Regression model for heat consumption monitoring and forecasting

Tatyana Dobrovolskaya[1*], and *Valery* Stennikov[1]

[1]Melentiev Energy Systems Institute of Siberian Branch of the Russian Academy of Sciences (ESI SB RAS), Pipeline Energy Systems Department, 130, Lermontov str., Irkutsk, Russia, 664033

Abstract. Heat supply is socially and economically important in our country. In this regard, high-quality monitoring and planning of the development of heat supply systems are a strategic vector of scientific research. This paper is focused on the studies demonstrating how to choose a methodological approach to describe changes in heat consumption in the retrospective. The change in heat consumption is described using multiple regression models. In the first part of the paper, the parameters for the regression model are determined and a statistical analysis of the obtained model is performed. In the second part of the paper, to eliminate the multicollinearity of the regression equation, the number of dependent variables in the model is reduced. A statistical analysis of the new regression model and the exponential regression model are carried out. The heat consumption values obtained using these models are compared with the statistical data. The conclusions about the quality of the obtained regression models are made. In the third part of the article, we make a forecast of heat consumption in the medium term by using a linear regression model and an exponential model.

1 Introduction

This paper is a continuation of the research on heat consumption monitoring and forecasting [1]. It is devoted to the generalization and analysis of data on energy consumption and economic indices, which are represented by time series. The studies were carried out to determine the closeness and nature of the relationship between the selected indices and the type of heat consumption dependence on the selected variables.

An analysis of the literature dealing with the prediction of energy consumption [2,3] indicates that the energy consumption is most often calculated using specific indicators for planning. However, the average heat consumption in residential buildings often does not reflect the real picture, as it does not take into account the climatic features of the region, the actual state of building envelopes, quantitative and qualitative characteristics of energy-saving measures. In industry, the forecast is based on indicators of energy intensity of various industries, that are adopted based on the foreign experience. In this case, it is quite difficult to draw analogies due to significant differences in economic development and technological potential of the countries. The researchers from other countries often present

* Corresponding author: makarova@isem.irk.ru

the studies on the use of regression models to analyse the current state and predict the levels of energy consumption [4-10]. They make an emphasis on the forecasts of electricity consumption for a short period of time (some hours or days).

This paper presents a research into a methodological approach to describe a change in heat consumption for a retrospective-calculated period. The coefficients of the multiple regression model of heat consumption are determined. A comparison of linear and nonlinear regression models of the heat consumption is performed. The research was conducted at the highest hierarchical level. The total heat consumption of the country was taken as the research object.

2 Multiple regression models of heat consumption

2.1 Determination of variables for regression model

A regression analysis is one of the most widely used statistical tools to describe the variations in a dependent variable (annual heat consumption) with independent variables used as inputs for the functions. The main goal of the regression analysis is to find an appropriate mathematical model and determine the best coefficients of the model from the given data. The major objective of the study was to identify the input parameters for the models that would describe the heat consumption in the best way.

In the first stage, we selected the parameters for the multiple regression model. A specific feature of regression models is the need to have prospective estimates of regression equations (independent variables). Consequently, it is much more difficult to choose the parameters as independent variables in the regression equation. The most authoritative forecasts of social and economic indicators are forecasts of socio-economic development, made by the Ministry of Economic Development of the Russian Federation.

The research focused on the following indicators: installed electric capacity of power plants (N, 10^6 kW), electricity consumption (W, 10^9 kW·h), heat consumption (Q, 10^6 Gcal), population (P, 10^6 people), investments in electric power industry (including district heating) (I, 10^9 RUB), gross domestic product (GDP) (GDP, 10^9 RUB), time interval (T, year). The considered indicators are represented by time series from 1990 to 2014 [11, 12].

An initial analysis of the system of the considered indicators is to identify strong relationship between the indicators in the sample. Correlation analysis allows us to determine if it is necessary to include the indicators in the multiple regression model. The correlation coefficient is a mathematical measure of correlation between two values. We use a linear correlation coefficient (Pearson correlation coefficient) [13, 14], because the indicators are represented by absolute values. Table 1 demonstrates linear correlation coefficients of selected indicators.

Table 1. Matrix of correlation coefficients of the considered indicators from 1990 to 2014.

Indicator	T	N	W	Q	P	I	GDP
T	1						
N	0,8559	1					
W	0,4079	0,6007	1				
Q	-0,928	-0,6522	-0,0667	1			
P	-0,858	-0,5382	-0,3786	0,8358	1		
I	0,8818	0,8946	0,4088	-0,7454	-0,606	1	
GDP	0,987	0,8698	0,4524	-0,8952	-0,8542	0,8905	1

The object of our research is heat consumption, therefore we can draw conclusions about the existence of relationship between heat consumption for the population, GDP and time (the values of the time period). Along with the data of the study, the autocorrelation coefficients of the considered indicators are calculated. Various methods are used to reduce the autocorrelation. The methods aim to exclude the main development trends in the initial data, i.e. linear trends. In this paper, time is introduced in the multiple regression equation as an independent variable [15].

Linear one-parameter regression models of the selected indicators were considered earlier in [1]. Further in the paper, we will consider in more detail the dependence of heat consumption on the selected indicators, by using a multiple linear regression.

2.2 A multiple linear regression model of heat consumption

A linear regression equation in a general form can be written as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon \tag{1}$$

where **X** – vector of independent (explanatory) variables, **B** – vector of parameter coefficients (to be determined), ε – random error (deviation), Y – dependent (explained) variable, m – the number of explanatory variables, n – the number of observations.

In order to uniquely solve the problem of finding the regression equation coefficients, the inequality $n \geq m+1$ should be met. To estimate a multiple linear regression, the statistical reliability requires that the number of observations be at least three times greater than the number of indicators to be estimated [13-15]. In our case, in this connection, further calculations are aimed at finding the regression coefficients. Equation (1) in the case of linear multiparameter regression of heat consumption can be written as:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + b_4 x_{4i} + b_5 x_{5i} + b_6 x_{6i} + e_i \tag{2}$$

where y – values of the explained variable (heat consumption); x – values of explanatory variables ; b – coefficients of the considered linear multiparameter regression; e – values of deviations of sample values of the explainatory variable from the values obtained from the regression equation.

The most common method of estimating the coefficients of the multiparameter linear regression equation is the least-squares method (LSM) [13-16]. The essence of the method consists in minimizing the sum of squares of deviations of the observed values of the dependent variable from its values obtained from the regression equation. Based on our requirement, the standard error should be minimal, which can also be written in the following form:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (\varepsilon_i)^2 \to \min . \tag{3}$$

Then the equation of multiple regression for description of a change in heat consumption will have the following form:

$$Q_i^{\text{mod}} = 3093{,}67 - 50{,}31 \cdot T_i + 10{,}47 \cdot N_i + 0{,}775 \cdot W_i - 24{,}46 \cdot P_i + 0{,}199 \cdot I_i - 0{,}048 \cdot GDP_i + e_i .$$

A statistical analysis was made to estimate the statistical error, the statistical significance of regression coefficients and the overall quality of the model. The statistical error of the model in the retrospective period was about 1%. The coefficients of the multiple linear regression model are statistically significant. The overall quality of the

model is evaluated by determination coefficient R^2. For the considered multiple regression equation, the determination coefficient is 0.994. The closer this ratio to 1, the closer the regression equation describes the behaviour of the dependent variable *Y*.

The determination coefficient, however, can be quite high in the presence of coinciding trends in the variables under consideration, and as a consequence a high degree of multicollinearity of the variables. The main method to eliminate multicollinearity of the regression model is the method of excluding variables. Further work with the obtained regression model will be aimed at excluding correlated variables.

The studies have indicated that the change in heat consumption is best described by the regression model with two variables (GDP and time), which can be written in the following form:

$$Q_i^{\mathrm{mod}1} = a_0 + a_1 T_i + a_2 GDP_i + e_i^1 . \quad (4)$$

We calculated the values of the regression coefficients *a* of equation (4) in accordance with the LSM, and this equation has the following form:

$$Q_i^{\mathrm{mod}1} = 1971{,}5436 - 59{,}887 \cdot T_i + 0{,}01297 \cdot GDP_i + \varepsilon_i^1 .$$

A statistical analysis of the heat consumption model (4) was performed. Statistical error of the regression equation is 94.13 or 4.1%. The regression equation coefficients are statistically significant. The determination coefficient of the equation is 0.877. Based on the calculations, we can conclude that the obtained regression equation explains 87.7% of variation in the dependent variable (heat consumption).

2.3 A non-linear regression model of heat consumption

Changes in the heat consumption can also be described by an exponential equation. The variables (regressors) as well as in the previous model, will be represented by time and GDP indicators. For this study, however, we will not use the absolute values of heat consumption and GDP, but the base growth rates of these indicators. The values of the indicators in the final year are chosen as the base value *n*, due to the fact that logarithm will be made exactly according to the basic growth of heat consumption. The general equation of exponential dependence of heat consumption has the following form:

$$Q^{\mathrm{mod}2} = e^{c_0 + c_1 \cdot T + c_2 \cdot GDP^2} \cdot \varepsilon_i^2 , \quad (5)$$

where $Q_i^2 = \dfrac{Q_i}{Q_n}$ – the basic growth of heat consumption, $i = 1,2,..n$; $GDP_i^2 = \dfrac{GDP_i}{GDP_n}$ – the basic growth of GDP, $i = 1,2,..n$; c_0, c_1, c_2 – the regression equation coefficients.

The LSM is used to calculate the values of the coefficients of equation (5). The equation of heat consumption depending on time and GDP will be as follows:

$$\ln Q_i^{\mathrm{mod}2} = 0{,}2463 - 0{,}0315 \cdot T_i + 0{,}4681 \cdot GDP_i + \ln \varepsilon_i^2 .$$

A statistical analysis was performed for this model of heat consumption. Based on the calculation results we estimated the statistical errors and the statistical significance of the regression equation coefficients. The determination coefficient R^2 was 0.91. These calculations are correct for the logarithm of the basic heat consumption growth. Further, after reduction to exponential function and multiplication by value of heat consumption in base year, the estimated determination coefficient was 0.9071.

3 Case study: heat consumption monitoring and forecasting

The considered regression models quite well describe the change in heat consumption in the retrospective period. These regression models can be used to make a short-term forecast of heat consumption.

In accordance with [17, 18], the forecast of GDP change was adopted and heat consumption levels were calculated using the obtained regression models. Two options of the GDP change forecast were approved, and interval estimates of heat consumption for a prospective period were obtained. The length of the forecast period covered by the regression model depends directly on the retrospective period for which the model is derived. It is assumed that the forecast period should not exceed 1/3 of the retrospective period. According to this circumstance, the heat consumption forecast is made until 2020. Figure 1 presents the results of the heat consumption calculation in accordance with the two options of economic development.

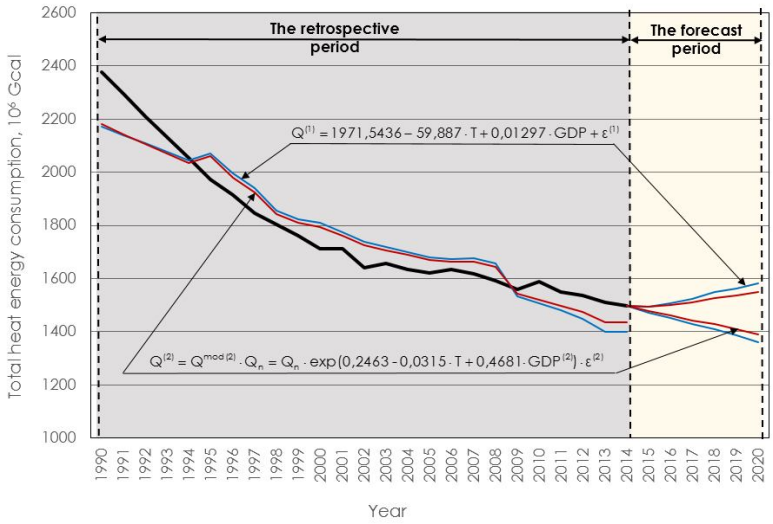

Fig. 1. The use of a multiple linear regression model and non-linear regression model for heat consumption monitoring and forecasting

4 Conclusions

The studies have demonstrated that in accordance with the economic development forecast [17, 18] based on the multiple linear regression model, a decrease in heat consumption by 2020 may be 9.2% in the first option of economic development or it can increase by 5.6% in the second option compared to 2014. According to the exponential model used to predict heat consumption, it can decrease by 2020 to 7.1% in the first option or increase by 3.4% in the second option compared to 2014. The change range of heat consumption, according to the exponential model, has narrower boundaries, which allows us to more accurately determine the possible level of heat consumption. The discrepancy between the projected values of heat consumption at the level of 2020, using linear and nonlinear regression models, is no more than 2% (32×10^6 Gcal). Thus, we can conclude that the obtained regression models can be used for monitoring and medium-term forecasting of heat consumption. In the future, we plan to continue the research in this area. Our plan is to

expand the models with the explanatory variables, and describe the structure of heat consumption, using regression models.

The research was performed at Melentiev Energy Systems Institute SB RAS in the framework of scientific projects III.17.4.1 №AAAA-A17-117030310432-9 and III.17.4.3 №AAAA-A17-117030310437-4 of the Fundamental Research Program of SB RAS.

References

1. T.V. Dobrovolskaya. Monitoring of heat consumption levels using regression models. System research in power engineering. Proceedings of young scientists of ISEM SB RAS. **Vol.45** 124-130 (2015)
2. A.M. Mastepanov. *A fuel and energy complex of Russia at the turn of the century: state, problems and prospects*. 793 (2010)
3. A.S. Nekrasov. *An analysis and forecasts of the fuel and energy sector development. Selected works*. 592 (2013)
4. Ming Meng, Dongxiao Niu. Annual electricity consumption analysis and forecasting of China based on few observations methods. Energy conversion an management, **52**, 953-957 (2011)
5. N. Fumo, M.A. Rafe Biswas. Regression analysis for prediction of residential energy consumption. Renewable and sustainable energy reviews. **47**, 332-343 (2015)
6. S. Smeekes, E. Wijler. Macroeconomic forecasting using penalized regression methods. International journal of forecasting. **34**, 408-430 (2018)
7. T. Catalina, V. Iordache, B. Caracaleanu. Multiple regression model for fast prediction of the heating energy demand. Energy and buildings. **57**, 302-312 (2013)
8. Tingting Fang, Risto Lahdelma. Evaluation of a multiple linear regression model and SARIMA model in forecasting heat demand for district heating system. Applied Energy. **179**, 544-552 (2016)
9. G.J. Tsekouras, E.N. Dialynas, N.D. Hatziargyriou, S. Kavatza. A non-linear multivariable regression model for midterm energy forecasting of power systems. Electric power systems research. **77**, 1560-1568 (2007)
10. V.Bianco, O. Manca, S. Nardini. Electricity consumption forecasting in Italy using linear regression models. Energy. **34**, 1413-1421 (2009)
11. Russian statistical yearbook. 2013. 717 (2013)
12. Federal state statistics service (http://www.gks.ru/)
13. E. Ferster, B. Renz. *Methods of correlation and regression analysis*. 302 (1983)
14. S.A. Borodich. *Econometrics: study guide*. 408 (2001)
15. G.A. Ivashchenko, G.S. Kildishev, R.A. Shmoilova. *Statistical study of the main trends of development and interrelation in the ranks of dynamics*. 168 (1985)
16. S.V. Solodusha. *Typical problems of the basic course of econometrics*. 42 (2007)
17. *Energetics of Russia: a view to the future* (Substantiating materials to the Energy strategy of Russia to 2030). 616 (2010)
18. The energy strategy of Russia to 2030 (https://minenergo.gov.ru/node/1026)