Proceedings of the 3rd International Conference on Business and Policy Studies
DOI: 10.54254/2754-1169/80/20241835

Analysis and Forecasting of Energy Data and GDP Based on Pearson Correlation Analysis-Random Forest Algorithm

Jiahong He[1,a,*]

[1]Business School, Shenzhen Technology University, Shenzhen, Guangdong, 518118, China
a. hejiahong024@gmail.com
**corresponding author*

Abstract: The relationship between energy data and GDP is an important economic issue because energy is the foundation that supports modern economic development. This paper examines the relationship between energy data and GDP and uses Pearson correlation analysis to calculate the relationship between each variable in energy data and annual GDP for 52 regions. The results show that total electricity consumption is the factor with the highest degree of influence on GDP, followed by natural gas consumption, biomass consumption, hydroelectricity consumption, LPG consumption and geothermal energy consumption. This indicates that there is an extremely strong correlation between local energy data and GDP. To further explore this correlation, this paper attempts to predict the GDP of the locality based on the local energy data using a machine learning approach.The Random Forest algorithm is chosen as the machine learning model, and the training, validation and test sets are divided according to the ratio of 6:2:2. By using MSE, RMSE, MAE, MAPE and R^2 to evaluate the model, it is found that the random forest machine learning algorithm can predict the local GDP well.The predicted value of GDP and the actual value of GDP are very close to each other, meanwhile, the value of MSE is 62.779.These findings show that the local energy data are closely related to economic development. The impact of energy factors on the economy should be taken into account when formulating economic policies. In addition, machine learning algorithms can provide strong support for economic forecasting and provide more accurate information for decision makers.

Keywords: Energy data, Random forest, Pearson correlation analysis

1. Introduction

The relationship between energy data and GDP is an important economic issue because energy is the foundation that supports modern economic development [1]. Energy data include information on production, consumption, trade and storage of various energy sources, while GDP is an important indicator of the level of economic development of a country [2].

Firstly, energy is an important support for modern economic development. Whether it is industrial production, transport, construction or daily life, a large amount of energy is needed to maintain operation. Therefore, the energy supply situation of a country or region directly affects its economic development level [3]. If a country lacks sufficient energy supply, then its economic activities will be limited, which will have a negative impact on its GDP [4].

Secondly, GDP also affects energy data. The more active the economic activity of a country or region, the greater its demand for energy [5]. For example, during a period of rapid economic growth, areas such as industry, transport and construction require large amounts of resources such as electricity, oil and gas to support their development [6]. Therefore, in this case, the relationship between energy data and GDP is positive [7].

However, the relationship between energy data and GDP is not a simple linear relationship [8]. In fact, there is a complex interaction between them [9]. For example, during a period of slowing economic growth, a country or region may take measures to reduce its demand for energy, thereby lowering its economic costs. These measures may include improving energy efficiency, promoting clean energy, etc. In this case, while GDP declines, energy data may remain stable or decline.

In addition, the relationship between energy data and GDP is affected by other factors, such as policies, technological advances, and environmental factors. For example, in the context of the growing global problem of climate change, more and more countries are taking measures to reduce carbon emissions and promote the use of renewable energy sources [10]. These policies may lead to a reduction in the consumption of certain traditional energy sources, but at the same time they may bring about new industries and employment opportunities. Therefore, in this scenario, while traditional energy data may decline, emerging industries and job opportunities are expected to drive GDP growth.

In summary, there is a complex interaction between energy data and GDP. The relationship between them is not only affected by economic factors, but also by policy, technology and environmental factors. Therefore, we need to consider various factors comprehensively in order to better understand the relationship between energy data and GDP and formulate development strategies accordingly. In this paper, we will discuss the relationship between energy data and GDP and explore the interaction between them, and finally try to use machine learning algorithms to forecast local GDP based on energy data.

2. Introduction to the dataset

The dataset used in this paper is a publicly available dataset from the UCL database, which includes a total of five years' worth of data, and the dataset consists of three types of data consisting of census and geographic data, energy data and economic data. In addition, the data used in this paper contains geographic data, energy data and economic data for 52 districts, which are available for comparative analysis. Energy data is shown in Table 1. Economic data is shown in Table 2.

Table 1: Energy data description.

Parameter name	Parameter Meaning
TotalC{year}	total energy consumption
TotalP{year}	Total energy production
TotalE{year}	Total energy expenditure
TotalPrice{year}	Average price of total energy
TotalC	Yearly percentage change in total energy consumption
TotalP	Yearly percentage change in total energy production
TotalE	Yearly percentage change in total energy consumption
TotalPrice	Yearly percentage change in total average energy prices
BiomassC{year}	Total biomass consumption
CoalC{year}	Total coal consumption
CoalP{year}	Total coal production
CoalE{year}	Total coal expenditure

Table 1: (continued).

CoalPrice{year}	Average coal price
ElecC{year}	Total electricity consumption
ElecE{year}	Total expenditure on electricity
ElecPrice{year}	Average price of electricity
FossFuelC{year}	Total fossil fuel consumption
GeoC{year}	Total geothermal energy consumption
GeoP{year}	Net geothermal energy generation in the power sector
HydroC{year}	Total utilities consumption
HydroP{year}	Total net hydropower generation
NatGasC{year}	Total natural gas consumption
NatGasE{year}	Total expenditure on natural gas
NatGasPrice{year}	Average price of natural gas
LPGC{year}	Total LPG consumption
LPGE{year}	Total expenditure on liquefied petroleum gas
LPGPrice{year}	Average price of liquefied petroleum gas

Table 2: GDP data description.

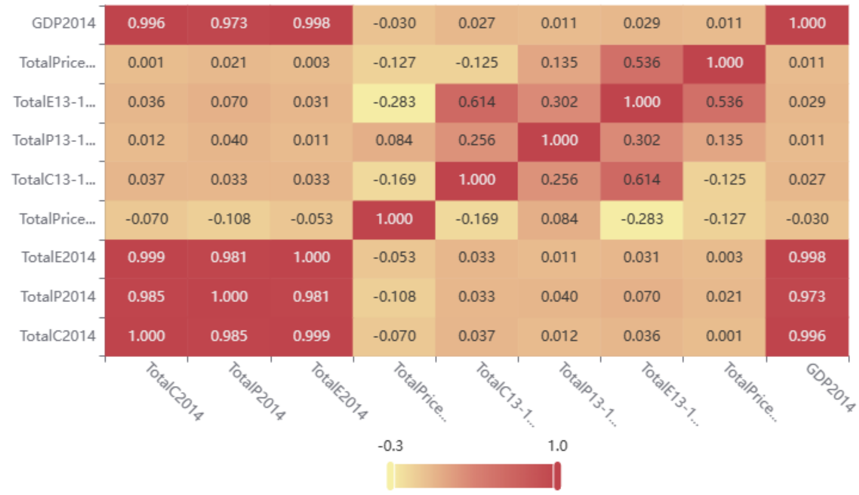
Parameter name	Parameter Meaning
GDP{year}{quarter}	Quarterly GDP
GDP{year}	Average GDP for the year

3. Relevance analysis

Pearson correlation analysis is a statistical method used to measure the strength of a linear relationship between two variables. Its principle is based on the Pearson correlation coefficient, which can be used to measure the degree of linear correlation between two variables. The Pearson correlation coefficient is obtained by calculating the product of the covariance of two variables and their respective standard deviations.

The principle of Pearson's correlation analysis is based on the assumption that there is a linear relationship between two variables and the degree of correlation between them is determined by calculating their covariance and standard deviation. If two variables are positively correlated, their covariance is positive; if two variables are negatively correlated, their covariance is negative. Also, by dividing the covariance by the product of their respective standard deviations, the effect of different units on the results can be eliminated, resulting in a unitless Pearson correlation coefficient with values ranging from -1 to 1. In this way, the strength and direction of the linear relationship between two variables can be judged by comparing the size of the Pearson correlation coefficient.

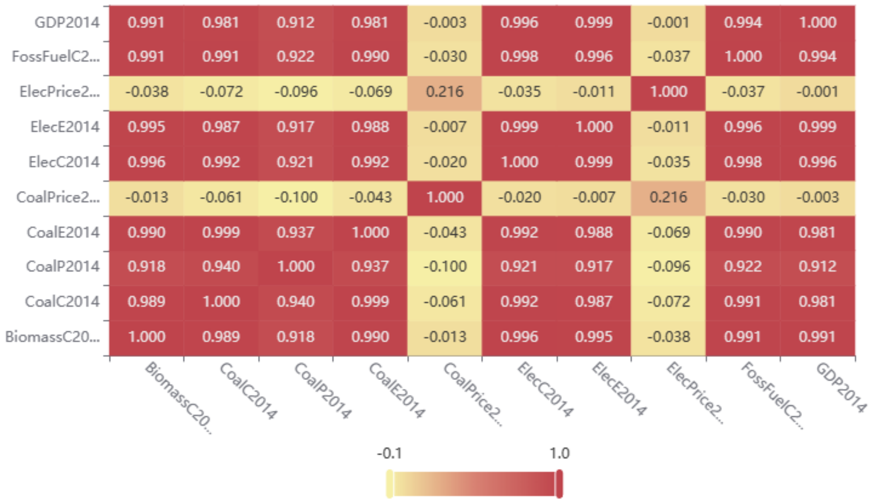
In this paper, Pearson correlation analysis is used to calculate the relationship between each variable in the energy data and annual GDP in 52 regions respectively, to calculate the correlation coefficient between the variables, and to draw heat maps. The energy data is divided into three parts, the first part is, the heat map of total consumption and GDP is shown in Fig. 1; the second part is the coal and electricity data, the correlation coefficient between coal and electricity data and GDP is shown in Fig. 2; the third part is the hydroelectricity, natural gas and liquefied petroleum gas (LPG) data, and the relationship between hydroelectricity, natural gas and LPG data and GDP is shown in Fig. 3.



	TotalC2014	TotalP2014	TotalE2014	TotalPrice...	TotalC13-1...	TotalP13-1...	TotalE13-1...	TotalPrice...	GDP2014
GDP2014	0.996	0.973	0.998	-0.030	0.027	0.011	0.029	0.011	1.000
TotalPrice...	0.001	0.021	0.003	-0.127	-0.125	0.135	0.536	1.000	0.011
TotalE13-1...	0.036	0.070	0.031	-0.283	0.614	0.302	1.000	0.536	0.029
TotalP13-1...	0.012	0.040	0.011	0.084	0.256	1.000	0.302	0.135	0.011
TotalC13-1...	0.037	0.033	0.033	-0.169	1.000	0.256	0.614	-0.125	0.027
TotalPrice...	-0.070	-0.108	-0.053	1.000	-0.169	0.084	-0.283	-0.127	-0.030
TotalE2014	0.999	0.981	1.000	-0.053	0.033	0.011	0.031	0.003	0.998
TotalP2014	0.985	1.000	0.981	-0.108	0.033	0.040	0.070	0.021	0.973
TotalC2014	1.000	0.985	0.999	-0.070	0.037	0.012	0.036	0.001	0.996

Figure 1: Correlation heat map.
(Photo credit: Original)

From the correlation heat map, it can be seen that the correlation coefficient between the total energy consumption and the local GDP of the 52 regions in the energy data reaches 99.6%, and the correlation coefficients between the total energy production and the total energy expenditure in the energy data and the local GDP reach 97.3% and 99.8, respectively.From the results of the correlation coefficients, it can be reflected that there are strong positive correlations between a region's energy consumption and expenditure and the local The results of the correlation coefficients can reflect that the energy consumption and expenditure of a region has a strong positive correlation with the local GDP.



	BiomassC20...	CoalC2014	CoalP2014	CoalE2014	CoalPrice2...	ElecC2014	ElecE2014	ElecPrice2...	FossFuelC2...	GDP2014
GDP2014	0.991	0.981	0.912	0.981	-0.003	0.996	0.999	-0.001	0.994	1.000
FossFuelC2...	0.991	0.991	0.922	0.990	-0.030	0.998	0.996	-0.037	1.000	0.994
ElecPrice2...	-0.038	-0.072	-0.096	-0.069	0.216	-0.035	-0.011	1.000	-0.037	-0.001
ElecE2014	0.995	0.987	0.917	0.988	-0.007	0.999	1.000	-0.011	0.996	0.999
ElecC2014	0.996	0.992	0.921	0.992	-0.020	1.000	0.999	-0.035	0.998	0.996
CoalPrice2...	-0.013	-0.061	-0.100	-0.043	1.000	-0.020	-0.007	0.216	-0.030	-0.003
CoalE2014	0.990	0.999	0.937	1.000	-0.043	0.992	0.988	-0.069	0.990	0.981
CoalP2014	0.918	0.940	1.000	0.937	-0.100	0.921	0.917	-0.096	0.922	0.912
CoalC2014	0.989	1.000	0.940	0.999	-0.061	0.992	0.987	-0.072	0.991	0.981
BiomassC20...	1.000	0.989	0.918	0.990	-0.013	0.996	0.995	-0.038	0.991	0.991

Figure 2: Correlation heat map.
(Photo credit: Original)

From the correlation heat map between coal and electricity data and GDP, it can be seen that there is a very strong positive correlation between local GDP and total biomass consumption, total coal consumption, total coal production, total coal expenditures, total electricity consumption, total electricity expenditures, and total fossil fuels consumption in 52 regions, especially the correlation coefficient of GDP and total electricity consumption reaches 99.9 per cent, followed by total biomass consumption with a correlation coefficient of 99.1 per cent.

In conclusion, electricity consumption and biomass consumption in energy data are the most critical factors affecting local GDP.

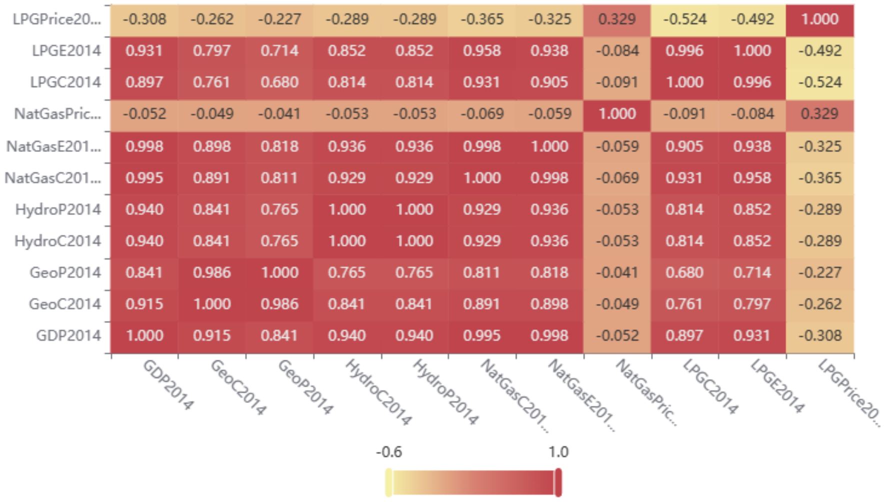


Figure 3: Correlation heat map.
(Photo credit: Original)

As shown by the heat map of the correlation between hydropower, natural gas and LPG data and GDP, natural gas consumption is also a key factor influencing local GDP, followed by hydropower consumption LPG consumption and geothermal energy consumption.

In summary, ranked according to the degree of influence on local GDP, the total consumption of electricity among energy factors is the most influential on GDP, followed by natural gas consumption, biomass consumption, hydroelectricity consumption, LPG consumption and geothermal energy consumption. The specific correlation ranking is shown in Table 3.

Table 3: Correlation coefficient between energy consumption and GDP.

Energy consumption	Correlation coefficient
Total electricity consumption	99.9%
Natural gas consumption	99.8%
Biomass consumption	99.1%
Utilities consumption	94.0%
Consumption of liquefied petroleum gas	93.1%
Geothermal energy consumption	91.5%

4. Machine Learning Prediction

Random Forest is an integrated learning method based on decision trees, which performs tasks such as classification, regression and feature selection by constructing multiple decision trees. Compared to a single decision tree, Random Forest has higher accuracy and robustness.

The core idea of random forest is to randomly divide the dataset into multiple subsets and then construct a decision tree for each subset. At each node, only part of the features are considered for division, and a portion of the samples from the current node are randomly selected to train the subtree. This avoids overfitting and increases the stability and generalisation of the model. When predicting, for classification problems, Random Forest votes the results of each decision tree and takes the category with the most votes as the final prediction; for regression problems, the average of all decision tree outputs is taken as the final prediction. Meanwhile, since each subtree is constructed independently, it can be computed in parallel, which improves the efficiency of the model. In addition to classification and regression, Random Forest can also be used for feature selection. When constructing a decision tree, the number of times each feature has been used can be recorded, and the features can be ranked according to their importance, so that the most discriminating features can be selected.

Based on the conclusion obtained earlier in this paper that there is a strong correlation between the GDP of a region and the local energy data, this paper tries to use machine learning methods to predict the GDP of the region based on the local energy data.Random forest algorithm is chosen as the machine learning model, and the training set, validation set and test set are divided according to the ratio of 6:2:2, and MSE, RMSE, MAE, MAPE and R² to evaluate the prediction effect of the model, and the results are shown in Table 4 and Fig. 4.The prediction effect of the model is tested using the test set, and the line graph of the predicted value GDP versus the actual value GDP is plotted, as shown in Fig. 5.

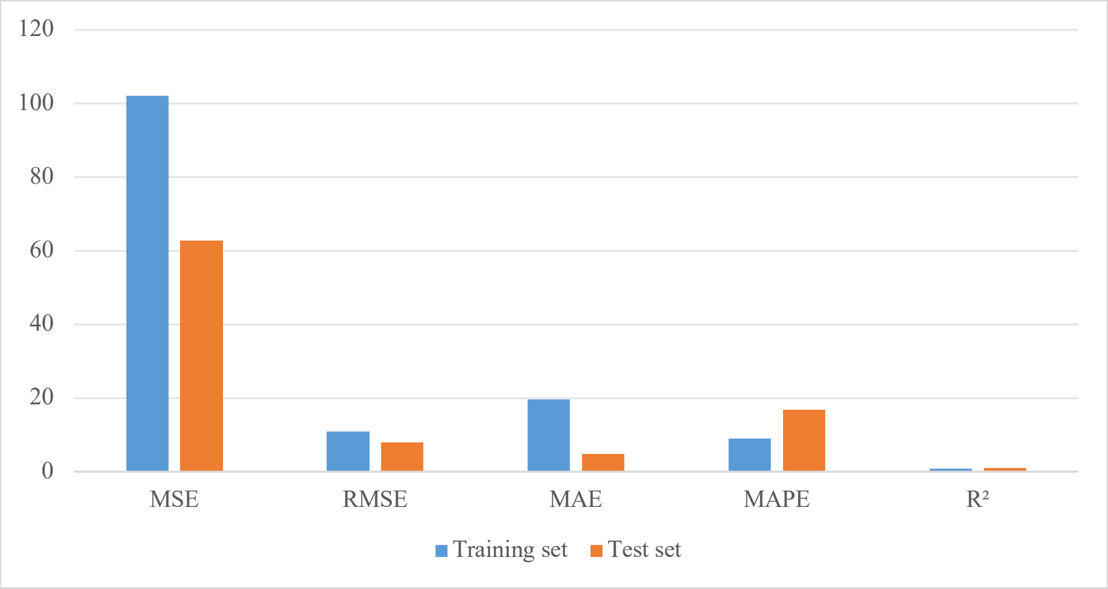


Figure 4: Modelling evaluation.
(Photo credit: Original)

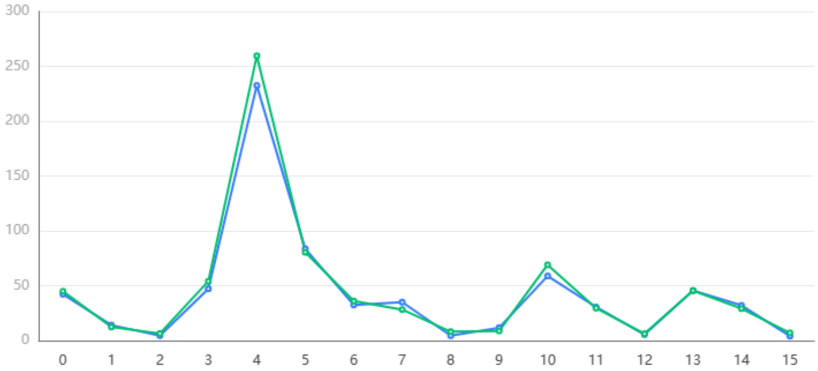


Figure 5: Forecasted versus actual GDP.
(Photo credit: Original)

As can be seen from Fig. 5, the prediction of GDP using energy data through the Random Forest machine learning algorithm achieves good prediction results, the predicted value of GDP and the actual value of GDP are very close to each other, and at the same time, the MSE value is 62.779, and the model based on the energy data can predict the local GDP very well.

5. Conclusion

By analysing the relationship between energy data and GDP in 52 regions, this paper finds that total electricity consumption has the greatest degree of influence on GDP among energy factors, followed by natural gas consumption, biomass consumption, hydroelectricity consumption, liquefied petroleum gas (LPG) consumption and geothermal energy consumption. This indicates that there is an extremely strong correlation between a region's GDP and local energy data.

In order to predict the local GDP more accurately, this paper uses the Random Forest algorithm to make predictions based on local energy data and evaluates the predictive effectiveness of the model. The results show that the prediction of GDP using energy data by the Random Forest machine learning algorithm achieves a very good prediction effect, and the predicted value of GDP and the actual value of GDP are very close to each other. Meanwhile, the MSE value is 62.779, which indicates that the model can predict the local GDP well based on energy data.

Combining the above results, it can be concluded that among the 52 regions, the total consumption of electricity has the greatest degree of influence on the local GDP, therefore, the investment and development of the power industry should be emphasised in the formulation of economic policies. In addition, the model constructed based on the Random Forest algorithm can make good use of local energy data to predict the GDP of the region, which provides an effective decision support tool for the government and enterprises.

References

[1] *Libya Energy Economic Summit (LEES) 2024: Global Energy Companies Committed to Sustainable Energy in Libya[J].M2 Presswire,2024,*

[2] *Kecai F ,Mao Z ,Yanan S , et al.Nexus between economic recovery, energy consumption, CO2 emission, and total natural resources rent[J].Resources Policy,2023,87(PB):*

[3] *African Energy Chamber (AEC) Endorses Libya Energy Economic Summit 2024[J].M2 Presswire,2024,*

[4] *Ava L ,Seungmoon S ,B B V , et al.Optimizing exoskeleton assistance to improve walking speed and energy economy for older adults.[J].Journal of neuroengineering and rehabilitation,2024,21(1):1-1.*

[5] *Sean K ,Chunxu W ,Megan M , et al.A comprehensive analysis of the energy, economic, and environmental impacts of industrial variable frequency drives[J].Journal of Cleaner Production,2024,434140474-.*

[6] *Lei L ,Wenjie L ,Jian Y , et al.Life cycle energy, economic, and environmental analysis for the direct-expansion photovoltaic-thermal heat pump system in China[J].Journal of Cleaner Production,2024,434139730-.*

[7] *Kecai F ,Mao Z ,Yanan S , et al.Nexus between economic recovery, energy consumption, CO2 emission, and total natural resources rent[J].Resources Policy,2023,87(PB):*

[8] *[8] Price J ,Shah N .Ten Resolutions for the Energy Industry in 2024[J].Climate and Energy,2023,40(6):17-20.*

[9] *Brun K, Kurz R. Reciprocating vs Centrifugal Compressors for Carbon Capture and Sequestration[J]. Turbomachinery International,2024,64(6):*

[10] *Barlybayev A ,Zhetkenbay L ,Karimov D , et al.Development neuro-fuzzy model to predict the stocks of companies in the electric vehicle industry[J].Eastern-European Journal of Enterprise Technologies,2023,4(4):72-87.*