US Energy Production Forecast Using Time Series Models

By

Zijun Wu

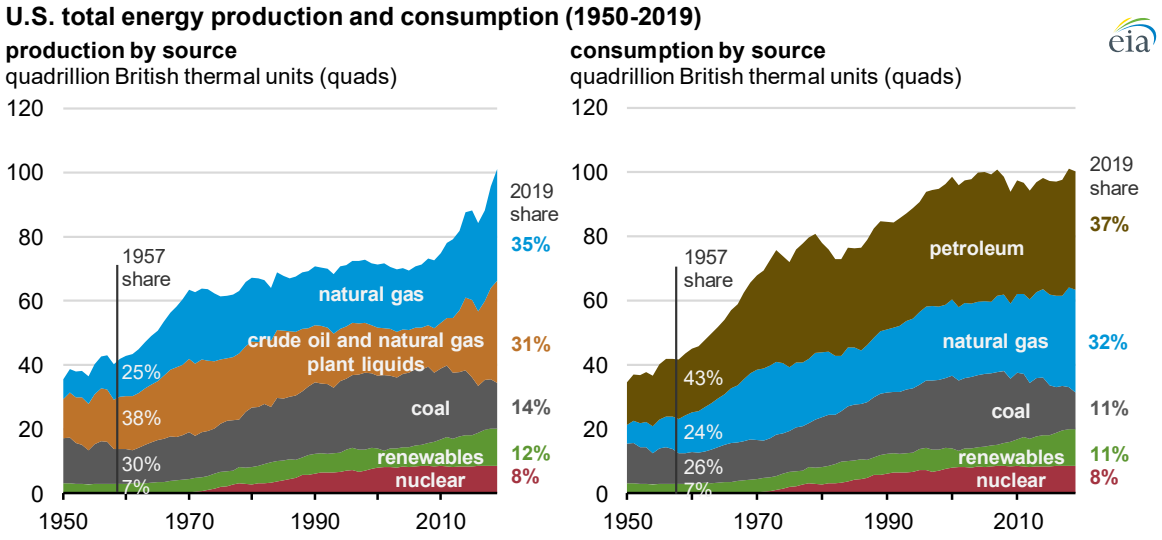
Writing sample

October 7, 2022

Problem Statement

According to the graph from U.S. Energy Information Administration, US energy production and consumption continuously increased from 1950s to 2019, with a peak in 2019. However, the energy production growth fluctuates from 1990 to 2019 while the energy consumption growth seems to gradually decelerate between the same period. For the next 10 years, regarding the energy consumption growth slowdown, I am going to forecast if the energy production will be affected and decelerate growth as well.

Figure 1. *US Energy Production and Consumption (1950-2019)*



Source: U.S. Energy Information Administration

Analytical Goal

If US energy production will smooth out or decrease in the next 5 years, United States will expect to reduce energy production emissions since 88% of the energy are non-renewable. If US energy production grows in the next 5 year, we will expect energy production emissions to grow in the future.

Data Sources

General description:

US_Energy_Production data is retrieved from FRED, Federal Reserve Bank of St. Louis. The data frame has the following properties:

- Two columns: Energy_Production and DATE
- Datetime frequency: Monthly
- Time window: 1985 – 2019
- Unit: Index 2017=100 (base year 2017)

Figure 2. *Energy Production Data Frame Information*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 392 entries, 0 to 391
Data columns (total 2 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   date               392 non-null    object
 1   Energy_Production  392 non-null    float64
dtypes: float64(1), object(1)
memory usage: 6.2+ KB
```

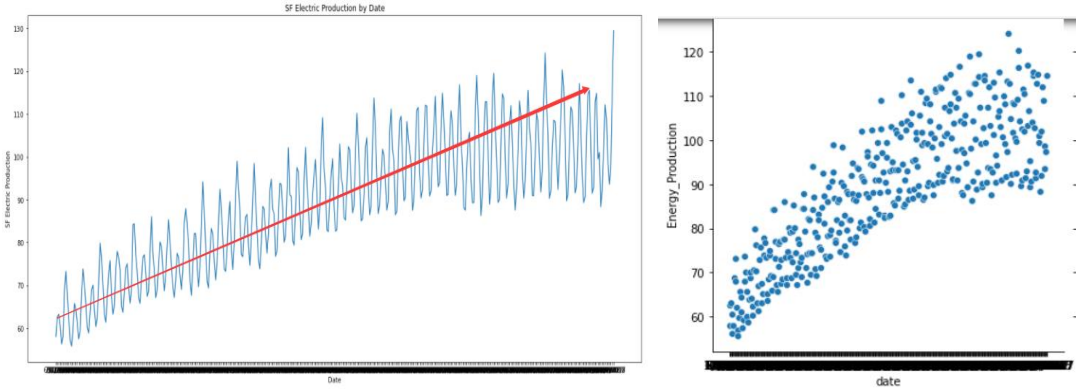
Figure 3. *Data Summary Statistics*

	count	mean	std	min	25%	50%	75%	max
Energy_Production	392.0	89.168193	15.198902	55.8137	77.50675	90.17825	100.72185	129.4048

Graphical representation:

From the graph, we can see that date and energy production has a positive relationship.

Figure 4. *Energy Production Line plot and Dot plot*



Data Processing

Data Frame Alternation

- Rename the column “DATE” as “date”
- Transform the date column to datetime class (format='%m/%d/%Y’).
- Create two columns with column name “ds” and “y” before we train the prophet model.

Figure 5. *Data Frame after Data Processing*

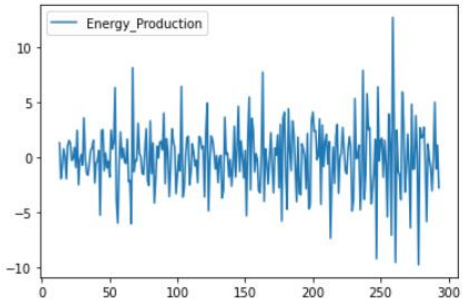
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 392 entries, 0 to 391
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   date               392 non-null    datetime64[ns]
 1   Energy_Production  392 non-null    float64
 2   y                  392 non-null    float64
 3   ds                 392 non-null    datetime64[ns]
dtypes: datetime64[ns](2), float64(2)
memory usage: 12.4 KB
```

Check Stationary for the differenced data

Figure 6. *ADF and KPSS test*

```
stationarity from ad_fuller test: True
stationarity from KPSS test: True
```

Figure 7. *Autocorrelation plot*



Model Assumptions

- For ARIMA model, before we train the data, we should preprocess data until data is stationary, which refers to time series data that mean and variance do not vary across time.
- For both Prophet and Holt Winter's models: Time series data should have seasonal effects. In other words, it takes into account the trend and seasonality while doing the forecasting. Because energy production data has seasonal effects.
- For GARCH model, we should consider heteroskedasticity, which refers to the expected value of all error terms when squared, is not the same at any given point. In other words, variance of error when the variance of the error term is not constant across data points.

Model Selection

Table 1. *Time Series Model Selection*

#	Model	Reasons
1	**Seasonal-ARIMA**	Higher AIC and BIC score than Auto-ARIMA
2	**Auto-ARIMA**	
3	**Prophet**	Account for seasonality or "change points"
4	**Holt Winters**	Control over adjustment on trend and seasonality
	Additive Trend	
	Multi. Trend	
	Addi. Trend & Addi. Seasonality	
	Multi. Trend & Multi. Seasonality	
5	**GARCH**	Minimize errors in forecasting

- There is seasonal patterns and trend in the energy production data. As a result, I select seasonal-ARIMA to manage the seasonal patterns.
- For auto-ARIMA, I want to compare the results with seasonal-ARIMA and see in what degree can auto-ARIMA capture the data patterns.
- For prophet model, I want to capture the seasonality and change points in the data

trend. Since we can observe data trend fluctuation, prophet will be a great model for automatically detects changes in trends by selecting changepoints from the data.

- I also include 4 Holt winters sub models, because I want more control over different trend and seasonality parameters in holt winters model.
- I choose GARCH model, because I observe irregular pattern of variation of an error term in the energy production data. I want to use GARCH to model time series with heteroscedastic errors.

Model Comparison

Forecast Error

Table 2. *Time Series Model Forecast Error*

#	Model	MAE	MSE	MAPE	sMAPE
1	**Auto-ARIMA**	2.8490	14.9680	0.0266	0.0270
2	**Seasonal-ARIMA**	2.7876	13.7157	0.0264	0.0266
3	**Prophet**	3.3636	18.4883	0.0332	0.0326
4	**Holt Winters**				
	Additive Trend	10.3731	168.1676	0.0957	0.1022
	Multi. Trend	22.5524	631.7153	0.2124	0.2433
	Addi. Trend & Addi. Seasonality	4.1027	25.5474	0.0403	0.0395
	Multi. Trend & Multi. Seasonality	3.6916	23.3954	0.0359	0.0355
5	**GARCH**	51.7910	5,496.1475	172.3663	0.8006

Table 3. *Time Series Model AIC BIC Score*

#	Model	AIC	BIC
1	**Auto-ARIMA**	1,324.170	1,349.931
2	**Seasonal-ARIMA**	1,322.224	1,340.625
3	**GARCH**	2,790.220	2,806.090

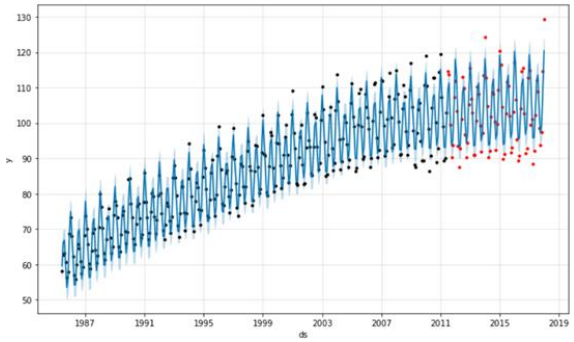
- In this case, I need to choose 4 best-performing models for further analysis according to my project requirement.
- For the first model, I choose SARIMA model over Auto-ARIMA model because SARIMA model has lower AIC and BIC scores. SARIMA model also has slightly lower time series forecast error.
- For the second model, I choose prophet model to account for seasonality or “change points” in the dataset.
- For the third model, I try 4 Holt Winters sub-models and choose Holt Winters with multiplicative trend and multiplicative seasonality because of lowest SMAPE score.
- For the fourth model, I try 5 GARCH models and choose the one with lowest AIC and BIC score (model p=1, q=1).

Forecast Accuracy

- In this step, I need to pick two best-performing models according to project requirement.
- According to error matrix, Seasonal-Arima has the lowest score in error matrix (MAE, MSE, MAPE, SMAPE, MASE) out of all models.
- Prophet model has the second lowest error matrix score.

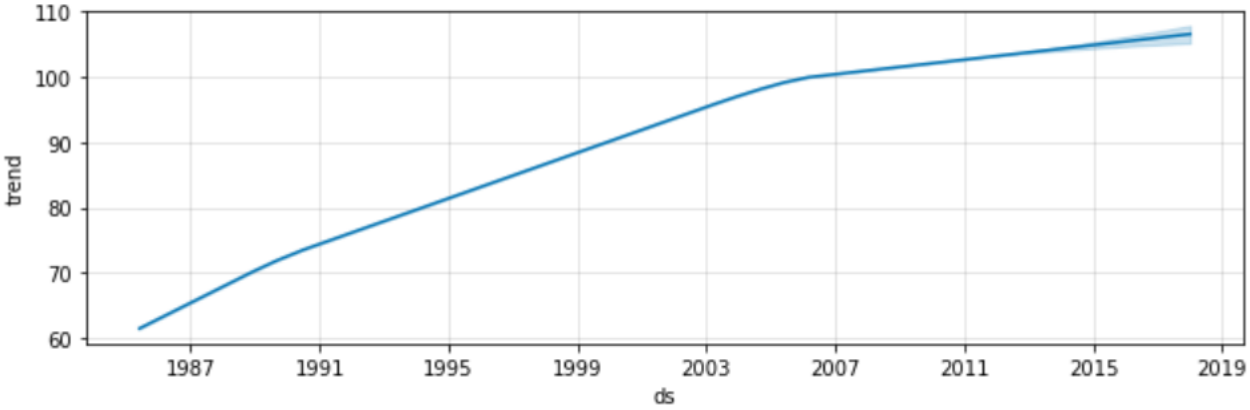
- As a result, I will use Seasonal-Arima and prophet model for forecast accuracy analysis.

Figure 8. *Prophet Prediction Graph (plot scatter)*



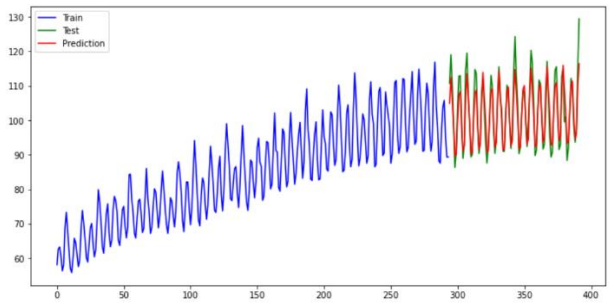
- In figure 8, I plot prophet prediction scatter graph. Red dots accounts for the actual test data, and the line under the test data is prophet model prediction line. We can see the actual data and prediction line are strongly correlated, which indicates a good prediction.

Figure 9. *Prophet Prediction (plot component)*



- In figure 9 prophet prediction component graph, we plot actual data and prediction range generated by prophet model for better visualization. We can see that, from year 2014 to 2019, the light blue area that stick very close to the data line is the forecast data range. The line that goes across the light blue range is the actual data. As a result, from this graph, we can obviously see that the actual data and data prediction range are strongly associated.

Figure 10. *SARIMA Prediction*



- In figure 10 SARIMA prediction, blue line is train data, green line is test data, and red

line is prediction data. We can see that prediction red line is strongly correlated with the actual test data green line. As a result, from this graph, we can obviously see that the SARIMA model made a good prediction on test data.

Analysis Result

- We can see that performance of both SARIMA model and Prophet model are very good because our data, energy consumption, has seasonal patterns.
- Energy consumption data is not volatile or random, and thus avoid the risks of low prediction accuracy in time series models.

Conclusion

- As a result, I expect US energy production continuously growing in the next 5 year. Unless renewable energy growth can cover all the energy production growth, energy production emissions will also be expected to grow in the next 5 years since 88% of energy production are non-renewable.

References

Marohl, Brett. “In 2019, U.S. Energy Production Exceeded Consumption for the First Time in 62 Years.” *U.S. Energy Information Administration (EIA)*, U.S. Energy Information Administration, Monthly Energy Review, 28 Apr. 2020, https://www.eia.gov/todayinenergy/detail.php?id=43515.