Prediction of Singapore’s Monthly Electricity Generation Using Time Series Machine Learning Techniques

Ryan Ueda Teo Shao Ming   
*Singapore Polytechnic*  
*School of Computing*Singapore  
ryanueda.21@ichat.sp.edu.sg

*Abstract*—Singapore's electricity generation has been steadily increasing over the years. This is due to rapid advancements in technology, allowing us to generate electricity more efficiently from various sources. Singapore's electricity generation data is also easily accessible and available via the singstat website, allowing us to analyze the data. In this paper we will explore the data, plot decomposition plots, Autocorrelation (ACF), Partial Autocorrelation (PACF) plots, and use time series machine learning techniques to attempt the future of Singapore’s electricity consumption capabilities.

Keywords—machine learning; decomposition plots; Autocorrelation; Partial Autocorrelation; energy; electricity

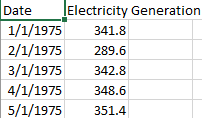
# **Introduction**

Recently, Singapore’s electricity generation capacity has been shown to be steadily increasing, with it being 289.6 GwH in February of 1975, and 4968.7 GwH in May of 2022. This is about a 1615% increase over the course of under 50 years, which is an exponential increase. EMA (Energy Market Authority) has taken steps in recent years to deploy equipment or technologies that are more energy efficient. As of today, 95% of Singapore’s electricity is generated using natural gas. Singapore is exploring ways to access regional power grids that are cost competitive. In addition, Singapore been hitting their targets for solar electricity production, and capitalising on emerging low-carbon alternatives. Although the general trend over the years seems to be increasing, there are dips in values quite often in the data. In this paper we aim to predict accurately the future of Singapore’s electricity generating capacity.

# **DATASET**

This dataset was obtained from the Singapore Government’s data statistics website – singstat.gov.sg. This dataset contains 569 rows and only two columns, Date and Electricity Generated. For every date, the electricity generated is presented in the adjacent column. The data is presented to us in a CSV (Comma Separated Values) file format. A sample of the dataset is shown below:

**Fig. 1. singstat.gov.sg Dataset**



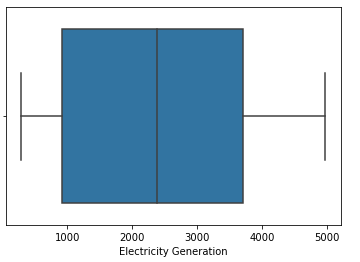
**Fig 2. Attributes in singstat.gov.sg Dataset**

Date: Presented in a Year-Month format for every month from 1975-01-01 to 2022-05-01

Electricity Generated: Amount of electricity generated per month, in gigawatts

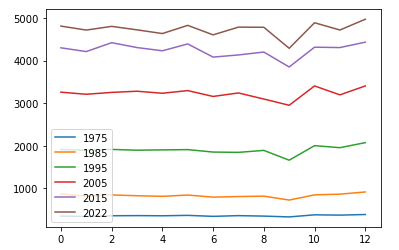
Before proceeding, we performed some EDA (exploratory data analysis) on the dataset to identify any null, nonsensical or extreme (outliers) values that have to be taken care of to ensure that the performance of our model is not affected.

**Fig 3. Boxplot Of Electricity Generation**



From the boxplot, we can see that the highest value is about 5000, while the lowest is a extremely small number, less than 500. The mean electricity generated across the years is around 2500. There are no outliers in the data and the data range seems to be relatively reasonable.

**Fig 4. Electricity Generation Capacity from 1975-2022**

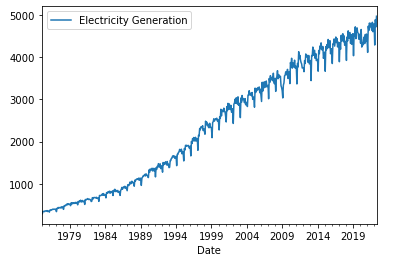


From the line graph plotted above, we can see that Singapore’s electricity generation capacity increases. Around the third quartile of the year however, electricity generation seems to dip slightly, before increasing again.

# **METHODOLOGY**

After performing EDA (Exploratory Data Analysis), we would clean the dataset and impute the values. However, this dataset does not contain any null values or irregular data such as outliers. Hence, we do not have to perform any data pre-processing on this dataset other than changing the index of the dataframe to be Date. We will then fit our data into 3 models: Simple Moving Average (SMA), Holt-Winters Exponential Smoothing, (HWES) and Seasonal Auto Regressive Integrated Moving Average with eXogeneous factors (SARIMAX).

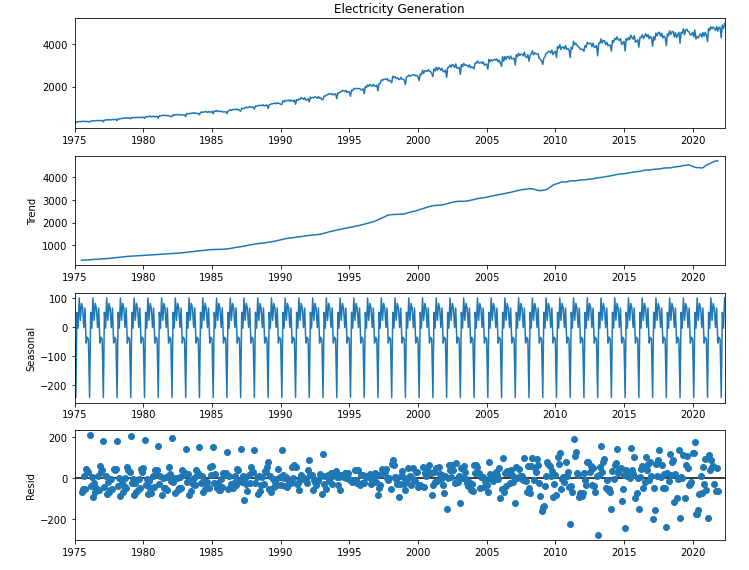
**Fig 5. Electricity Generation Over The Years**



From the run-sequence plot of the time series above, we can see that there is a very strong positive trend in the data and is very likely to be non-stationary. We can confirm this using a decomposition plot, as well as checking the means/variances of the data along with the ADF (Augmented Dickey-Fuller) test to check for stationarity.

## Decomposition Plot

**Fig 6. Decomposition Plot**



Based on the decomposition plot above, we can visually conclude that there is a clear positive trend in the data, and some form of seasonality. This can be confirmed by a quick check.

**Fig 7. Mean and Variance Check**





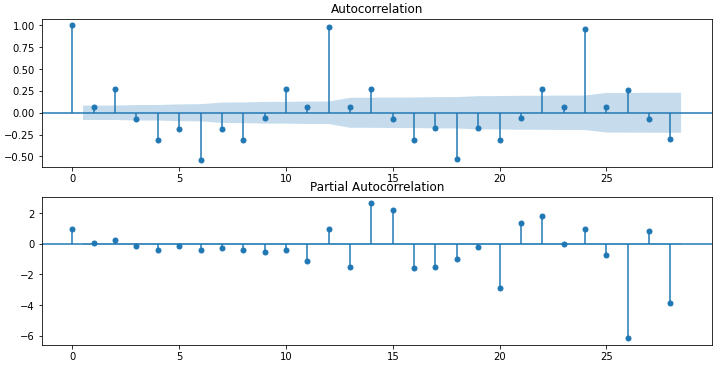
A quick check on the means and variances of the data shows that they have both drastically different means as well as variances, implying that the data is not stationary.

To confirm this, we will perform the ADF (Augmented Dcikey-Fuller) test. Upon conducting this test, it returns a p-value of 0.990521, which is extremely high. This is sufficient evidence to show that we cannot reject the nulll hypothesis that the data is stationary. Hence, we can conclude that the time series is not stationary.

## ACF & PACF Plot (Seasonal Decomposition)

To find the seasonality of the dataset, we can plot an ACF plot.

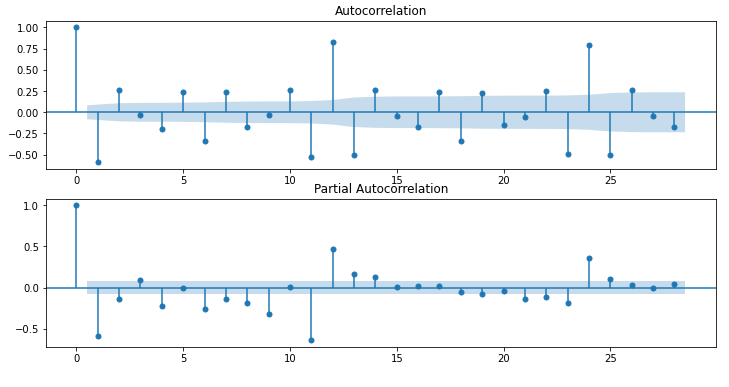
**Fig. 8 ACF and PACF Plots of Seasonal Decomposition**

****

The ACF plot shows that there is a very strong lag at every 12 points. This tells us that the data has seasonality with a period of 12.

## ACF & PACF Plot of the Temperature

## **Fig. 9 ACF and PACF Plots**



Since the time series is not stationary, the time series was differenced once before plotting the ACF and PACF plots.

Based off the ACF and PACF plots above, we are able to identify the p and q values used in SARIMAX, which is the model we will be using in this paper.

From the PACF plot, the values that exceed the confidence interval range from 1-14, excluding 5.

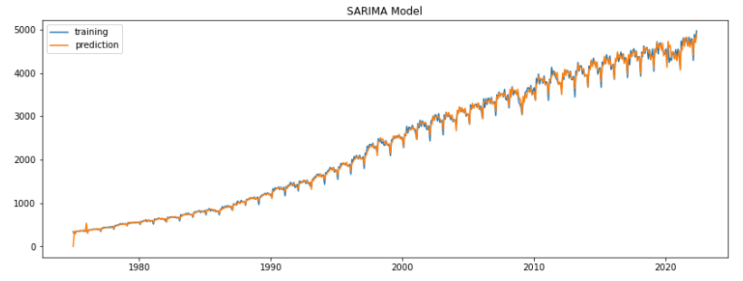
From the ACF plot, almost all the values exceed the confidence interval.

However, by general rule of thumb, we will only look at the first 5 values to be used as p and q values in SARIMAX.

## SARIMAX Model

In the previous section, we have derived our (P,D,Q,S) values, with a differencing order of 1 and seasonality of 12. In the first fitting of our data into the SARIMAX model, we will use an order of (1,1,1) and a seasonal order of (1,1,1,12)

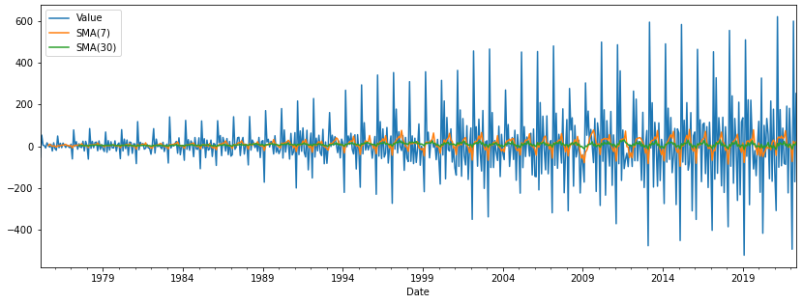
**Fig. 10 SARIMAX Results**



After fitting our data to the SARIMAX model, we can see that the predicted values fit extremely well to the training data. The predicted line fits almost entirely on the training data. This shows that the SARIMAX model is effective in handling our time series data.

## Simple Moving Average Model

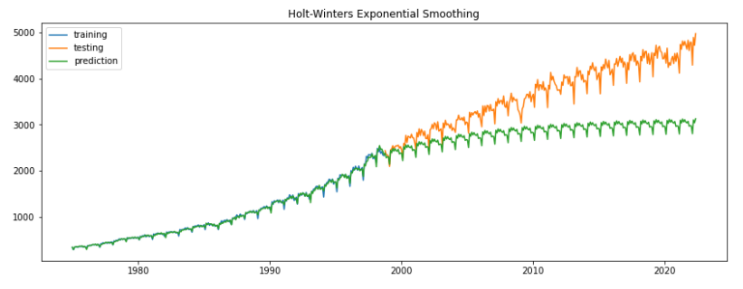
**Fig. 11 Simple Moving Average Model**



Upon plotting a simple moving average model of 7 days and 30 days, we can see that the predictions made do not fit the time series data at all. Although it roughly captures the general trend of the data, the predicted values are very far off, and minimized. This is likely due to the fact that the model only predicts based off taking the average of past values and plotting on them.

## Holt-Winters Exponential Smoothing Model

**Fig. 12 Holt-Winters Exponential Smoothing Model**

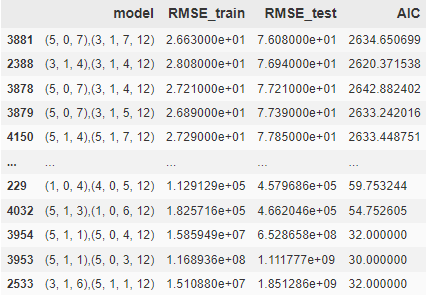


Upon plotting the HWES model, we can see that it fits very well to the train data. However, when it starts predicting on the test data, the predicted values stray off completely and do not converge with the test data at all. This shows that the model is not a good fit for our time series.

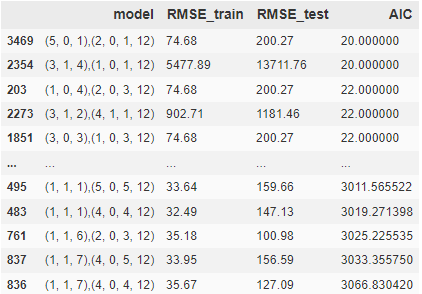
# **Model Improvement – SARIMAX GridSearch**

Upon comparing the 3 models, it is evident that the SARIMAX model is far superior in handling our time series data as compared to the Simple Moving Average, as well as Holt-Winters Exponential Smoothing model. However, we obtained a AIC score of 6283.067 and a RMSE score of 58. Although it boasts an impressive RMSE score, the AIC score can be further improved. To do this, we will perform a simple Grid Search algorithm to find the ideal hyperparameters to fit for the SARIMAX model.

**Fig. 13 GridSearch Results Sorted By RMSE\_test**



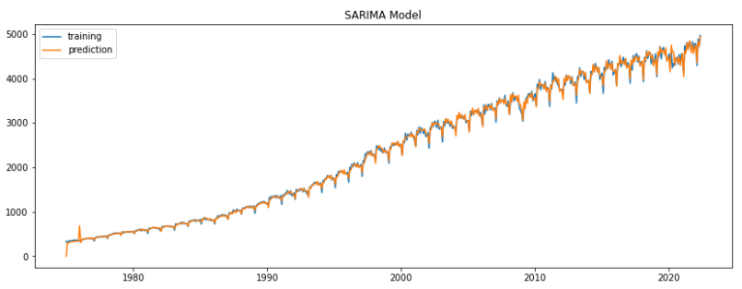
**Fig. 13 GridSearch Results Sorted By AIC**

****

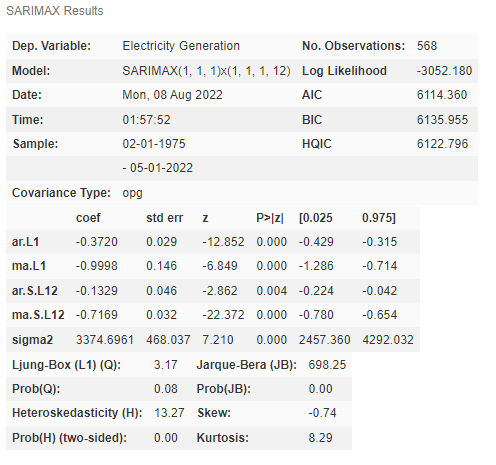
The GridSearch returns a dataframe with all the combinations of orders and seasonal orders, with the RMSE train, test as well as AIC scores. By comparing these two dataframes, we are able to identify the ideal combination of order and seasonal orders to fit into the SARIMAX model. Although there are some combinations with a very low AIC score, it possesses an rather high RMSE. What we are looking for is a nice balance between RMSE and AIC.

# **SARIMAX With Ideal Parameters**

**Fig. 14: SARIMAX With Ideal Parameters**



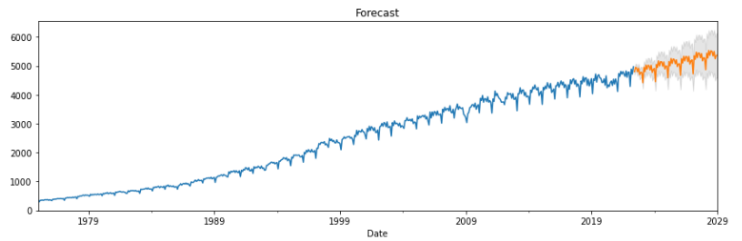
Once again we plot the train data along with the predicted values. As shown from the plot above, we can see that the model works very well with our data as the two lines are effectively on top of each other.





Furthermore, the model possesses an AIC score of 6082.607 and a RMSE of 56, both of which are improvements from the original model.

**Fig. 15: SARIMAX Forecast**

****

Finally, we plot the forecast plot to forecast the next 10 years of electricity generation in Singapore. As shown above, it is forecasted for Singapore’s electricity generation capacity to steadily increase over the next 10 years. The forecast was plotted with a confidence interval, which means that even if the true values do not exactly match the predicted values, it will not exceed the confidence interval. As we predict further into the future, the with of the confidence interval gets proportionately larger. This is natural as predicting into the future on lesser data is bound to result in more errors.

# **CONCLUSION**

In this paper, we have discussed multiple ways to forecast the electricity generation capacity of Singapore along with various machine learning techniques. We explored the methods of checking for stationarity of the time series, such as the Augmented Dickey-Fuller test (ADF), used ACF and PACF plots to understand the components of our time series, used the Simple Moving Average (SMA), Holt-Winters Exponential Smoothing (HWES), as well as Seasonal Auto-Regressive Integrated Moving Average (SARIMAX) models to try to forecast the future of Singapore’s electricity generation capacity. Lastly, we finetuned our model by performing a simple gridsearch to obtain the ideal order and seasonal orders for our SARIMAX model.

# **DATASET LINK**

<https://tablebuilder.singstat.gov.sg/table/TS/M890831>