Feasibility Analysis of China's Electricity Supply and Carbon Neutrality Using LSTM-ARIMA Model

Chang Guo[1, #], Tianyi Bao[2, #, *], Ding Ma[3, #]

[1]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China, 221116

[2]School of Mechanical and Electrical Engineering, China University of Mining and Technology, Xuzhou, China, 221116

[3]School of Economics and Management, China University of Mining and Technology, Xuzhou, China, 221116

*Corresponding author: 19803132944@163.com

#These authors contributed equally.

Abstract. This article focuses on the modeling and algorithm design of China's photovoltaic power generation industry. This article mainly uses ARIMA-LSTM to study the influencing factors of China's power supply and predict its development trend; Using the GPCA-EWM evaluation model to evaluate the possibility of constructing a photovoltaic power station in a certain area. In terms of factor selection, The paper integrated the five major indicators of economy, production, consumption, transportation, and population for factor selection and data acquisition. Finally, the paper obtained feasibility score maps for various provinces in China, And based on this analysis, it is more suitable to build photovoltaic power stations in the northwest region of China.

Keywords: ARIMA-LSTM, Spearman correlation coefficient, SA-PSO.

1. Introduction

Due to the energy crisis and ecological problems are becoming more and more prominent, coupled with the photovoltaic power generation technology in the continuous improvement, perfect, the development of photovoltaic industry is strongly supported by various countries around the world, scientific and reasonable siting is a key step in the construction of solar photovoltaic power plant.

Domestic research on the siting of photovoltaic power plants is relatively small, which may be related to the late start of the photovoltaic industry in China. Ma Yifei[1] and others through the review of relevant information, to determine the main factors affecting, and through the hierarchical analysis method to determine the weight of each factor, and then combined with the GIS technology, and ultimately came up with the appropriate siting area. Duan Ying[2] and others in order to analyse the suitability of the construction of photovoltaic power plants in rocky desertification areas, the use of GIS technology on the number of hours of sunshine, slope, slope direction and rocky desertification areas for the superposition of the analysis, and then get the area suitable for photovoltaic project development.

In order to analyse the suitability of siting PV power plants in the eastern region of Morocco, Alami Merrouni[3] and others used GIS technology to assess the suitability of constructing PV power plants in the region in terms of solar energy resources, land availability, water resources and infrastructure. Uranchimeg[4] et al. used GIS technology to analyse the suitability of PV sites in Mongolia by applying hierarchical analysis to find out the weights of seven major influencing factors such as annual global horizontal radiation, average annual temperature, and altitude.

In this paper, the most suitable areas for PV power plants in China are selected by modelling the factors affecting power supply and the factors affecting the construction of PV power plants in China.

2. ARIMA-LSTM modelling to predict power supply trends

In order to study the relationship between factors that affect China's electricity supply and predict the development trend of China's electricity supply from 2024 to 2060, the paper use the method of calculating the Spearman coefficient to select the main influencing factors. Each influencing factor is a typical time series, and a weighted average prediction model, ARIMA-LSTM model, can be used for analysis and prediction, ensuring prediction accuracy while better fitting the data linearly and nonlinearly.

2.1. Data preprocessing

The paper downloaded data on China's gross domestic product, hydroelectric power production, thermal power production, electricity imports, and electricity exports from the official website of the National Bureau of Statistics from 2000 to 2022.

Then, by calculating the Spearman correlation coefficients between various influencing factors and power supply, the correlation between power supply and various influencing factors is measured.

$$r_s = 1 - \frac{6\sum (x_i - y_i)^2}{n(n^2-1)} \tag{1}$$

Among them, x_i and y_i are the ranks of the two variables according to their size, and n is the sample size. When n >20, statistical tests can be used.

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$
$$|t| > t_{a/2}(or\ p < a)shows\ that\ the\ r\ statistics\ are\ significant,$$
$$|t| \leqslant t_{a/2}(or\ p \geqslant a)shows\ that\ the\ r\ statistics\ are\ not\ significant; \tag{2}$$

The Spearman coefficient heatmap for each influencing factor is as Fig 1:

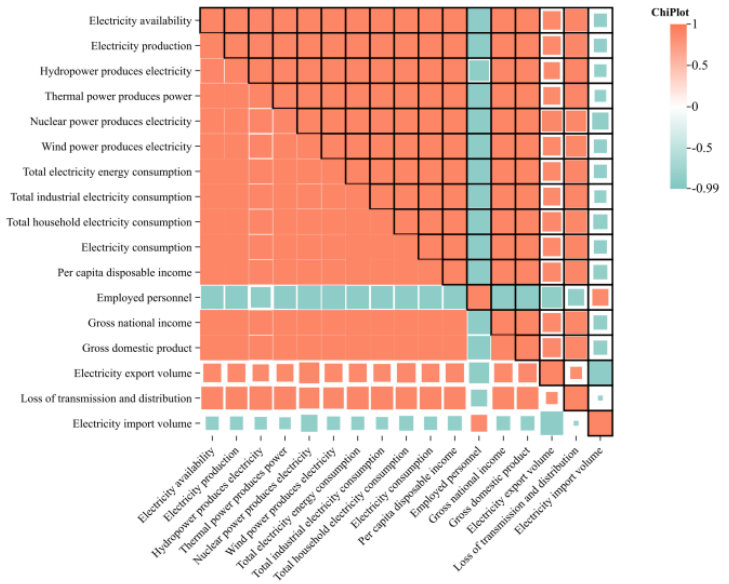


Fig 1 The Spearman coefficient heatmap for each influencing factor

Based on relevant literature, select the influencing factors with high Spearman correlation coefficient: Electricity generated by thermal power plants, residential electricity total consumption, hydroelectric power generation capacity, resident disposable income, Electricity generated by hydropower plants, and further analyze their relationship with power supply.Then, perform cubic

spline interpolation on the data of the selected influencing factors from 2000 to 2022 to obtain complete data.

2.2. Prediction of power supply based on ARIMA-LSTM model

The ARIMA model is a linear sequence prediction model that is limited by the input time series characteristics. It has good performance in handling stationary time series, but there is bias in its prediction results when dealing with non-stationary data. The LSTM model is a nonlinear model that can better handle the nonlinear parts in time series. The ARIM-LSTM hybrid model can improve the accuracy and robustness of predictions.

ARIMA model establishment

Draw a sequence diagram and perform a stationarity test to observe whether the sequence is stationary; For non-stationary sequences, first perform d-order differencing and convert them into stationary time series.

$$\Delta yx = y(x+1) - y(x), (x = 0, 1, 2, ...) \tag{3}$$

Obtain the autocorrelation coefficient (ACF) and partial correlation coefficient (PACF) of a stationary time series, and obtain the optimal order p and q by analyzing the autocorrelation and partial correlation graphs or searching through AIC and BIC.

ARIMA formula definition:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{4}$$

The ARIMA model prediction chart is as shown in Fig 2:

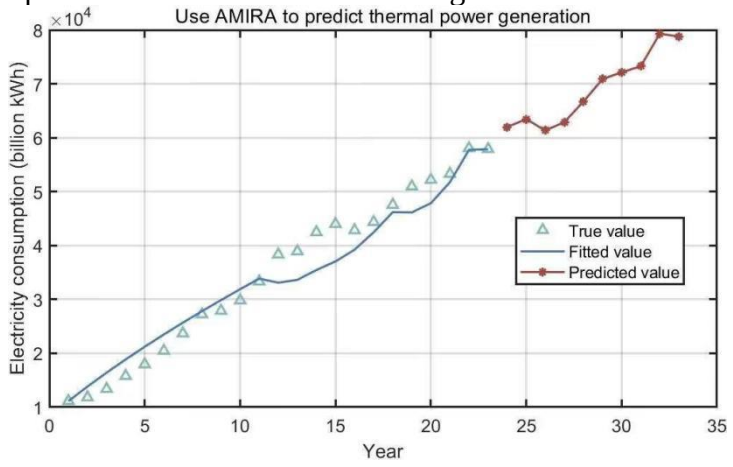


Fig 2 ARIMA model prediction chart

LSTM model establishment

LSTM [4]controls the transmission of information by introducing three gates: forget gate f_t, input gate i_t, and output gate o_t. The functions of the three doors are:

First, use the external state h_{t-1} from the previous moment and the input x_t from the current moment to calculate the values of the three gates and candidate state $\tilde{c}_t$;

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{5}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{6}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{7}$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{8}$$

Combining forget gate f_t and output gate i_t to update internal state c_t, also known as memory unit;

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{9}$$

Combined with output gate o_t, transmit the information of internal state c_t to external state h_t.

$$h_t = o_t \odot \tanh(c_t) \tag{10}$$

LSTM[5] unit structure and LSTM timing chart are as shown in Fig 3 and Fig 4.

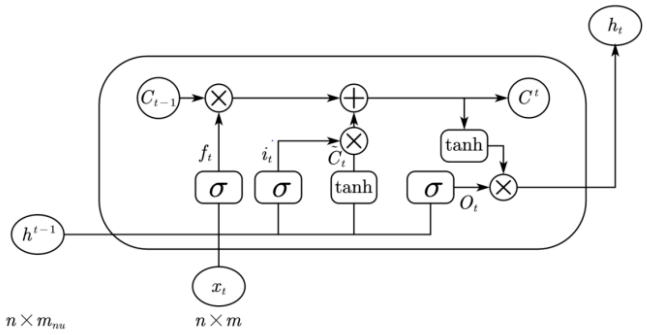


Fig 3 LSTM unit structure

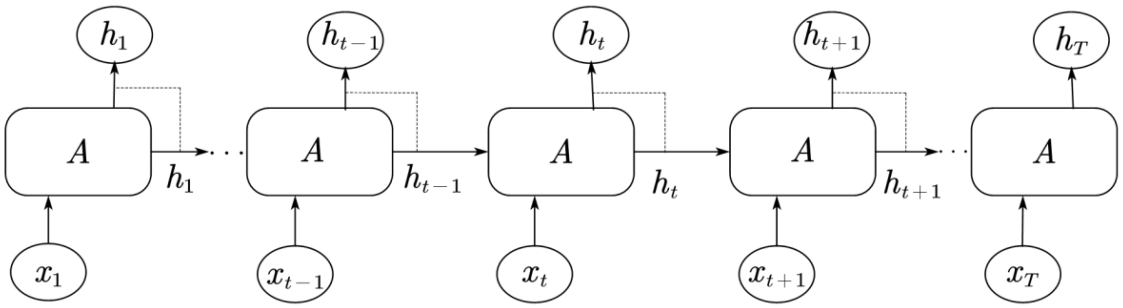


Fig 4 LSTM timing chart

ARIMA-LSTM model establishment

Firstly, the ARIMA model is used to filter out the linear trend prediction results from the influencing factor data, while the non-linear trend is retained in the residual value of the ARIMA model; Then, the residual value is transmitted to the LSTM model, and the LSTM model is trained to output the error prediction results of the nonlinear trend; Finally, the prediction results of the two models are superimposed to obtain the combined prediction model prediction results.

In the ARIMA model, the sequence value of the influencing factors at time t is Y_t, and the prediction result output by the ARIMA model is $\hat{Y}_t$. Therefore, the final residual result of the ARIMA model input to the LSTM model can be expressed as:

$$e_t = Y_t - \hat{L}_t \tag{11}$$

In the LSTM model, the residual sequence result $\{e_t\}$ is modeled, and the LSTM algorithm is used to describe the nonlinear relationship of the residual values, resulting in the predicted value $\hat{N}_t$ of the residual values:

$$\hat{N}_t = f(e_{t-1}, e_{t-2}, ..., e_{t-m}) + \varepsilon_t \tag{12}$$

In the formula, $f(\cdot)$ is the relationship function of the LSTM model, ε_t is the error value.

Finally, based on the predicted values of the ARIMA model and the residual predicted values output by the LSTM model, the predicted values of the combined model are calculated:

$$\hat{Y}_t = \hat{L}_t + \hat{N}_t \tag{13}$$

Structure of ARIMA-LSTM model is as shown in Fig 5.

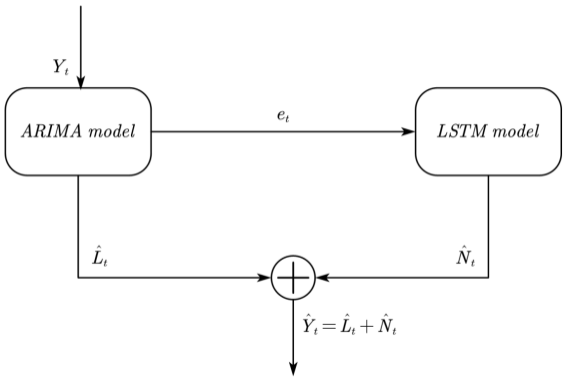


Fig 5 Structure of ARIMA-LSTM model

2.3. Simulation results and analysis

The paper will use approximately 80% of the influencing factor data as the training set and 20% of the data as the testing set. Then, the paper verified the R values of the training and testing sets as shown in Fig 6 and found that they were close to 1, which means The paper have a good fit.

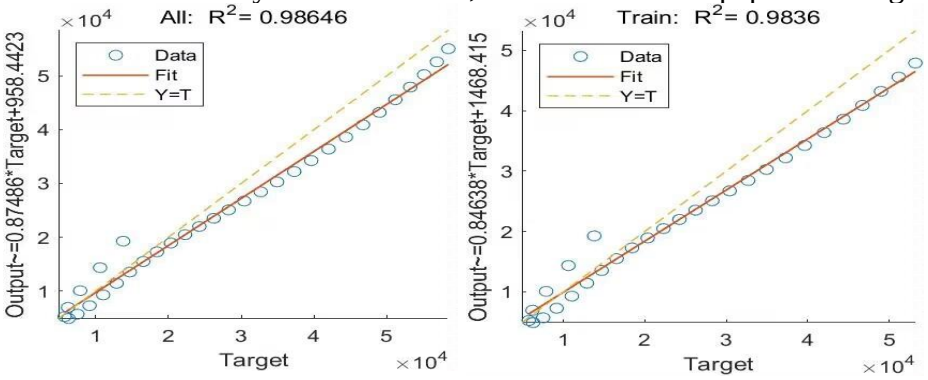


Fig 6 R test

Substitute the selected 5 influencing factor data into the ARIMA-LSTM model to obtain the and prediction result graph as shown in Fig 7.

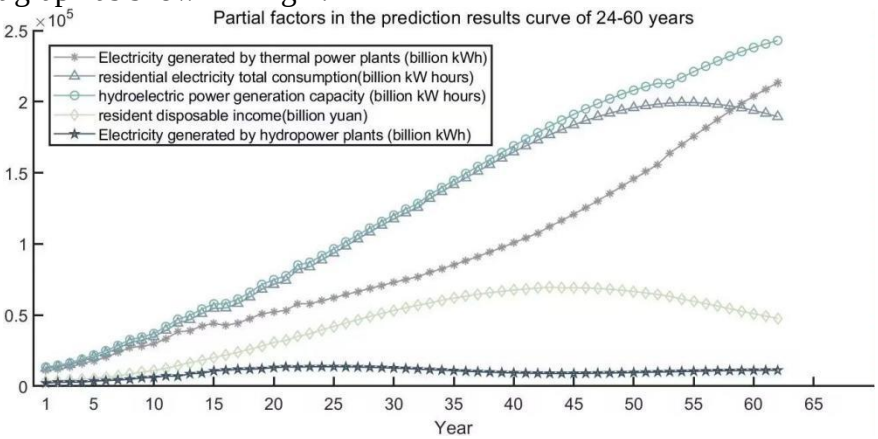


Fig 7 2024-2060 Power Supply Forecast Chart

Upon careful consideration of the provided analysis, several key insights can be articulated. Firstly, there is a consistent and upwards trajectory in the generation of diverse forms of electrical power, as well as in the disposable income levels of the populace. This corroborates findings fromprevailing scholarly research and is in harmony with the observed patterns of growth within the energy sector. Secondly, it is projected that China's matrix of electrical energy provision will, in the absence of extraordinary external influences, remain relatively stable through the forecasted period. This prediction is perhaps contingent upon the exclusion of potential shifts in policy, which are recognized as crucial variables that could significantly influence the energy landscape. These dynamics will be given further scholarly examination in subsequent analytical frameworks to account for the comprehensive understanding required in such complex policy and market environments.

3. Finding the most suitable areas for building PV power plants in China using the GPCA-EWM model

3.1. GPCA-EWM model

The analysis of data on the influencing factors of photovoltaic power stations is used to determine the feasibility of constructing photovoltaic power stations in a certain region, with the focus on selecting indicators and determining their weights[6]. Principal Component Analysis (PCA) is a multivariate statistical method that uses a few linear combinations of the original variables to reveal the vast majority of information about the original variables. The variance contribution rate is used as the basis for selecting principal components to simplify the system structure and grasp the essence of the problem.

Select indicators that affect the construction of photovoltaic power stations. Select the impact factor data for all provinces in China in 2022.

Calculate eigenvalues, contribution rates, cumulative contribution rates, and ultimately determine the principal components. Select the top m principal components with eigenvalues greater than 1, where m is the selected principal component score.

Since X is centralized, the h-th principal component is $F_h = \mu'_h X$; The variance contribution rate of principal components $a_k = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$; The cumulative variance contribution rate of principal components F_1, $F_2, \cdots, F_p$ is

$$a_1 + a_2 + \cdots + a_n = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{14}$$

Select the top m principal components (0<m ≤ p) with eigenvalues>1 or cumulative contribution rate>85%.

Calculate the numerical values of principal component factors for each province. Select n provinces and m indicators, then X_{ij} is the value of the jth indicator in the i-th province $(i = 1, 2, \cdots, n; j = 1, 2, \cdots, m)$.

Standardize the data of each principal component factor. The purpose is to eliminate the differences in attributes and dimensions between different indicator data. To avoid the meaningless logarithm when calculating entropy, the data is shifted.

$$X'_{ij} = \frac{X_{ij} - \min(X_{ij})}{\max(X_{ij}) - \min(X_{ij})} + 0.1 (i = 1, 2, \cdots, n; j = 1, 2, \cdots, m) \tag{15}$$

Calculate the proportion of the i-th factor under the j-th indicator to that indicator.

$$P_{ij} = \frac{X'_{ij}}{\sum_{i=1}^{n} X'_{ij}} (i = 1, 2, \cdots, n; j = 1, 2, \cdots, m) \tag{16}$$

Calculate the entropy value of the jth indicator.

$$e_j = -k \sum p_{ij} \ln(p_{ij}), k > 0, k = 1/\ln(n), e_j \geq 0 \tag{17}$$

Define the difference function.

$$d_j = \frac{1 - e_j}{m - \sum_{j=1}^{m} e_j} \left(0 \leq d_j \leq 1, \sum_{j=1}^{m} d_j = 1 \right) \tag{18}$$

Calculate the weight.

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j} (1 \leq j \leq m) \tag{19}$$

Calculate the comprehensive score for each sample.

$$s_i = \sum_{j=1}^{m} w_j p_{ij} (i = 1, 2, ..., n) \tag{20}$$

3.2. Selection of influencing factors

(1) Economic benefits[7]:

$$R = g \times p \tag{21}$$

Among them, R is power generation revenue, g is power generation, and p is grid electricity price. The formula for power generation is:

$$EP = HA \times \frac{P_{AZ}}{E_S} \times K \tag{22}$$

In the formula: EP :an online power generation (kW.h)

HA:one horizontal total solar irradiation ($kW \cdot h/m^2$, peak hours) $H = 3.6 \times 10^{-3} \sum_{i=1}^{n} g_i h$

Es:one standard irradiation capacity (con constant = $1 kW \cdot h/m^2$)

Paz:one-component installation capacity (kWp);

K:one comprehensive efficiency coefficient.

(2) Environmental benefits: The construction of photovoltaic power plants can reduce dependence on fossil fuels, improve local energy structure, and also consider the impact of pollutants such as noise generated during the operation of photovoltaic power plants on the environment.

(3) Cost[7]: The construction cost of photovoltaic power stations is one of the important factors affecting the feasibility of constructing photovoltaic power stations in a certain region. If the

construction cost is too high, it may make the area unable to bear the construction cost of photovoltaic power stations, thereby reducing the feasibility of constructing photovoltaic power stations in the area. The construction cost of photovoltaic power stations includes expenses related to equipment procurement, land leasing, construction and installation. Among them, the availability and cost of land resources are also important factors affecting the construction cost of photovoltaic power stations. In areas with scarce land resources, land costs may be high, which directly affects the construction cost and economic benefits of photovoltaic power stations.

LCOE in China from 2010 to 2021 is as shown in Fig 8.

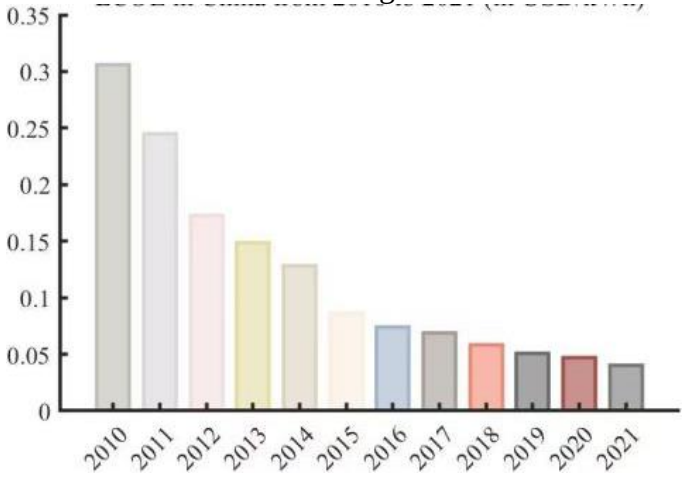


Fig 8 LCOE in China from 2010 to2021(in USD/kWh)

(4) Government policies and regulations: The government's policy support and subsidy intensity for photovoltaic power generation will directly affect the investment return and economic benefits of photovoltaic power stations, thereby affecting the feasibility of constructing photovoltaic power stations in the region.

The weight distribution of the main influencing factors is shown in Table 1.

Table 1 Principal Component Weight Allocation Table

Impact factors	weight
policy subsidies	0.6187
horizontal radiation exposure	0.0609
forest coverage	0.0730
GDP	0.1278
cost	0.0737
urbanization rate	0.0460

GPCA-EWM calculates the scores for each province. This article downloaded a map of China's administrative regions from Google Maps and used the depth of colors to represent the scores, as shown in Fig 9.

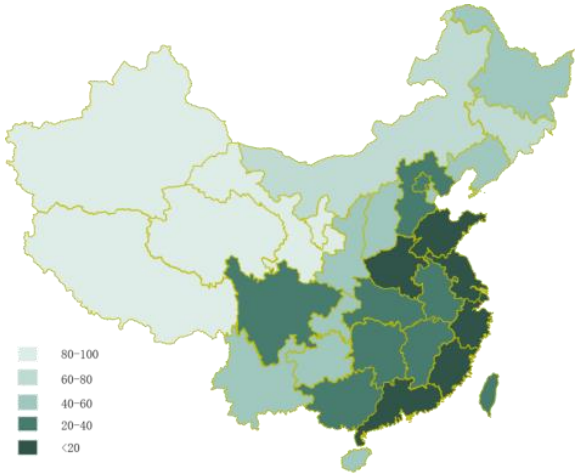


Fig 9 Score map

As can be seen from the Fig above, some provinces in western China generally have high scores, which may be due to their good lighting conditions and geographical conditions, sufficient available land, and high national policy subsidies[9][10]. For the eastern and southeastern coastal areas of China, although they may be economically developed, convenient transportation and fast power transmission, their generally high urbanization rate may lead to little available land. For the Sichuan Basin, the paper think it may be caused by the basin topography and climate conditions. The results obtained by the team are consistent with the results of existing studies, which proves that the accuracy of the model has been fully verified.

4. Conclusions

This paper examines the main influencing factors of China's electricity supply as electricity generated by thermal power plants, residential electricity total consumption, hydroelectric power generation capacity, resident disposable income, electricity generated by hydropower plants five kinds, and predicts that their development trend is steadily increasing in the next 60 years. Considering the various influencing factors affecting photovoltaic power plants selected China's most suitable region for the construction of photovoltaic power plants for the western region, which has good lighting conditions and geographic conditions, sufficient available land, and high subsidies of state policy. Understanding the influencing factors of China's power supply as well as the scientific and reasonable selection of the location of PV power plants can make better use of China's energy resources, improve economic efficiency, reduce construction costs, and reduce the difficulty of maintenance and other issues.

Reference

[1] Ma Yifei. Research and application of GIS-based solar photovoltaic energy plant siting method[D]. Northern Nationalities University, 2020.

[2] Duan Ying, Liao Liufeng, Yang Juan, Zhan Peigang. GIS-based estimation of suitable capacity for construction of photovoltaic projects in rocky desertification areas - An example from Zhenning, Ziyun and Guanling counties[J]. Middle and low latitude mountain meteorology, 2021, 45(06):26-30.

[3] Alami Merrouni A, Mezrhab A, Mezrhab A. CSP Sites Suitability Analysis in the Eastern Region of Morocco[J].Energy Procedia, 2016, 18:6-15.

[4] Munkhbat U, Choi Y. GIS-Based Site Suitability Analysis for Solar Power Systems in Mongolia[J]. Applied Sciences. 2021; 11(9):3748.

[5] Liu Yibo. Analysis of Electricity Quantity Prediction Based on LSTM Neural Network [J]. Integrated Circuit Application, 2023, 40 (08): 78-79.

[6] Zhu Yidong Research and Application of Photovoltaic Power Station Site Selection [D]. Anhui University of Science and Technology, 2022.

[7] Yu Tao. Analysis of Economic Indicators of Photovoltaic Power Generation Projects and Their Impact on Investment Risks [J]. Engineering Economics, 2019, 29 (05): 28-31.

[8] Ye Jianchun Research on Budget Management of Photovoltaic Power Station Construction Projects [D]. Southeast University, 2021.

[9] Qu Zhun, Yang Suchang, Xiao Jianhua. Analysis of the Current Situation and Potential of Photovoltaic Power Station Construction in Northwest China under the Dual Carbon Background [J]. Resources and Environment in Arid Areas, 2024, 38 (02): 20-26.

[10] Qu Ming Evaluation of Carbon Saving and Emission Reduction in the Qinghai Tibet Plateau Region Based on the Potential of Solar Power Generation [D]. Tianjin University of Urban Construction, 2023.