Project 1: The consumption of non-renewable and renewable energies in USA

Contents

1. Introduction

Nowadays, we use a lot of energy for our daily activities at home, in businesses and industry as well as for transportation. To satisfy the demand of energy, different energy sources have been used in the United States. The energy can be classified into renewable (an energy source that can be replenished) and non-renewable (an energy source that cannot be easily replenished) types.

In 2017, it was found that domestic energy production is around 90% of U.S. energy consumption. Currently, energy in the U.S. was consumed in five main purposes (commercial, residential, industrial, transportation and electric power). Depending on economic growth and other factors (such as weather and fuel prices), energy consumption for these sectors has been changed over time.

Due to the environmental issues, renewable energy has been considered to deploy more and more in practical uses. Many studies proved that renewable energy plays an important role in reducing greenhouse gas emission. Using renewable energy can reduce the use of fossil fuels, which are major imported sources of U.S.

a. Goals

The aim of this project is to predict the USA energy consumption of non-renewable and renewable energies in the next 5 years.

b. Data source

All data for my report were imported from website of U.S. Energy Information Administration (EIA) (https://www.eia.gov/). Total energy consumption of non-renewable and renewable energies for each state in USA from 1960-2016 will be used for further analysis.

c. Approach

- Collect data from EIA website via API query, csv and excel files.
- Analyze and extract required data for this report.
- Explore and visualize data of energy consumption.
- Construct a model for prediction.

2. Description of the dataset

Data were found and downloaded from EIA website (belong to U.S. Energy Information Administration). The format of data files could be in csv or excel. The API query was also used to collect data (using Requests library).

Pandas package was used to load all data files (used functions: pd.read_xls, pd.read_csv). This problem is related to time series. So, the index of dataframe was set to be datetime index when loading date files.

I only kept the necessary information in data file. Function pd.drop was used to remove unwanted data.

For missing values, it depends on the purpose of using. I could handle them as follows:

- Remove missing value data by pd.dropna()
- Replace them with appropriate values by pd.fillna()
- Keep it as NaN

After cleaning all data, I saved my data into a csv file (pd.to_csv function) for further analysis.

3. Initial findings from exploratory analysis

a. Trend of energy consumption

In general, the total energy consumption of USA has kept increasing rapidly since 1950. It reached the highest level in 2007. After that, the annual increasing trend in energy consumption changed sharply because of the economic recession. It seems to be that the energy consumption becomes stable after 2009. Total annual energy consumption increased in five of the years between 2009 and 2017 and decreased in five of the years. Economic growth and other factors (such as weather and fuel prices) can influence energy consumption.

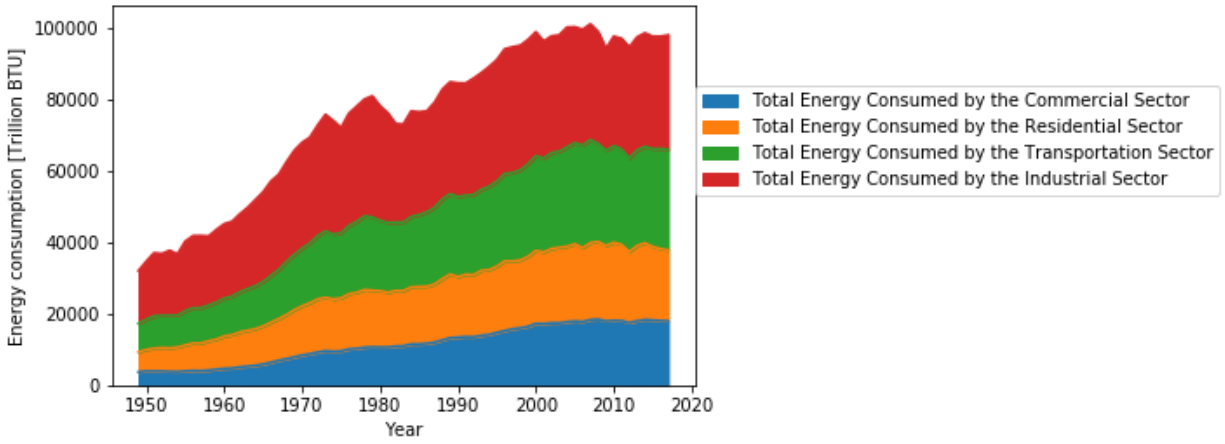


Fig 1: Total energy consumption in USA from 1949 to 2017

From decomposed plot of monthly total primary energy consumption, we can see the trend, seasonal and noise behavior for the monthly energy consumption in a single year.

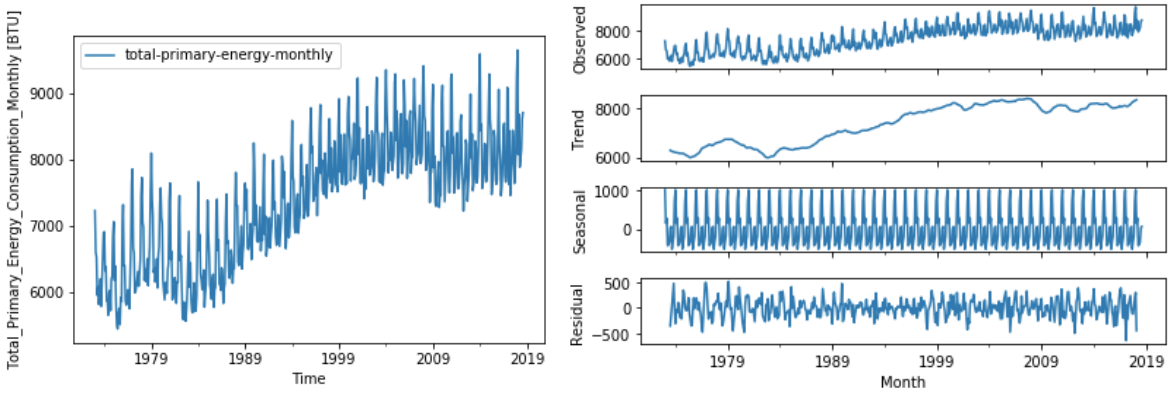
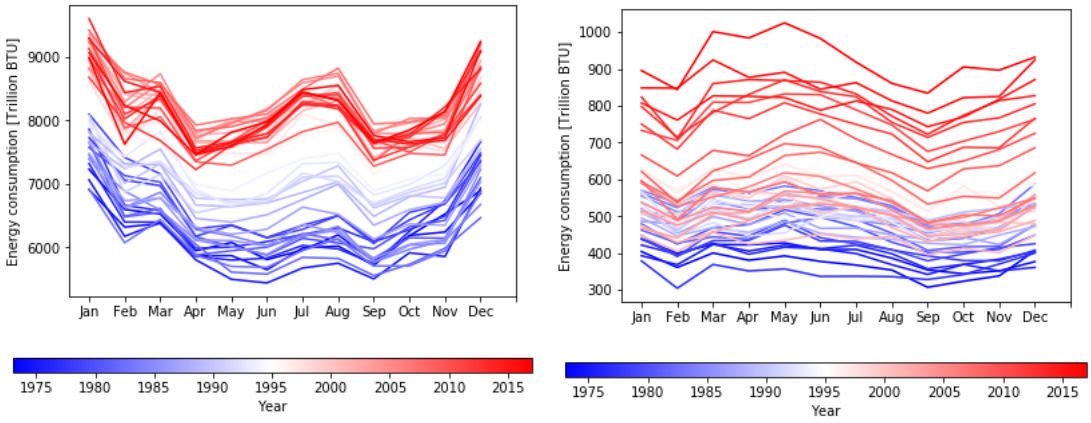


Fig 2: Decomposition of monthly energy consumption in USA from 1973 to 2017

Looking at the monthly energy consumption, it is clearly indicated the stable trending of energy consumption in USA have occurred since 2010. In one year, we can draw a rule for monthly energy consumption as follows: it began decreasing from January to May (first local minimum is in April or May), then increasing until August (local maximum). After that, it decreased again in September (2^{nd} local minimum). From October to December, energy consumption returned to increase. The maximum energy consumption in one year could be in January or December (around 80% in January since 1973). For monthly renewable energy consumption, it has a maximum value around May or June and two local minimums around February and September.



(a) (b)

Fig 3: Distribution of monthly total energy (b) and total renewable energy (b) consumption from 1973 to 2017

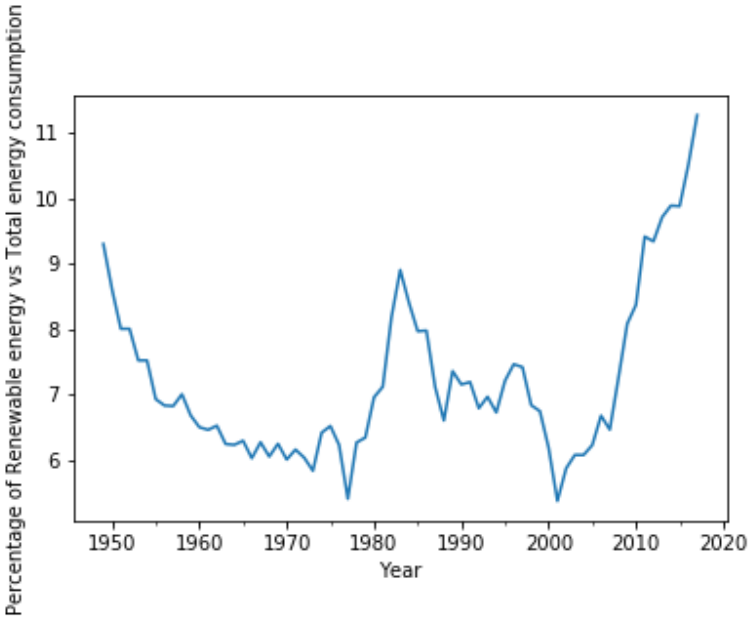


Fig 4: Percentage of renewable energy consumption from 1949 to 2017

The percentage of renewable energy in total energy consumption was around 11% in 2017. Before the year of 2000s, the percentage of renewable energy is small (from 6 to 9%). After 2000, renewable energy has been received more and more attention due to the sustainable strategy for energy development. It is displayed the percentage of renewable energy kept increasing and expect to increase in the next coming years.

b. The distribution of energy consumption

The energy consumption is not uniformly distributed across the United States and has a skewed-right distribution. In 2016 (our latest available data), it showed that top five highest energy consumption states are Texas, California, Florida, Louisiana and Illinois. Top five lowest energy consumption state are Vermont, Columbia (DC), Island, Delaware and Hawaii.

In general, it can classify all of states into three groups based on energy consumption. They are low (smaller than 2 million BTU), medium (from 2 to 6 million BTU) and high (larger than 6 million BTU) group. Even though there are other factors that affects the energy consumption for each state (such as area, population, geography…), this classification can provide a quick overview of total energy consumption in the whole country.

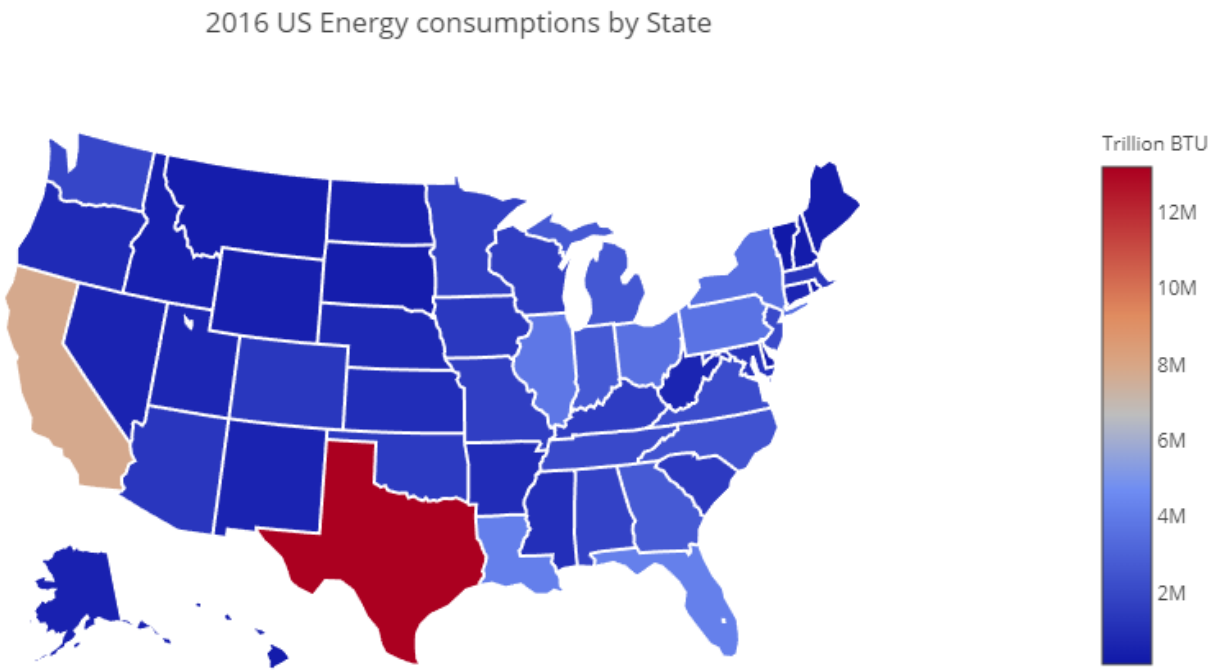


Fig 5: Total energy consumption by State in 2016

c. Engineering features

For total energy consumption, we can point out that 17 engineering features can be considered in our report. They included GDP, population, prices of materials (crude oil, natural gas, coal, electricity), motor vehicles mileage & fuel economy, producer price index for all commodities, energy import & export, CO_2 emission, heating & cooling degree days, renewable production & consumption, total energy production. Pearson coefficient indicates that most of these features

have a strong relation with total energy consumption. It could be neglected the feature of coal price due to the weak correlation.

For total renewable consumption, all 17 features can be considered as total renewable energy consumption, total renewable energy production, hydroelectric power, hydroelectric power consumption per production, geothermal energy consumption per production, geothermal energy consumption by the electric power, solar per PV energy consumption per production, total biomass energy consumed by the electric power, total biomass energy production, total renewable energy consumed by the electric power, waste consumption for electricity, wind energy consumed by the electric power, wind energy consumption per production, wood consumption for electricity, biomass exports, biomass imports, biofuels consumption and biofuels production. Similarly, the values of the Pearson correlation coefficient for renewable energy consumption were also determined. It indicated that all features have positive Pearson coefficient with total renewable energy consumption in the range of 0.6 and 1.

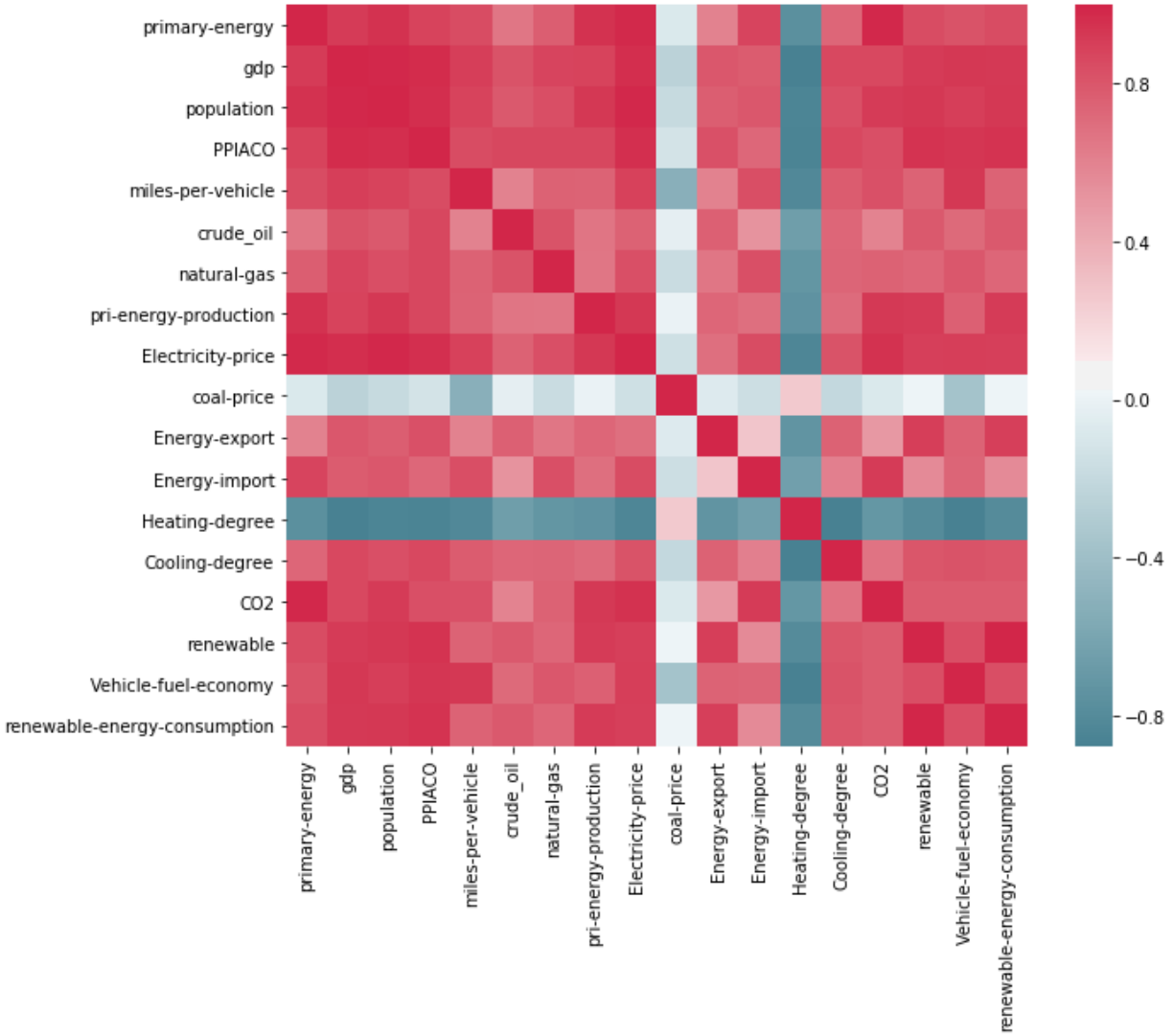


Fig 6: Pearson coefficient of all features for total energy consumption

For time series problem, it is easy to make the prediction on the stationary series. The stationary process means that it doesn't change its statistical properties over time. For example, mean and variance do not change over time (constancy of variance is also called homoscedasticity). The covariance function also does not depend on the time (should only depend on the distance between observations). We used Augmented Dickey–Fuller test (ADF) for stationary test in our study. The null hypothesis of the Augmented Dickey-Fuller is that there is a unit root, with the alternative that there is no unit root in 5% of significance level.

a. Seasonal ARIMA model

Univariate time series model was tried to predict the monthly total energy and renewable energy consumption. The seasonal ARIMA model was used for this case because of the decomposition plot. Values of AIC for training set was used to determine the optimum parameters of the model. Prediction values from this model was not fitted well with test set. I think that the multivariate time series could be a better choice. We need to evaluate and choose the important features from 17 features.

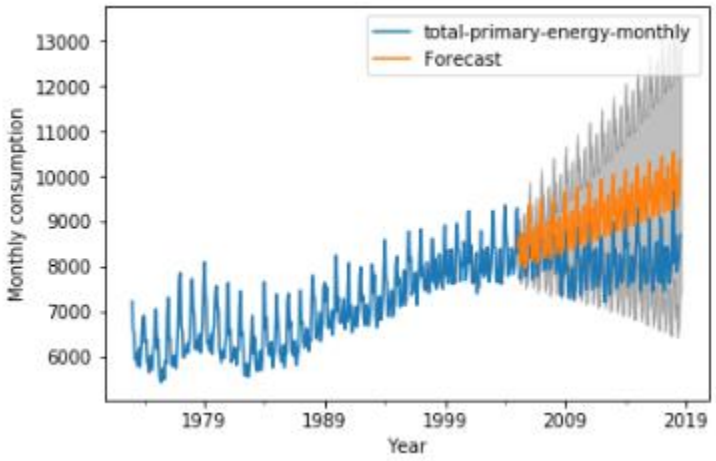


Fig 7: Seasonal ARIMA model for monthly energy consumption

4. Results and in-depth analysis

a. Feature importance

It is necessary to use multivariate time series model for predicting the total energy and renewable energy consumption. There are 17 engineering features for each problem. If I use all these features, the model will be completely failed for prediction. The idea is that I only choose the important features which are strongly influenced in the target variables for carrying out prediction. The random forest regressor was used to determine the important features. Note that all data was scaled and transformed before applying random forest classifier.

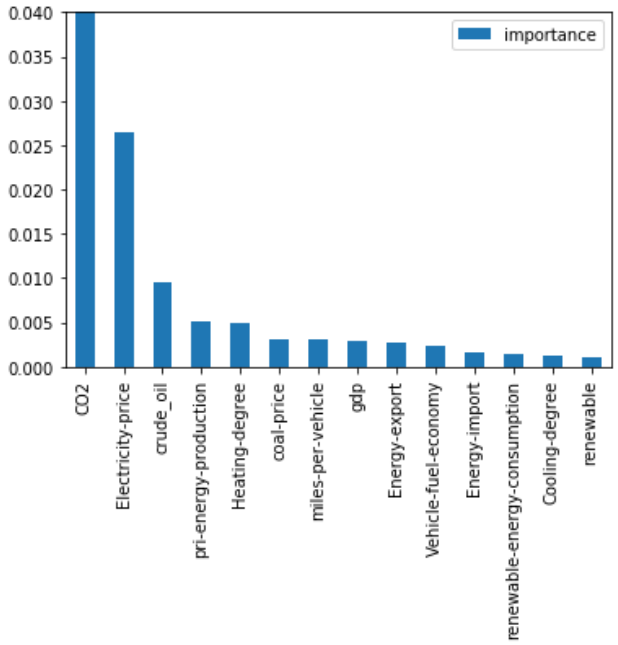


Fig 8: Ranking importance scores for all features of total energy consumption

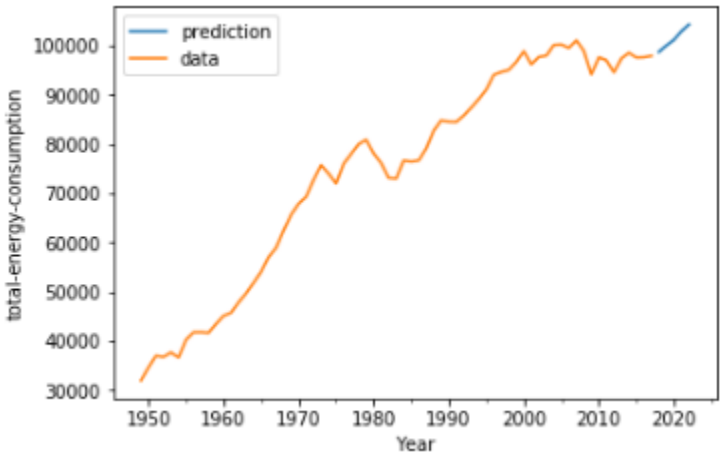
For total energy consumption, it found out that CO_2, electricity price and crude oil price are the most important features. Ratio of train and test set is 7/3. Accuracy of training and test set are 0.97 and 0.83, respectively. Similarly, the important features of total renewable energy consumption are total renewable production, hydroelectric power consumption per production and wood consumption for electricity generation.

	importance
total-renewable-production	0.817505
Hydroelectric-power-Con-Pro	0.059751
Wood_Consumption_for_Electricity	0.040350
Hydroelectric-power	0.021367
Total_Biomass_Energy_Consumed_by_the_Electric_Power	0.016894
Biofuels_Consumption	0.014514

Table 1: Ranking importance scores for all features of total renewable energy consumption

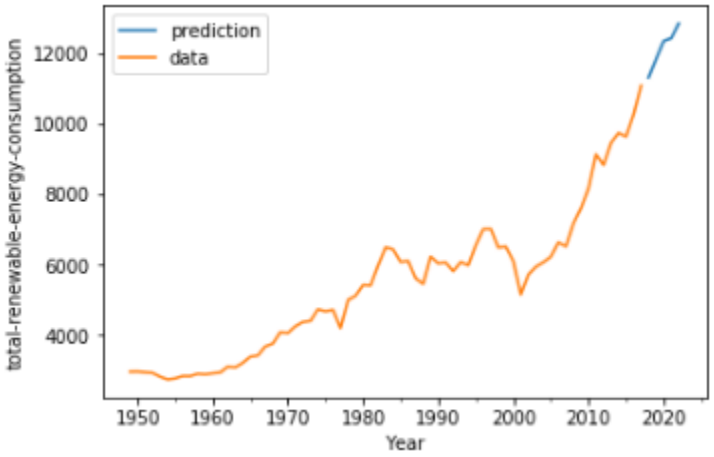
b. Prediction using VAR model

Vector Autoregressions (VAR) model was used to predict the total energy and total renewable energy consumption. Only important features were considered in the prediction. Results indicated that both total energy and total renewable energy consumption increase in next 5 years.



	Predicted-values
2018-01-01	98684.037609
2019-01-01	99987.343257
2020-01-01	101125.865984
2021-01-01	102879.853733
2022-01-01	104247.378982

Fig 9: Prediction for total energy consumption

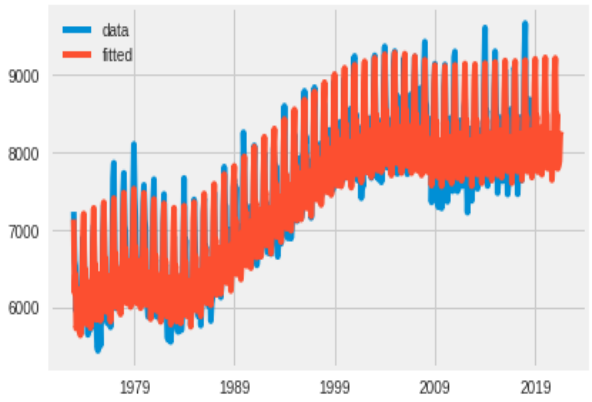


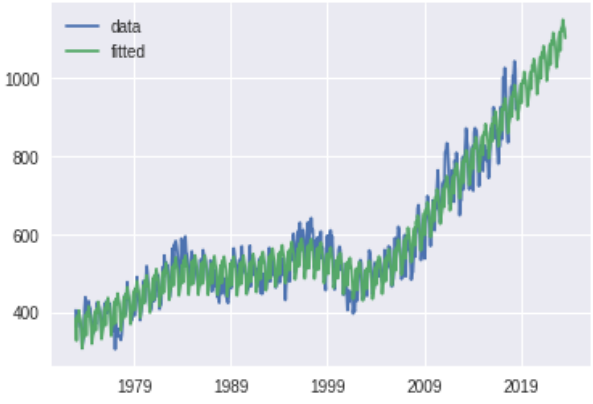
	Predicted-values
2018-01-01	11306.844667
2019-01-01	11833.068270
2020-01-01	12337.983842
2021-01-01	12416.785156
2022-01-01	12836.197791

Fig 10: Prediction for total renewable energy consumption

c. Using Prophet

For monthly energy consumption, it has a yearly seasonal trending. Besides using seasonal ARIMA model, the fbprophet model was also considered. It works well with time series that have strong seasonal effects and several seasons of historical data.





(a) (b)

Fig 11: Prediction for monthly total energy consumption (a) and montly total renewable energy consumption (b) by fbprophet

Figure 11 showed the true and predicted values from fbprophet model for monthly energy consumption. There is a good agreement between true and fitted data for both cases. The root mean squared errors are 217.8 and 33.2 (BTU), respectively. The predictions in next 5 years present that monthly total energy consumption seems to be into stable state. Even though it keeps increasing, the percentage of increasing is small. In contrast, monthly renewable energy consumption prefers to increase strongly in next coming years. After 2000, renewable energy consumption has been used more and more in USA. It indicates the tendency of using energy in the future is to pay attention in cleaner and friendlier environmental energy resources.

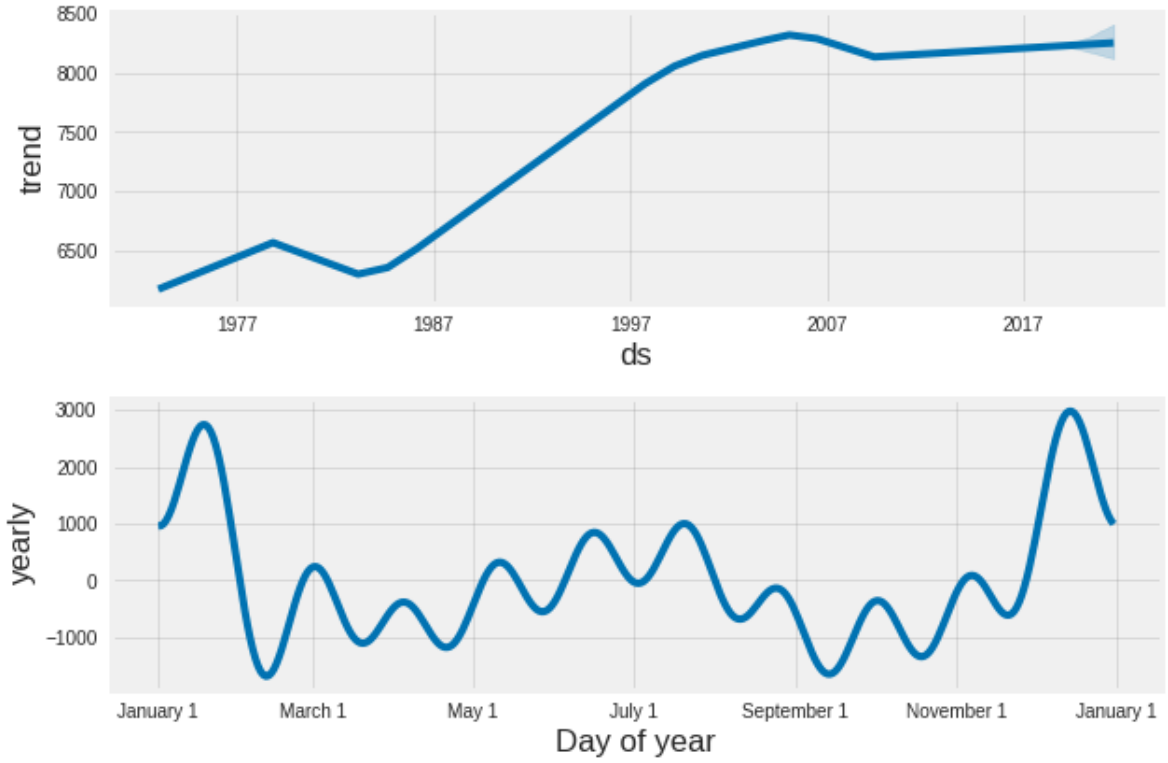


Fig 12: Prediction of trend for monthly total energy consumption by fbprophet

Fig 13: Prediction of trend for monthly total renewable energy consumption by fbprophet

The distribution of energy consumption in a single year was also included in figure 12 and 13. For total energy consumption, the distribution is similar to what we explored in Figure 3a. It often reaches maximum monthly consumption in January and December. However, it does not have an apparent trend for total renewable energy consumption. The maximum values can be in the month of January, March, May or October.

5. Conclusions

In this report, multivariate and univariate time series models were used to predict the total energy and renewable energy consumption of USA in next 5 years. It was found that both total energy and renewable energy consumption still keep increasing. However, the rate of increasing for renewable energy consumption is higher than total energy consumption. In respect of monthly energy consumption, the total energy has currently been in the stable condition and expected to be in the steady state in next 5 years. For renewable energy, the monthly consumption has gone into the growing steps and predicted to be enlarged in the future.

In addition, the random forest regressor was also deployed to determine the important features for multivariate time series prediction. It exhibited that CO_2 emission, electricity price, crude oil price are the important features for total energy consumption. And, total renewable production, hydroelectric power consumption per production, wood consumption for electricity generation are the most influenced features for total renewable energy consumption. Moreover, the trend of energy consumption throughout a year was also determined by using fbprophet model. The forecast values indicated that fbprophet is the good model for predict seasonal time series problem.