2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) | 978-1-7281-5244-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ITAIC49862.2020.9338976

Research On Prediction Of Electricity Consumption In Smart Parks Based On Multiple Linear Regression

Zhiyang Zhao,Yue Peng ,Xianxun Zhu,Xiong Wei,Xu Wang, Jiancun Zuo*
School of Computer and Information Engineering, Shanghai Polytechnic University
Shanghai ,China
13509703508@163.com ,pengyue@sspu.edu.cn,1591694407@qq.com ,291176867@qq.com, weixiongcn@163.com
*Corresponding author, E-mail: jczuo@sspu.edu.cn

***Abstract*—With the construction of cities and the development of smart parks, the energy efficiency improvement and the low-carbon operation realization have been developed by the people-oriented smart parks as a recent stage strategy based on the park informatization. The multiple linear regression method is adopted in this paper for the establishment of time -power consumption relationship model. The power consumption in the park can be predicted finally with the accuracy rate of 0.826. This model can serve for managers and decision makers further with strong data support.**

Keywords—multiple linear regression; power consumption; prediction

I. INTRODUCTION

At present, China is in the peak period of urban construction. The rapid development of urban construction leads to greater energy demand. The energy loss of huge public structures is much higher than that of residential structures, resulting in energy waste and high energy consumption. Therefore, the state proposes to implement an energy consumption limit usage system for huge public structures. Therefore, the state proposes to implement an energy consumption quota system for large public buildings. The establishment of the energy consumption quota system must first obtain the energy consumption level and characteristics of large public buildings. At this stage, it is not feasible to install a real-time monitoring system for each large public building. Therefore, this article considers using energy consumption models to study the energy consumption of large public buildings, and predicts the energy consumption of electricity to provide for the establishment of an energy consumption quota system.[1]

Many scholars at home and abroad have conducted in-depth research on energy consumption analysis and prediction, which can be roughly divided into two directions: First, building structure is the research object, mostly based on thermodynamic theory, comprehensively considering internal and external disturbance factors, and giving thermodynamic equations And solve.[2-4] The second is to analyze energy consumption data, such as using artificial intelligence and data mining algorithms to find the relationship between energy consumption data and influencing factors to make predictions.[5-8]

In this article, it uses a multiple linear regression model to predict the energy consumption of smart parks. Through the stepwise regression of multiple independent variables, the accuracy of the prediction is improved. The second part of the article details the basic principles of the multiple linear regression model; then the third part describes the model establishment process; in the fourth part, the experimental results are given and comparative analysis and discussion are made; finally, the research results are summarized and prospected.

II. SYSTEM MODEL INTRODUCTION

The linear regression model is a classic statistical model. The application scenario of the model is to predict a continuous dependent variable based on the known independent variables. From the perspective of data mining, linear regression is a supervised mathematical model, in other words, the independent variable x and the dependent variable y must be available in the modeling process. Linear regression models are divided into unary linear regression and multiple linear regression models.

The univariate linear regression model is also named the uncomplicated linear regression model, which means that the model contains only one independent variable and one dependent variable. The data set used for modeling can be expressed as $\{(x_1,y_1),(x_2,y_2), \ldots ,(x_n,y_n)\}$. Among them, x_i represents the i-th value of the independent variable x, yi represents the i-th value of the dependent variable y, and n represents the sample size of the data set.

Based on the study of linear correlation, the data change relationship between two or more independent variables and one dependent variable is called multiple linear regression analysis, and the mathematical formula obtained is called multiple linear regression model. The multiple linear regression model is an extension of the unary linear regression model.[9]

978-1-7281-5244-8/20/$31.00 ©2020 IEEE

The purpose of multiple linear regression is to establish a linear regression equation of multiple elements of the dependent variable on multiple independent variables based on the test values of the dependent variable and multiple independent variables; verify and analyze the significance of the simple linear impact of each independent variable on the dependent variable. Select independent variables that only have a significant linear effect on the dependent variable, set up the optimal multiple linear regression equation; evaluate the relative importance of each independent variable's impact on the dependent variable, and determine the deviation of the optimal multiple linear regression equation. Fig.1 shows the prediction process of the multiple linear regression model [9].

Suppose the dependent variable y and the independent variables $x_1, x_2, \ldots, x_{m-1}$ have n sets of actual observation data, see Table I. y is an observable random variable, which is affected by m-1 non-random factors $x_1, x_2, \ldots, x_{m-1}$ and ε random factors. If y has the following linear relationship with $x_1, x_2, \ldots, x_{m-1}$ [9]:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_{m-1} x_{m-1} + \varepsilon \qquad (1)$$

Among them, $\alpha_{0,}\alpha_{1,}\alpha_{2,\ldots,}\alpha_{m-1}$ is m unknown parameters; the unknown random variable whose mean value of ε is zero and degree of deviation $\sigma^2 > 0$ is named error term, and $\varepsilon \sim N(0, \sigma^2)$ is usually supposed.

TABLE I. OBSERVATION DATA TABLE

variable	y	x_1	x_2	…	x_{m-1}
1	y_1	x_{11}	x_{21}	…	x_{m1-1}
2	y_2	x_{12}	x_{22}	…	x_{m2-1}
⋮	⋮	⋮	⋮	…	⋮
n	y_n	x_{1n}	x_{2n}	…	x_{mn-1}

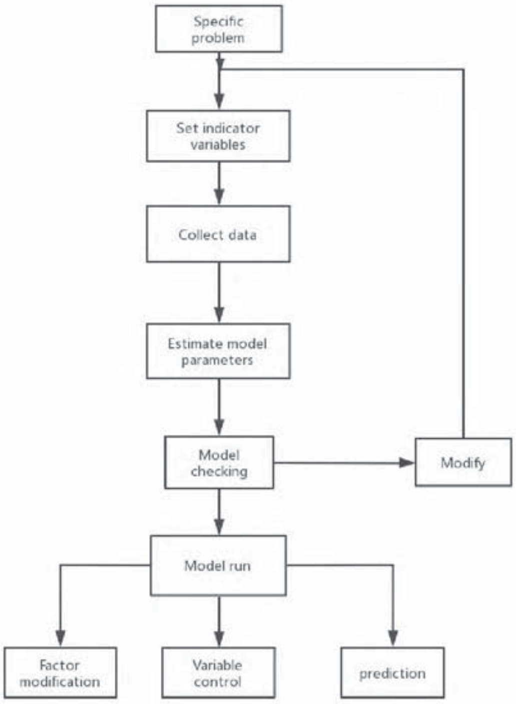


Fig. 1. The predictive model of multiple regression.

For n (n ≥ p) individual tests, n sets of data are obtained [9]:

$$\begin{cases} y_1 = \alpha_0 + \alpha_1 x_{11} + \ldots + \alpha_{m-1} x_{1m-1} + \varepsilon_1 \\ y_2 = \alpha_0 + \alpha_1 x_{21} + \ldots + \alpha_{m-1} x_{2m-1} + \varepsilon_2 \\ \qquad \vdots \\ y_n = \alpha_0 + \alpha_1 x_{n1} + \ldots + \alpha_{m-1} x_{nm-1} + \varepsilon_n \end{cases} \qquad (2)$$

Among them, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are not related of each other and obey $\varepsilon \sim N(0, \sigma^2)$ distribution. Order,

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m-1} \\ 1 & x_{21} & x_{22} & \vdots & x_{2m-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm-1} \end{pmatrix}_{n\times m},$$

$$\alpha = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{m-1} \end{pmatrix}_{m\times 1}, \quad \varepsilon = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$$

The formula (1) is expressed in matrix form as [9]:

$$\begin{cases} Y = X\alpha + \varepsilon \\ \varepsilon \sim N\ (1,\ \sigma^2 I_n) \end{cases} \tag{3}$$

III. Model establishment

A. Selection of independent variables

When performing regression analysis, the selection of independent variables is the basis of regression. Generally speaking, two principles should be considered when selecting independent variables: 1) The independent variable must be closely related to the dependent variable; 2) A strong linear relationship should be avoided as far as possible between independent variables.

The data set used in this paper is the park power consumption measured and controlled by the smart park electromechanical control management platform. Taking into account the representativeness, completeness and data availability of influencing factors in the experiment, the final selected variables are shown in Table II. The data set contains 8 variables, namely time, active power, reactive power, voltage, current, electric power for lighting, electric power for water heaters, and electric power for air conditioners.

TABLE II. Part of the data set

datetime	*active_power*	*reactive_power*	*Voltage*	*Global_intensity*	*metering_1*	*metering_2*	*metering_3*
2010/12/16 13:03	374	10.8	236.93	16.4	16	18	28.333332
2010/12/16 13:04	492.8	20.2	235.01	21	37	16	29.133331
2010/12/16 13:05	605.2	19.2	232.93	26.2	37	17	46.86667
2010/12/16 13:06	675.2	18.6	232.12	29	36	17	59.533333
2010/12/16 13:07	647.4	14.4	231.85	27.8	37	16	54.9
2010/12/16 13:08	630.8	11.6	232.25	27	36	17	52.13333

B. Estimation of model parameters

After the regression model is determined, the unknown parameters of the model are estimated by the universal least square method. For regression problems that do not meet the basic assumptions of the model, methods such as ridge regression, principal component regression, and partial least square estimation are used. But these are based on the universal least squares method [9]. The equation $(X^T X)\hat{\beta} = X^T Y$ or $A\hat{\beta} = B$ in matrix form assumes that the coefficient matrix A is full rank, and the regression coefficient β obtained by solving the above matrix equation is estimated by the least square method:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{4}$$

It is the least square method of regression coefficient β.

C. Model checking

After the unknown parameters of the model are estimated, a regression model is initially established, and the hypothesis of the linear relationship between the dependent variable and multiple independent variables must be tested for significance.

The fit is generally used to verify the fit of the sample regression line to the independent variables. In the unary linear regression equation, the coefficient of determination R^2 is used to estimate the degree of fit of the estimated equation to the sample observation value, which is also applicable to the multiple linear regression equation. The mathematical model is as follows [9]:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{m}\left(y_i - \bar{y}\right)^2}, \quad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i \tag{5}$$

y_i is the regression value on the i-th sample point $(x_1, x_2, \ldots, x_p)$. $\bar{y}$ is the sample average of y of the sample. If the value of R^2 is closer to 1, it indicates that the regression equation fits the actual observations better. Mean square error MSE and root mean square error RMSE can also be used as measurement standards. The greater the deviation between the real value and the predicted value, it indicates that the prediction effect is not accurate [9].

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2 \tag{6}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2} \tag{7}$$

IV. Analysis of Experimental Results

The experiment uses the sklearn framework, which is a machine learning tool based on the Python language. The data types it requires are numerical and nominal data, which can perform data mining and data analysis simply and efficiently.

The experiment first divides the data set into a training set and a test set at a proportion of 8:2, and then standardizes the data. Among them, fit_intercept: Boolean, the default is True. If the parameter value is True, it means that the training model needs to add an intercept term; if the parameter is False, it means that the model does not need to add an intercept term. Finally, perform polynomial expansion.

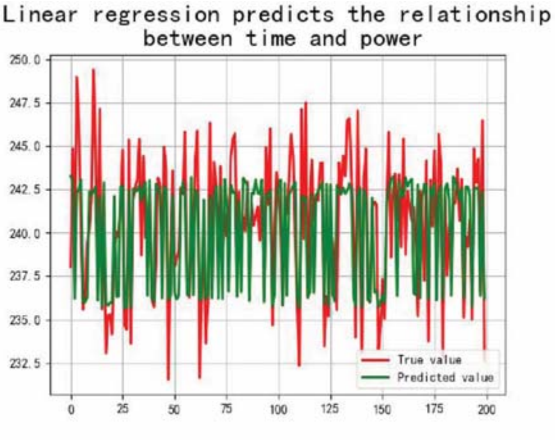


Fig. 2 Time-power diagram

Fig.2 shows the relationship between linear regression prediction time and power, where the x-axis represents time and the y-axis represents power. The red and green lines in the figure are the display of the true value and the predicted value, respectively. From Fig.2 above, it can be reflected that the linear relationship between the time and power of the data set is not very large, there is no obvious linear relationship, and the accuracy is only 0.485.

Fig.3 is a linear regression predicting the relationship between power and current. According to physical knowledge, there is a certain relationship between power and current. The experimental results in Figure 3 also verify this. The degree of fit is high, and the current prediction accuracy is 0.992. Table III below shows the experimental result data.

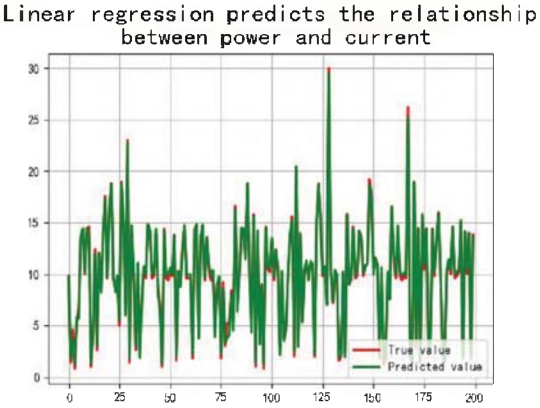
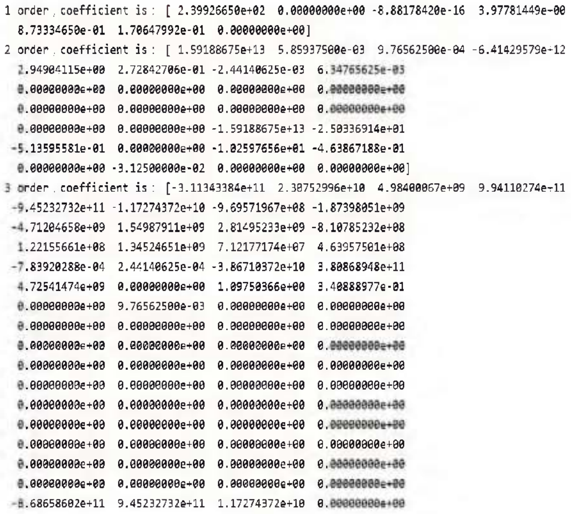


Fig. 3 Power-current diagram

TABLE III. Experimental Results

Accuracy	0.48506578175142046
MSE	7.62171282580588
RMSE	2.760744976597056
Current prediction accuracy rate	0.9920420609708968
Current parameters	[5.07744316 0.07191391]

As shown in Fig.2, there is no obvious linear correlation between time and power, so the experiment was expanded by polynomial, using the pipeline function pipeline module. The Pipeline module can be given multiple different operations and execute them in the order from front to back. During the execution of the Pipleline object, this article assumes that the data is first expanded by polynomial, and then stepwise regression. The results are shown in Fig.4 and 5.



```
1 order , coefficient is : [ 2.39926650e+02  0.00000000e+00 -8.88178420e-16  3.97781449e-00
  8.73334650e-01  1.70647992e-01  0.00000000e+00]
2 order , coefficient is : [ 1.59188675e+13  5.85937500e-03  9.76562500e-04 -6.41429579e-12
  2.94904115e+00  2.72842706e-01 -2.44140625e-03  6.34765625e-03
  0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00 -1.59188675e+13 -2.50336914e+01
 -5.13595581e-01  0.00000000e+00 -1.02597656e+01 -4.63867188e-01
  0.00000000e+00 -3.12500000e-02  0.00000000e+00  0.00000000e+00]
3 order , coefficient is : [-3.11343384e+11  2.38752996e+10  4.98400067e+09  9.94110274e-11
 -9.45232732e+11 -1.17274372e+10 -9.69571967e+08 -1.87398051e+09
 -4.71204658e+09  1.54987911e+09  2.81495233e+09 -8.10785232e+08
  1.22155661e+08  1.34524651e+09  7.12177174e+07  4.63957501e+08
 -7.83920288e-04  2.44140625e-04 -3.86710372e+10  3.80868948e+11
  4.72541474e+09  0.00000000e+00  1.09750366e+00  3.40888977e-01
  0.00000000e+00  9.76562500e-03  0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
 -8.68658602e+11  9.45232732e+11  1.17274372e+10  0.00000000e+00
```

Fig. 4 Display of partial result coefficients

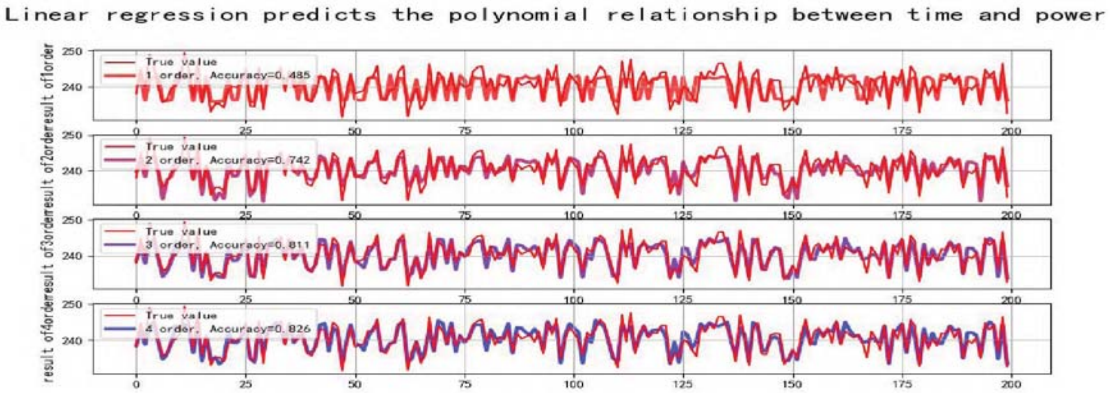


Fig. 5 Time-power polynomial relationship

Fig.4 is a partial display diagram of the coefficients generated by each order of regression. Fig.5 is the polynomial relationship between linear regression prediction time and power. It is divided into four parts according to the set order. From top to bottom, the results of the first, second, third and fourth orders are compared. The x-axis of the part represents time, and the y-axis of the ordinate represents power.

The experimental results show that the higher the order, the more coefficients are generated. Figure 4 shows some of the result coefficients. The first-order coefficients are 7 terms, the second-order coefficients are 28 terms, and the third-order coefficients are 84 terms, and the accuracy is Also significantly improved, as shown in Table IV, the accuracy rate gradually tends to 1. The closer the value of is to 1, the better the fit of the regression equation to the actual observations

TABLE IV. COMPARISON OF RESULTS ACCURACY

1st order accuracy	0.485	*2nd order accuracy*	0.742
3rd order accuracy	0.811	*4th order accuracy*	0.826

V. CONCLUSIONS AND PROSPECTS

Based on the electricity consumption data of the smart park, this paper conducts multiple statistics and analysis on the data, establishes a multiple linear regression equation of the dependent variable against multiple independent variables based on the actual observations of the dependent variable and multiple independent variables, and forms a predictive model . The biggest advantage of this model is that by first selecting a part of the independent variables for regression testing, and then gradually regressing the unselected independent variables under limited conditions, and the effect is significantly improved. This method is not only intuitive, but the test results are true and reliable, which can provide a strong data basis for the energy consumption quota system.

Because the purpose of this article is to predict the energy consumption of smart parks, and energy consumption is not only related to electricity consumption, but also concerned with the surroundings, such as changes in the natural environment of the building itself and the use of personnel, so follow-up research will consider Whether energy consumption will be affected by more factors.

ACKNOWLEDGMENT

This work was supported by the Graduate Program Foundation of Shanghai Polytechnic University under Grant
No. EGD19YJ0103.

REFERENCES

[1]. Research Center for Building Energy Efficiency, Tsinghua University. China Annual Development Research Report on Building Energy Efficiency 2009. Beijing: China Building Industry Press. 2009

[2]. Crawleya D B. Lawrieb L K. Frederick C Energyplus: creating a new-generation building energy simulation program, 2001 (4).

[3].Spatial regression analysis of domestic energy in urban areas[J]. Wei Tian,Jitian Song,Zhanyong Li. Energy. 2014

[4]. Hua Ben. The integrated innovation of China's urban building energy system[J]. Journal of South China University of Technology (Natural Science Edition), 2007(10): 111-116.

[5]. Yao Jian, Yan Chengwen, Ye Jingjing, etc. Building energy consumption prediction based on neural network. Building energy efficiency. 2007, (10): 31~33

[6]. Tian Wei, Wei Lai, Zhu Li, He Cheng, Sun Yu, Yang Song. Overview of research on urban-scale building energy consumption[J]. Building Energy Efficiency, 2016, 44(02): 59-64.

[7]. Ma Zhiliang, Teng Mingkun, Ren Yuan. Building energy consumption information model for big data analysis[J]. Journal of South China University of Technology (Natural Science Edition), 2019, 47(12): 72-77+91.

[8].Alberto Hernandez Neto, Flavio Augusto Sanzovo Fiorelli. Comparison Between Detailed Model Simulation and Artificial Neural Network for Forecasting Building Energy Consumption. Energy and Buildings. 2008,(40): 2169~2176

[9]. Wu Jinpei, Sun Deshan. Modern Data Analysis ［M］. Beijing: Mechanical Industry Press, 2006..