hypothesis is that the data series is nonstationary. The test results are shown in Table2. (△ expresses the first order difference).

Table 2 ADF unit root test results on variables

Variables	ADF Test Statistic	5% Critical Value	Conclusion
LNQ	-3.423	-3.710	non-stationary
ΔLNQ	-3.168*	-3.066	stationary
LNPCGDP	-1.217	-3.691	non-stationary
ΔLNPCGDP	-4.107*	-3.733	stationary
LNHIS	-1.093	-3.658	non-stationary
ΔLNHIS	-4.413*	-3.674	stationary
LNEI	-1.316	-3.691	non-stationary
ΔLNEI	-3.802*	-3.733	stationary

Note:"*" expresses MacKinnon critical values for rejection of hypothesis of a unit root under the 5% significance level

In table 2 all the original values of the variables are less in absolute value than the ADF test statistic's critical value at the 5% significance level; so we fail to reject the null hypothesis at the 5% significance level. But all the computed ADF test statistic values of the first difference of the variables are greater in absolute value than the ADF test statistic's critical value at the 5% significance level, and so the null hypothesis is rejected at the 5% significance level, and so all the variables are I(1), and this meets the conditions for cointegration analysis. In other words, from 1985 to 2005, there may be a cointegration relationship between electricity consumption and the explanatory variables.

The cointegration test needs to be run to find whether there is a cointegration relationship. The null hypothesis is that there is no cointegration relationship between electrical energy consumption and the explanatory variables. All the observed series contain a time trend; so the cointegration test model contains the intercept and time trend. The results of the Johansen cointegration test are shown in table 3.

Table 3 Results of Johansen Cointegration test

Eigenvalue	Likelihood	5 Percent Critical Value	Hypothesized No. of CE(s)
0.7561	60.950	47.856	None*
0.7184	34.141	29.797	At most 1*
0.3170	10.645	15.495	At most 2
0.1380	2.821	3.841	At most 3

Note: "*" expresses it is significant under 5% confidence level

The results in table 3 show that the Likelihood ratio of the first two eigenvalues is greater than the critical value at the 5% significance level; therefore there is a long-term equilibrium relationship between electricity consumption and the three explanatory variables. The normalized cointegration coefficients are shown in table 4.

Table 4 Normalized Cointegration Coefficients

LNQ	LNPCGDP	LNHIS	LNEI	C
1.0000	1.01 (0.028)	0.13 (0.065)	-0.86 (0.040)	1.27

Note: the number in parenthesis in the table is the asymptotic standard error.

So the cointegration function is stated as:

$$\hat{LNQ} = 1.27 + 1.01 LNPCGDP + 0.13 LNHIS - 0.86 LNEI \tag{4}$$

If the residual series of equation (4) is stationary, there is a cointegration relationship between electrical energy consumption and the three explanatory variables; otherwise, there is no cointegration relationship. So the Johansen cointegration test is run to test whether the residual series is stationary, and the test results are shown in table 5.

Table 5 ADF Unit Toot Test Results on Residual Series

ADF Test Statistic	1 Percent Critical Value	5 Percent Critical Value	10 Percent Critical Value
-4.01	-4.62	-3.71	-3.30

The 5% critical value of the ADF test statistic is -3.71; the computed ADF test statistic value of -4.01 indicates that there are no unit roots in the residual series; that is, the residual series is stationary. So there is a cointegratoin relationship between electricity consumption and the explanatory variables.

In equation (4) the coefficients of the explanatory variables are the elasticity of *Q* with respect to the three explanatory variables. That is, a 1% increase in *PCGDP* leads to, on average, a 1.01% increase in *Q*, a 1% increase in *HIS* increases *Q* by 0.13%, and a 1% improvement in *EI* decreases *Q* by 0.86% on average.

3 Multivariate SVM Model

3.1 Regression Arithmetic of SVM

Suppose $T = \{(x_1, y_1), \cdots, (x_i, y_i), \cdots, (x_p, y_p)\}$, where $x_i \in R^m$ is the input variable, $y_i \in R$ is the corresponding output value and p is the total number of the data points. Then the SVM regression function is:

$$f(x) = (\boldsymbol{\omega} \cdot \Phi(x)) + b \qquad (5)$$

where $\Phi(\cdot)$ is a non-linear mapping function, $\boldsymbol{\omega}$ is a weight vector, and b is the error term. $\boldsymbol{\omega}$ and b are estimated by:

$$\min R(\boldsymbol{\omega}) = \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{p} L_\varepsilon(y_i, f(x_i)) \qquad (6)$$

where C is the punishment parameter, which is considered to specifies the trade-off between empirical risk and the model's flatness. $\frac{1}{2}\|\boldsymbol{\omega}\|^2$ is the normalization term. $L_\varepsilon(y_i, f(x_i))$ is called the ε-insensitive loss function, which is defined as:

$$L_\varepsilon(y_i, f(x_i)) = \max(|y_i - f(x_i)| - \varepsilon, 0) \qquad (7)$$

In equation (7) the loss equals zero if the forecasting error is less than ε; otherwise the loss not less than ε. In order to represent the distance from actual values to the corresponding boundary values of the ε-band, two positive slack variables ξ and ξ^* are introduced. Then, equation (6) is transformed into the following form:

$$J = \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{p}(\xi_i^* + \xi_i) \qquad (8)$$

$$s.t. \begin{cases} y_i - [\boldsymbol{\omega}, \Phi(x)] - b \le \varepsilon + \xi_i^* & \xi_i^* \ge 0 \\ [\boldsymbol{\omega}, \Phi(x)] + b - y_i \le \varepsilon + \xi_i & \xi_i \ge 0 \end{cases}$$

This constrained optimization problem is solved by using the following Lagrangian form:

$$\max H(\partial, \partial^*) = -\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(\partial_i, \partial_i^*)(\partial_i, \partial_i^*)K(x_i, x_j) + \sum_{i=1}^{p}\partial_i^*(y_i - \varepsilon) - \varepsilon\sum_{i=1}^{p}(y_i + \varepsilon)$$

$$s.t. \begin{cases} \sum_{i=1}^{p}(\partial_i - \partial_i^*) = 0 \\ \partial_i, \partial_i^* \in [0, C] \end{cases} \qquad (9)$$

where ∂_i, ∂_i^* are Lagrangian multipliers, and $\partial_i - \partial_i^* \ne 0$ i.e. corresponding data points are a support vector. ∂_i and ∂_i^* calculated by the Lagrange multipliers, an optimal desired weight vector of the regression hyperplane is obtained:

$$\omega^* = \sum_{i=1}^{p}(\partial_i - \partial_i^*)K(x_i, x) \qquad (10)$$

Hence, the regression function is:

$$f(x) = \sum_{i=1}^{p}(\partial_i - \partial_i^*)K(x_i, x) + b \qquad (11)$$

where $K(x_i, x)$ is called the kernel function. The value of the kernel function equals the inner product of $\Phi(x_i)$ and $\Phi(x)$, which are produced by mapping x_i and x into a higher dimensional feature space; that is:

$$K(x_i, x) = \Phi(x_i, x) \qquad (12)$$

3.2 Multivariate SVM Model

For a univariate time series $\{x_1, x_2, \cdots, x_n\}$, training sample sets, $\{x_1, x_2, \cdots, x_m\} \to \{x_{m+1}\}$, $\{x_2, x_3, \cdots, x_{m+1}\} \to \{x_{m+2}\}$, $\cdots$, are established. Suppose

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \\ x_2 & x_3 & \cdots & x_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-m} & x_{n-m+1} & \cdots & x_{n-1} \end{bmatrix}$$

and $\mathbf{Y} = [x_{m+1} \quad x_{m+2} \quad \cdots \quad x_n]^{\mathrm{T}}$

where $\{x_i, x_{i+1}, \cdots, x_{i+m-1}\}$ is the input vector, $\{x_{i+m}\}$ is the output value and m is the embedded dimension.

Supposed that we have observed an l-dimensional multivariate time series:

$$\{X_n\}_{n=1}^{N} = \{(x_{1,n}, x_{2,n}, \cdots, x_{l,n})\}$$

As in the case of a univariate time series, we make a state space reconstruction:

$$\begin{aligned} \mathbf{V}_n = [&x_{1,n}, x_{1,n-1}, \cdots, x_{1,n-m_1+1}; \\ &x_{2,n}, x_{2,n-1}, \cdots, x_{2,n-m_2+1}; \cdots; \\ &x_{l,n}, x_{l,n-1}, \cdots, x_{l,n-m_l+1}]^{\mathrm{T}} \end{aligned}$$

m_i is the embedded dimension of ith variable, $i = 1, 2, \cdots, l$. The node quantity is the sum of the embedded dimensions in the multivariate time series:

$$m = m_1 + m_2 + \cdots + m_l$$

The structure of the multivariate SVM is shown in figure 1.

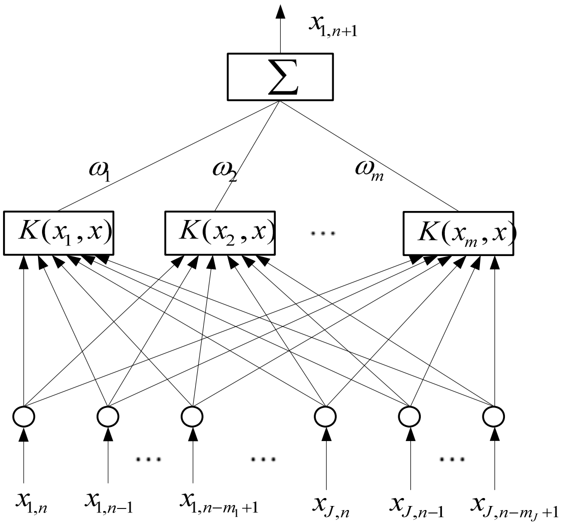

Fig.1 Multivariate Time Series SVM

4 Case Study

4.1 Kernel Function and Parameters

Comparing the results calculated by 4 kinds of kernel function, the Gaussian radial basis function, $K(x_i, x) = \exp(-|x_i - x|/2\sigma^2)$, was applied in the SVM.

In the model of SVM, the forecast results are sensitive to the parameter of σ, and insensitive to the other parameters. According to the simulation results and the principle of minimum error [14], let ε =0.0008, and C=10000. When ε and C are selected, the simulation results between σ and the mean of absolute proportional error (MAPE) of the forecast results is simulated and shown in figure 2. According to the figure 2, when the σ=3.5, the MAPE is minimal, so σ is set as 3.5.

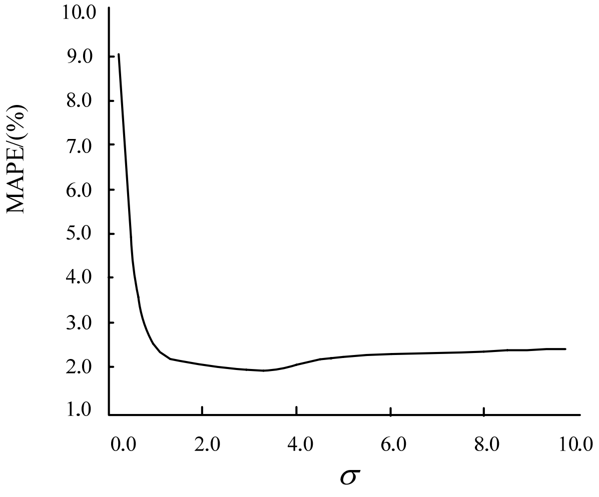

Fig.2 Simulated Relationship Between σ and MAPE

4.1 Forecast Results

The electrical consumption is the output variable. The model (4) indicates that the *PCGDP,HIS* and *EI* are the determining factors of *Q*, and there is a long term equilibrium relationship among them. So *PCGDP*, *HIS*, *EI* and actual *Q* are the input variables of the SVM, which is called multivariable SVM. In order to eliminate dimensional diversity in the variations in each time series, data is normalized into the interval [0, 1]. The forecasted results of electrical consumption by multivariable SVM are shown in table 6.

To compare the forecast results of the multivariable SVM with that of univariable SVM, the forecast results by univariable SVM are shown in table 6. There is only one input variable, the actual electrical consumption, in the model of univariable SVM. That is, the future electrical consumptions are obtained by inputting the previous electrical consumptions into the model of SVM.

Table 6 The Forecast Results of Electrical Consumption

Year	Multivariable SVM		Univariable SVM	
	Forecast value	APE	Forecast value	APE
1994	9141.1	1.2883	9221.3	0.4222
1995	9833.3	1.8966	9946.6	0.7662
1996	10529.2	2.1814	10867.9	0.9624
1997	11266.3	0.1613	11605.2	2.8420
1998	11652.4	0.4647	11519.1	0.6846
1999	12029.5	2.2405	11789.4	4.1917
2000	13059.2	3.0598	13817.2	2.5669
2001	14404.7	1.5635	14169.4	3.1715
2002	15890.8	2.6985	15902.1	2.6293
2003	18374.2	3.4543	18510.3	2.7391
2004	22794.1	3.7444	22738.3	3.4904
2005	24762.4	0.7137	25604.5	2.6627
2006	28727.5	1.6964	28759.6	1.8100

The absolute proportional error (APE) between the actual value and the forecasted value is calculated by: $\text{APE} = \left|(x_i - x_i^*)/x_i\right| \times 100\%$.

where x_i^* is the forecast value, and x_i is the actual value.

According to the table 6, the average values and the variances of APE forecasted by multivariable SVM and univariable SVM are shown in table 7.

Table 7 Average values and variances of APE

Multivariable SVM		Univariable SVM	
Average	Variance	Average	Variance
1.9359	1.2499	2.2261	1.4166

From table 6, the APE forecasted by multivariable SVM is in the open interval of [0.1613, 3.7444], and the APE forecasted by univariable SVM is in the open interval of [0.4222, 4.1917]. From table 7, the average value and variance of APE forecasted by multivariable are less than that forecasted by univariable SVM. It proves that the forecast effect by multivariate SVM is better than that by univariable SVM in forecasting electrical consumption in China.

5 Conclusion

Two conclusions are obtained:

(1) By undertaking a cointegration analysis with annual data over the period 1985~2005 in China, the estimation results show that there is cointegration relationship between electrical energy consumption and the three explanatory variables; the cointegration model (4) explains how the three explanatory variables influence the electrical energy consumption.

(2) Input *PCGDP*, *HIS*, *EI* and actual electrical energy consumption into the model of SVM, the electrical energy consumption are forecasted. Comparing the forecast results by multivariable SVM with that by univariable, we conclude that the forecast effect by multivariate SVM is better than that by univariable, because the crucial factors which influence the electrical energy consumption are considered in the multivariable SVM.

References:

[1] Yuan Jiahai, Ding Wei, Hu Zhaoguang, Cointegration and Co-feature Analysis of Electricity Consumption and Economic Growth in China, *Power System Technology*, Vol. 30, No.9, 2006, pp. 10-15.

[2] Chien-Chiang Lee, Chun-Ping Chang, Structural Breaks, Energy Consumption, and Economic Growth Revisited: Evidence from Taiwan, *Energy Economics*, Vol. 27, No. 6, 2005 , pp. 857-872.

[3] Erkan Erdogdu, Electricity Demand Analysis Using Cointegration and ARIMA Modeling: A Case Study of Turkey, *Energy Policy*, In Press, Corrected Proof, available on online at 17 April 2006.

[4] Ajith Abraham, Baikunth Nath, A Neuro-fuzzy Approach for Modeling Electricity Demand in Victoria, *Applied Soft Computing*, Vol.1, No. 2, 2001, pp.127-138.

[5] Yemane Wolde-Rufael, Electricity Consumption and Economic Growth: A Time Series Experience for 17 African Countries, *Energy Policy*, Vol.34, No. 10, 2006, pp.1106-1114.

[6] Taylor J W, Buizza R., Neural Network Load Forecasting with Weather Ensemble Predictions, *IEEE Trans.Power Syst*., Vol.17, No.3, 2002, pp.626-632.

[7] Liu Mengliang, Liu Xiaohua, Gao Rong, Short Term Load Forecasting Using Wavelet Transform and SVM Based on Similar-days, *Transactions of China Electrotechnical Society*, Vol.21, No.11, 2006, pp. 59-63.

[8] Xie Hong, Wei Jiangping, Liu Heli, Parameter Selection and Optimization Method of SVM Model for Short-term Load Forecasting, *Proceedings of the CSEE*, Vol.26,No.22, 2006, pp. 17-22.

[9] Zhang Qian-jin, Research on Electric-Power Load Forecasting Based on Support Vector Machine Regression Technique, *Aeronautical Computing Technique*, Vol.36, No.4, 2006, pp.105-110.

[10] Xie Hong，Chen Zhiye，Niu Dongxiao, et al, Research on a Daily Load Forecasting Model Based on Wavelet Decomposition and Climatic Influence, *Proceedings of the CSEE*，Vol.21, No.5, 2001, pp. 5-10.

[11] Zhao Dengfu，Pang Wenchen，Zhang Jiangshe，et al, SVM for Short Term Load Forecasting Based on Bayesian Theory and Online Learning, *Proceedings of the CSEE*，Vol.25, No. 13, 2005, pp. 8-13.

[12] Huang Yuansheng, Zheng Yan, Qi Jianxun, The Application of Electric Power Demand Forecasting Based on LS-SVM, *Chinese Journal of Management Science*, Vol.13, Special Issue, 2005, pp. 32-37.

[13] Wang Xiaohong, Wu Dehui, An Annual Electric Consumption Forecasting Model Based on Least-square Support Vector Machines, *Relay*, Vol.34, No.16, 2006, pp. 15-21.

[14] V Cherkassky，Y Ma, Practical Selection of SVM Parameters and Noise Estimation for SVM Regressions, *Neural Networks*，Vol.17, No.1, 2004, pp. 113-126.