Applying LS-SVM to Predict Primary Energy Consumption

Wang Yi
Economy and Management School
North China Electric Power University
Beijing, the People's Republic of China
wy6060@126.com

Li Ying
Economy and Management School
North China Electric Power University
Beijing, the People's Republic of China
Liying2005@126.com

***Abstract*—Since reform and opening up to the outside world, rapid industrialization and urbanization in China have played an important role in the increase of energy consumption. China became the second energy consumption country all over the world in 2008. As a big country in energy consumption, forecasting energy consumption is one of the most important tools for energy policy setting. Although there are many researchers has devoted into the relationship between energy and economy, the forecasting of energy consumption is still in its infancy. In this paper, economic indictors, such as real GDP, population, industrial structure, import and export, and government expenditure, are selected as input influencing factors on energy consumption, and then it is performed to find proper features from the data in terms of their statistical information. And then a novel forecasting model of energy consumption, LS-SVM regression prediction model, is presented. In the end, the case study is carried out to test the proposed model, which shows that it has many promising features that make it become a more reliable yet functional prediction tool for forecasting energy consumption.**

Keywords-Energy consumption prediction; influencing factors analysis; Least squares support vector machines

I. INTRODUCTION

Energy is one of important resources to human survival and growth. The growth of economics has been closely linked with the availability, extraction, distribution and use of energy [1]. Throughout the history of the development of human society, the energy problem has been always the vital problem for governments at all levels to formulate economic development strategies and policies. Since the reform and the opening to the outside world, rapid industrialization and urbanization in China have played an important role for energy consumption. The Chinese economy has achieved an average growth rate of 12% from 1979 to 2007, which makes China one of countries with fastest growth rate. Meanwhile, aggregated primary energy consumption has increased largely. According to the data from National Energy Administration, China became the second energy consumption country all over the world in 2008 with 70.2% coal, 18.4% petroleum, 3% natural gas and 8.4% primary electricity. With the rapid economic development, the energy demand has zoomed in recent years, which makes the gap between energy supply and demand become more and more serious. As a net energy importing country since 2006, our energy importing reliance increases gradually, so the accurate estimating of its energy demand is critical in energy policy-making. Although many researches are interested in studying energy consumption forecasting in other countries and regions, there are only a few literatures on energy consumption prediction in china, which makes research space in this field. Then the relative studies on energy consumption have strong realistic and theoretical significance.

So far, some soft computing techniques, such as artificial neural network (ANN), genetic algorithm and art colony optimization, autoregressive integrated moving average model (AIMAM), generalized autoregressive conditional heteroskedasticity (GARCH) and its extended models, etc, have been estimated national energy demand or sectional energy demand. Reference [2] established a forecasting model based on GA to estimate Turkey's transport energy demand. Reference [3] used ANN method to forecast South Korea's national energy demand. Reference [4] used GA method to estimate Turkey's energy demand based on economic indicator, such as GDP, population and import and export figures of Turkey. Reference [5] adopted ant colony optimization approach to estimate energy demand in Turkey based on economic indicators, such as GDP, population and import and export. Reference [6] used ARIMA method to forecast primary energy demand of Turkey. Reference [7] used the exponential form of the GARCH to predict energy demand of Taiwan. Every forecasting method has its advantages and at same time, has some weakness. For example, due to its self-studying and self-adapting ability, ANN has achieved comprehensive application in foresting, but the calculation speed of the neural network method is slow.

In the last decade, SVM, developed by Vapnik and his colleagues in 1995, has been introduced for pattern recognition and regression problems. These SVM algorithms have solid theoretical foundation rooted in statistical learning theory and been applied into many fields, including in energy demand forecasting [8]. As the algorithms have been applied widely, some problems of standard support vector machine appear. In detail, its modeling process involves a quadratic programming problem, but such QP problem is time consuming. To overcome this shortcoming, Least Squares Support Vector Machines (LS-SVM) have been proposed, which converts the inequality constraints of original SVM into equality ones. And this leads to solving a linear system instead of a QP problem, whose convergence performance is enhanced obviously.

There are many influencing factors on energy consumption in one country. On the whole, the energy demand relates to

many factors, among which are mainly gross domestic product (GDP), population, energy structure, industrial structure, government expenditure, etc. The goal of study is to provide a more reliable forecasting model of primary energy consumption using a novel machine-learning algorithm based on least squares support vector machines (LS-SVM), which is a relatively new and powerful machine-learning algorithm widely applied in regression, pattern classification, and other relative fields [9].

The structure of the paper is the following: Section II provides an introduction to LS-SVM regression forecasting model; Section Ⅲ offers the case study using the real-world data; Finally, Section IV presents the main conclusion and suggestions for future research.

II. LS-SVM FORCSTING MODEL

A. Data description

Similarly as previous achievements [3-7], real GDP (10^2 million yuan), population (10^4) are regarded as influencing factors in this paper.

As to industrial structure, we select the ratio of the tertiary sector to GDP (in percentage terms), which can affect primary energy consumption by the change of the tertiary sector.

Reference [10] provided some evidence and found that an increase in government expenditure in China leads to an increase in energy intensity. Therefore, in this paper, government expenditure (10^2 million yuan), standing for the total government expenditures on final goods and services which includes investment expenditure by the government, etc, is also an influencing factor.

Imports (10^2 million yuan) and exports (10^2 million yuan) are also influencing factor, because we can know the industulization and economic situation in China from these data.

All the data from 1990-2008 can be collected from China Statistical Yearbook. Real GDP, government expenditure, imports and exports are all converted according to the constant price of 1978.

B. Facror analysis on influencing factors

In the fields of data processing, statistical characteristics have special effects to denote data. For example, mean value reflects the average level of samples. And standard variance is often applied to express the centrality and the deviation from the mean value of the samples. Such characteristics fit for denoting Gaussian variables. In other words, if some variable is not Gaussian, mean value and standard variance may only represent part, not the whole of data characteristics. Moreover, many real variables are not Gaussian, some are approximately Gaussian, and some are far from that. So in this paper, higher-ranked statistical characteristics are introduced to illustrate the data tendency latent in the original influence factors of energy consumption.

On the whole, higher-ranked information includes skewness, kurtosis, and so on. Supposing x is a continuous variable, its kurtosis can be defined as below.

$$k_{skewness} = \frac{E(x-\mu)^3}{\sigma^3} = \begin{cases} >0 & right-leaning \\ =0 & Gaussian \\ <0 & left-leaning \end{cases} \tag{1}$$

$$k_{kurtosis} = \frac{E(x-\mu)^4}{\sigma^4} - 3 = \begin{cases} >0 & Super-Gaussian \\ =0 & Gaussian \\ <0 & Sub-Gaussian \end{cases} \tag{2}$$

where μ is the mean value of x. σ is the standard deviation of x. Here to examine the properties of primary energy consumption, the higher-ranked statistical characteristics of the influencing factors are listed in TABLE I.

TABLE I
HIGHER-RANKED STATISTICAL CHARACTERISTICS OF THE INFLUENCE FACTORS

Original influencing factor	skewness	Kurtosis
GDP	-0.438	-0.220
Population	-0.372	-1.074
Import	0.577	-1.068
Export	0.779	-0.601
The ratio of the tertiary industry to GDP	-0.170	-1.767
Government expenditure	-0.828	-0.223

From TABLE I, we find that all of influencing factors of primary energy consumption are sub-Gaussian. Generally speaking, combining skewness with kurtosis, these indicators show non-Gaussian properties, so we don't need to eliminate Gaussian or Gaussian-like noises. From influencing factors and samples, the number of samples is small, compared with six influence factors. Therefore, it forms a typical Gaussian nonlinear regression under the condition of small samples. LS-SVM shows good performance in forecasting under the condition of small samples. If we use neural network method, we need 10 times samples.

C. Regression forecasting by LS-SVM

In this part, the regression analysis on the expected aggregated energy consumption and its influencing factors are performed by Least Squares Support Vector Machines to accomplish the forecasting modeling.

Supposing the extracted samples are

$$D = \{(x_1, y_2), (x_2, y_2), \cdots, (x_N, y_N)\}$$

where N is the sample number, and $\boldsymbol{x}_i \in R^n, y_i \in R$, $\boldsymbol{x}_i$ is the extracted factor vector. The aggregated energy consumption model $y = w^T \varphi(x) + b$ is inferred from the data $D = \{(x_i, y_i)\}_{i=1}^N$ by minimizing the cost function

$$\min_{w,b} J_1(w,b) = \mu E_W + \zeta E_D = \frac{\mu}{2} w^T w + \frac{\zeta}{2} \sum_{i=1}^{N} e_i^2 \quad (3)$$

subject to the constraints

$$e_i = y_i - (w^T \varphi(x_i) + b), \quad i = 1, \cdots, N \quad (4)$$

The regularization and error term are defined as $E_W = \frac{1}{2} w^T w$ and $E_D = \frac{1}{2} \sum_{i=1}^{N} e_i^2$ respectively. The trade-off between regularization and training error is determined by the ratio $\gamma = \zeta / \mu$.

Standard SVM for regression involves the use of the so-called ε -intensive loss function which leads to a convex Quadratic Programming problem in the dual space. However the proposed aggregated energy consumption by LS-SVM uses a least squares cost function that results into a linear Karush-Kuhu-Tucker system in the dual space. To find the solution of (3), the Lagrangian function is constructed as (5) by introducing the Lagrange multipliers for the equality constraints (3) and taking the conditions for optimality,

$$L_1(w,b,e,\alpha) = J_1(w,e) - \sum_{i=1}^{N} \alpha_i [w^T \varphi(x_i) + b + e_i - y_i] \quad (5)$$

Then a linear Karush-Kuhu-tucker system for finding annual energy consumption model is obtained in the dual space

$$\begin{bmatrix} 0 & 1_v^T \\ 1_v & \Omega + \gamma^{-1} I^N \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \quad (6)$$

with $Y = [y_1; \cdots; y_N]$, $1_v = [1; \cdots; 1]$, $e = [e_1; \cdots e_N]$, $\alpha = [\alpha_1; \cdots; \alpha_N]$,and where Mercer's condition is applied within the Ω matrix $\Omega_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$. Possible kernel functions are, e.g.:

- Linear kernel $K(x_1, x_2) = x_1^T x_2$
- Radial Basis Function (RBF)-kernel

$K(x_1, x_2) = \exp(-\|x_1 - x_2\|_2^2 / \sigma^2)$,

where Mercer's condition holds for all possible choices of the kernel parameter. The aggregated energy consumption regressor is then constructed as follows:

$$y(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i) + b \quad (7)$$

In this way, the mapping relationship of energy consumption and its influencing factors are obtained.

III. EMPIRICAL WORK

In this section, we present experimental results on the proposed model based on LS-SVM for robust forecasting of aggregated primary energy consumption. And the real data of aggregated primary energy consumption of China from 1990 to 2008 are applier to test the forecaster. The forecasting procedure includes three steps.

1) First step

According to the proposed model, we will observe the higher-ranked statistical information of the influencing factors on primary energy consumption. The computed statistical data are listed in TABLE I, from which we find that the six influencing factors are all sub-Gaussian. But, if combining skewness with kurtosis, we will find they all shows non-Gaussian properties. But under the condition of small samples and six influencing factors, LS-VSM is an effective forecasting method. And all influencing factors have more obvious characteristics. That is consistent with the foregoing quantitative analysis.

2) Second step

Here LS-SVM is trained by the expected energy consumption and their influencing factors. In this paper, the Gauss kernel function is selected as kernel function and the corresponding parameters are determined by cross-validation method. The relative forecasting results are illustrated in TABLE II, from which we can find good performance of the proposed model.

3) Performance criteria

After the regression forecasting model is trained, we should select criteria to test performance of the proposed model. Reference [11] told us that the criterion used to select the most appropriate model is to maximize the goodness-of-fit using the simplest model or combination of models. We select the absolute percentage error (APE) and the mean absolute percentage error (MAPE) are used, which are defined as,

$$APE = \frac{\left| E_i^* - E_i \right|}{E_i} \times 100\% \quad (8)$$

$$MAPE = \frac{1}{N} \sum_{N} APE \quad (9)$$

where E_i is the real energy consumption, E_i^* is the predicted value of E_i by using the proposed model and N is the number of years in the predicting period. The smaller the APE value and MAPE value, the closer are the predicted values to the actual values.

At the beginning, the adjustment of parameters is somewhat large to find the optimal baseline of the parameters. After that, each change in parameter is small. This process continues until all MAPE from test results are less than the acceptable error. In this paper, the accepted error range should be 5% or so.

IV. CONCLUSION

In this paper, economic indicators, such as real GDP, population, gross exports, gross imports and government expenditure, are selected as influencing factors. A novel forecasting model of energy consumption, LS-SVM regression

prediction model, is presented, which makes a full of use of higher-ranked statistical information of input influencing factors to perform factor analysis. Although, all the influencing factors shows sub-Gaussian characteristics, if combining skewness with kurtosis, we will find LS-SVM is an effective method to forecast energy consumption under the condition of small samples and many influencing fators. From the forecasting results, we can conclude that the proposed model of LS-SVM shows promising features to become a more reliable yet functional prediction tool for forecasting energy consumption.

TABLE II
THE FORECASTING RESULTS OF THE PROPOSED MODEL

Year	Real value (10^6tce)	Forecasted value(10^6tce)	Relative error (%)
1990	987.0	1018.78	3.22
1991	1037.8	1013.51	-2.34
1992	1091.7	1137.88	4.23
1993	1159.9	1101.79	-5.01
1994	1227.4	1246.67	1.57
1995	1311.8	1368.73	4.34
1996	1389.5	1452.03	4.5
1997	1378.0	1318.33	-4.33
1998	1322.1	1279.00	-3.26
1999	1338.3	1405.21	5.00
2000	1385.5	1423.05	2.71
2001	1432.0	1392.19	-2.78
2002	1518.0	1583.58	4.32
2003	1749.9	1831.62	4.67
2004	2032.3	1942.06	-4.44
2005	2246.8	2319.60	3.24
2006	2462.7	2558.50	3.89
2007	2655.8	2529.12	-4.77
2008	2850.0	2987.94	4.84

ACKNOWLEDGMENT

The author will thank three anonymous contributors for their useful comments and suggestions.

REFERENCES

[1] Joanne Evans and Lester C, Hunt, International Handbook on the Economics of Energy. Edward Elgar Publishing Limited, 2009.

[2] Soner Haldenbilen, Halim Ceylan, "Genetic algorium approach to estimate transport energy demand in Turkey", Energy policy. pp. 89-98, 2005.

[3] Zong Woo Geem, William E. Roper, "Energy demand estimate of South Korea using artifical neural network", Energy Policy. pp. 4049-4054, May 2009.

[4] Halim Ceylan, Harrun Kemal Ozturk, "Estimating energy demand of Turkey based on economic indicators using genetic algorithm approach", Energy Conversion and management. pp. 2525-2537, Dec. 2004.

[5] M. Duran Toksari, "Ant coloby optimization approach to estimate energy demand of Turkey", Energy Policy. pp. 3984-3990, Mar. 2007.

[6] Volkan Ş. Ediger, Sertaç Akar, "ARIMA forecasting of primary energy demand by fuel in Turkey", Energy Policy. pp. 1701-1708, Jul. 2007.

[7] H.T.Pao, "Forecasting energy consumption in Taiwan using hybird nonlinear models", Energy. pp. 1438-1446, Aug. 2009.

[8] Bing Dong, Cheng Cao, Siew Eang Lee, "Applying support vector machines to predict building energy consumption in tropical region", Energy and Building. pp. 545-553, Sep. 2005.

[9] J.A.K. Suykens, J. Vandewalle, and B.De Moor, "Optimal control by least squares support vector machines," pp: 23-35, 2001.

[10] Karl Yuxiang, and Zhongchang Chen, "Government expenditure and energy intensity in China," Energy Policy. pp: 691-694. Nov. 2009.

[11] N. Draper, and H. Smith, Applied Regreeion Aanalysis, second ed. John Wiley and Sons, New York, 1981.