

Electricity Generation Time Series - Final Project

by Kyle Kim

PSTAT 174 - UCSB

Dr. Gareth Peters

8020554

March 13, 2023

Table of Contents:

1. Executive Summary	3
2. Introduction	4
3. Data Analysis and Experimental Design	5
3.1 Data Selection & Properties	5
3.2 Empirical Analysis	5
3.3 Pre-processing Data for Time Series	6
3.4 Hypothesis Testing on Pre-processed Data	8
4. Time Series Modeling & Forecasting	9
4.1 New York	10
4.1.1 New York - Model Fit & Forecasting	11
4.1.2 Forecasting Comparison & Best Model for NY	15
4.2 California	16
4.2.1 California - Model Fit & Forecasting	17
4.2.2 Forecasting Comparison & Best Model for CA	20
4.3 Florida	21
4.3.1 Florida - Model Fit & Forecasting	21
4.3.2 Forecasting Comparison & Best Model for FL	24
4.4 Texas	25
4.4.1 Texas - Model Fit & Forecasting	26
4.4.2 Forecasting Comparison & Best Model for TX	29
5. Discussion & Conclusion	30
6. Bibliography	32
7. Appendix	33

1. Executive Summary

Electricity generation in households are managed by state governments, agencies and utility companies to maintain the needs of consumers across homes in the U.S. This study explores and analyzes the electricity generation patterns in households from January 2003 to March 2021 for New York, California, Florida and Texas and forecasts the following year until April 2022. These datasets reflect two relatively hotter summer states (CA and FL) and two relatively colder winter states (NY and TX) and with Box-Jenkins methodology, SARIMA models were chosen for each of the datasets after assessing transformation and differencing to pre-process our data, while confirming stationarity and better normality using Augmented Dickey-Fuller, KPSS and Shapiro-Wilk tests. Three of the best models for each dataset were analyzed on their residuals using Ljung-Box-Pierce and Yule-Walker tests and forecasting performance measured by FMSE. As a result of our findings, New York, Florida and Texas had moderately accurate forecasts while California had a poor prediction, which are all explored through more research on the conditions of each state and their extraneous factors. The study shows that the states have variation in electricity generation between hotter summer and colder winter conditions and these extraneous factors can prove to be useful in providing a more accurate forecast based on a more complex model.

2. Introduction

From a household's perspective, electricity is used on a rapid and daily basis. Utility companies and governments help track the usage of electricity consumption in households and allocate funds towards projects to help stabilize the community's energy demand. As electricity utilities are due in increments of months (either monthly or quarterly), it is important for companies and state governments to forecast the estimated household demand in order to prepare for a possible shift in consumption or weather conditions. A possible scenario may be that weather conditions for one year in winter and summer are more extreme than previous years or spring and autumn have colder/hotter days than usual. In response, state governments and agencies can prepare to build and/or upgrade power plants to meet household demand.

This scenario may be true for some states. Note that it is important to note that electricity utility costs in month increments are highly dependent on the appliances used in response to the weather. For hotter summers, appliances like A/C and cooling fans, which utilize electricity, are rapidly used whereas space and water heaters and stoves, which utilize natural gas, are more used the colder the winters.

In 2021, 43% of the electricity generation in the United States was accounted for in the household sector - a large percentage that must be managed in order to maintain the basic needs of household consumers (EIA 2021). If that percentage were to decrease although consumer demand is increasing, these household consumers would face complications in health risks (such as the deaths in Texas's 2021 blackout crisis) and in tending to daily tasks. As high electricity generation for homes will be a consistent trend in the U.S., it is also important for state governments to consider combating factors such as climate change and weather disasters while maintaining their own policies to sustain the environment efficiently. California, for example, is a state that has shown a clear objective in utilizing more renewable sources for electricity and reducing the US's common trend of high natural gas consumption for energy (EIA 2022). The questions of interest explored are:

- (1) Are the forecasts for the models in each state accurate to follow as a basis for stabilizing the electricity generation for households and if not, why?
- (2) Is there a variation in electricity generation between hotter summer and colder winter temperature states in forecasting and what are extraneous factors we can explore that affect electricity generation in households between different states to help?

For the purpose of answering these questions, four states with close rates of average people per household are observed. Two states with hotter summers - California and Florida - and two states with colder winters - New York and Texas - and their electricity generation by month are used in this project to determine the future generation pattern. These states also have different types of trends which we will explore in the empirical analysis of each dataset. Research on different factors such as energy sustainability, household consumer behavior, government policies, and average household size is conducted and will be discussed in the conclusion after analysis and forecasting of the data.

The goal of this project is to perform time series modeling and forecasting on roughly 19 years worth of monthly data from January 2003 to March 2021 in order to forecast the monthly electricity generation for households from April 2021 to March 2022 (12 month period) and answer our questions of interest. The source of this data was from Kaggle, which conveniently had exported a .csv file from a reliable government website, the US Energy Information Administration, which collects different types of energy data across all sectors in the US and in each state.

3. Data Analysis and Experimental Design

In order to tackle these questions of interest, more contextual information on our data is needed. As previously mentioned, four coastal states - New York, California, Florida, Texas - and their monthly household electricity generation data were selected for this project.

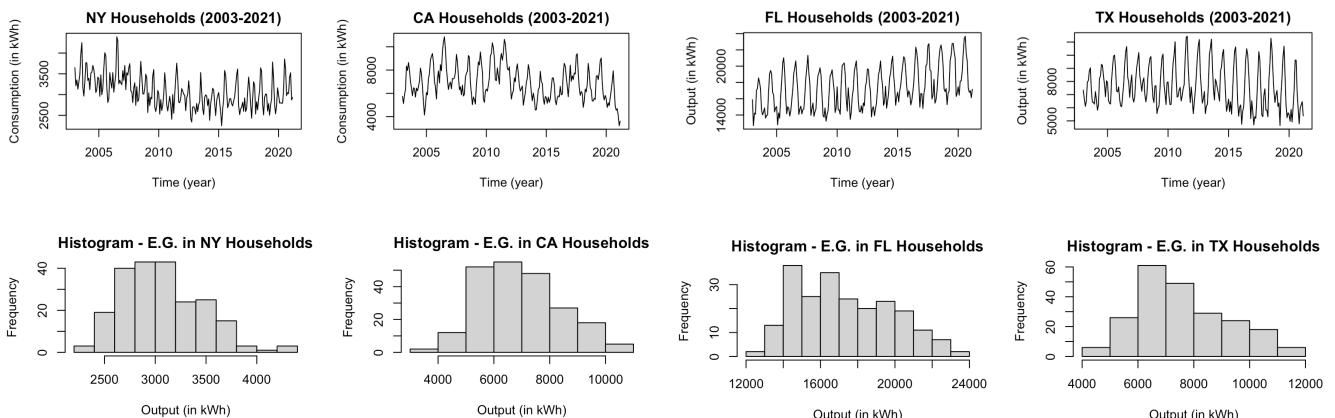
3.1 Data Selection & Properties

The sample rate, or time frame, for each dataset is from January 2003 to March 2022, with a total of 231 observations per dataset. The electricity generation data only looks at the state's household electricity generation per month, with units of kWh (or kilowatt-hours). Let's take a look at the 5 V's data for more context about our own: 1) Volume: There are 231 observations per dataset, equivalent to 924 observations in total. This is a rather smaller volume of data. 2) Value: The data provides an actual representation of electricity generation in households, overviewed by the US EIA, an agency that has primary federal government authority on energy statistics and analysis and is funded by Congress. The reliability of the data indicates high value. 3) Variety: The datasets are split by state: New York (NY), California (CA), Florida (FL), Texas (TX). The source of the datasets is consistent, from the US EIA. 4) Velocity: The datasets are measured at a monthly rate, ranging from January 2003 to March 2022. EIA updates their datasets with a month's worth of data continuing on the previous dataset, typically 2 months lagged from the actual date of update release. The data has moderate velocity. 5) Veracity: There are no missing values in any of the datasets and EIA self-proclaims a 98-100% accuracy on their data collection. The data has good veracity.

Now that we have the sufficient data, a training and test split was performed in order to assess model forecasting performance later. The training set consisted of data from January 2003 to March 2021, or roughly an 18 year period, and the testing set's time frame was from April 2021 to March 2022, roughly a 12 month period in order to perform a forecast for each model produced.

3.2 Empirical Analysis

Before the implementation of modeling on our data, more analysis was needed to figure out the components of pre-processing we need to achieve, which is the first step of Box-Jenkins methodology. These components include stationarity and seasonality and trend, which if are present then our time series is considered not stationary. An initial time series plot and histogram for each state were observed:



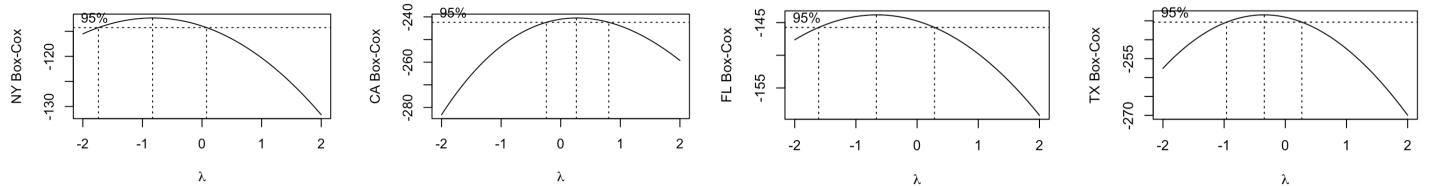
At first glance, there is a seasonal pattern across all four datasets, as there are peaks and valleys on a yearly basis. Additionally, the variance of each dataset is not stabilized which we will account for later in pre-processing. For New York (NY), there appears to be a downward trend then a small upward trend in the latter years of the data. New York, a relatively colder, has a high population density and the coldest summers out of any of the four states. For California (CA), there is a decreasing trend, highlighted by the latter years of the data. California is known to be a hotter state and its electricity generation is highly correlated to climate, as the state is consistently trying to find ways to combat climate conditions and blackouts. In the past, California has aggressively implemented solar panels and renewable energy resources to stabilize the tradeoff between demand and generation.

On the other hand, Florida (FL), typically a hotter state especially in the summers, exhibits an increasing trend and Texas (TX), a harsh winter state, has a slight decreasing trend, with both having slight increases in variance over time. The seasonality aspect is more consistent over time for FL and TX compared to CA and NY. Florida and Texas are also not as ambitious as New York and California in combating climate in relation to energy consumption, using more non-renewable resources.

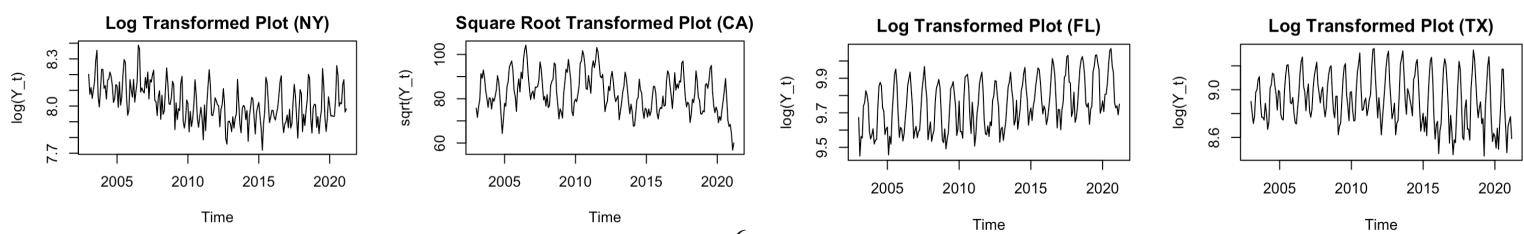
Because of these components of seasonality and trend in each state's dataset, non-stationarity describes the current state of the data. The histograms for each state slightly resemble a normal distribution (with slight skewing) but ultimately a decision to transform and better normalize the data was made.

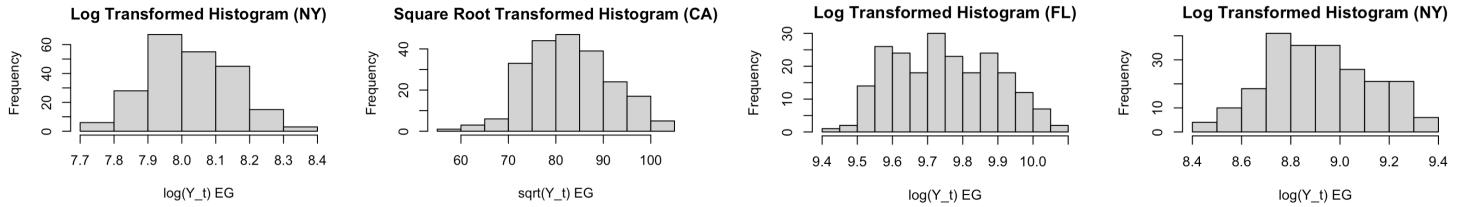
3.3 Pre-processing Data for Time Series

Before dealing with differencing for seasonal and trend structures, the data could be better normalized to help stabilize variance. Our data is positive, meaning that we can use the box-cox transformation to estimate a proper λ in order to transform our data to be more normalized. The plots of the box-cox transformation are given below, with confidence intervals to give a rough estimate for λ :



In terms of transforming the data, it was found that doing log transformation was better suited as the interval for each transformation included 0 for λ for the states of NY, FL, and TX (consulted with Dr. Peters in OH). However, when performing box-cox transformation on CA's data, it was found that log transforming the data actually skews it noticeably, so instead a box-cox transformation where λ for California ($\lambda \approx 0.2626$) is slightly closer to 0.5 works better on our data with a square root transformation. The same logic was applied to NY, FL, and TX data, but logarithmic transformation proved to be the most optimal in all 3 cases. We can also take a look at the plots and histograms for each of the newly transformed datasets:





Now, we are able to assess that the transformed data in each state's histogram is better normalized than the pre-transformed data. However, our time series plots for each state's transformed data still show signs of trend and seasonality and stationarity still needs to be achieved.

After analyzing, differencing was needed across all four transformed datasets to perform modeling since trend and seasonality were still present. In order to remove seasonality and trend to satisfy the rules of stationarity and to be able to fit our models, differencing was decided to be applied. A differencing method was applied at lag 12; this seasonal differencing satisfied the seasonality component of our monthly data. Now we have:

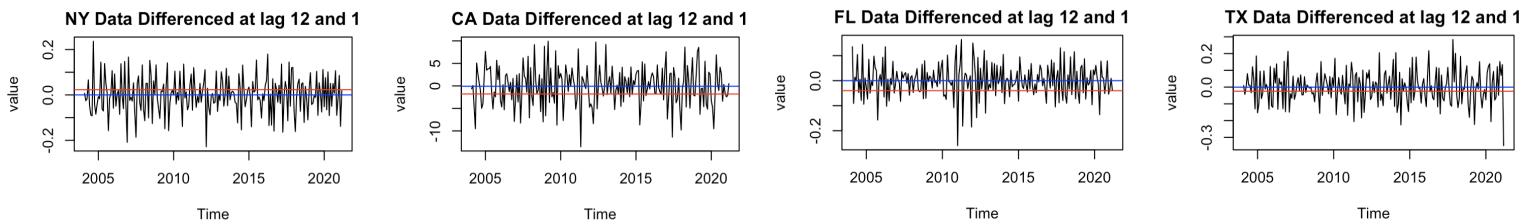
$$(1 - B^{12})Y_t = (Y_t - Y_{t-12})$$

(Plots of the seasonally differenced data at lag 12 is presented in the appendix.)

After differencing at lag 12, trend still may be present in each of our seasonally differenced data plots. We cannot conclude we have fully pre-processed our data yet, as trend being present is a sign for non-stationarity. Let's now apply non-seasonal differencing on our transformed data, differenced at lags 12 and 1. Now we have our new equation:

$$(1 - B^{12})(1 - B)Y_t = (Y_t - Y_{t-1})(Y_{t-12} - Y_{t-13})$$

Let's also take a look at the time series plots of the transformed and differenced data:



We can see now that the variance looks more stabilized after differencing the data at lag 12 and at lag 1. The plots seem to be deseasonalized and overall detrended. The data appears stationary after pre-processing and ACF and PACF plots (shown in time series modeling stage) are more interpretable for modeling. However, before we can begin the next step of model selection, let's run a few hypothesis tests to verify stationarity for each of our time series.

3.4 Hypothesis Testing on Pre-processed Data

To verify stationarity in each of our new transformed and differenced data, we will use the '*tseries*' package in order to utilize the Augmented Dickey-Fuller test, the KPSS test, and the Shapiro-Wilk test to confirm stationarity and normality.

* For each of the Augmented Dickey-Fuller tests, our p-value for each test was < 0.05 , meaning we can reject the Dickey-Fuller's null hypothesis stating non-stationarity or the presence of a unit root and can conclude that each of our time series are stationary after box-cox transformation and differencing.

$$Y_t = \mu + \beta(t) + \alpha(Y_{t-1}) + \sum_{j=1}^p \phi_j \Delta Y_{t-j} + \epsilon_t$$

where $\alpha = 1$ with $p = 0$ as the null hypothesis, α is the coefficient of the first lag on Y (Y_{t-1}) and βt is the trend, and includes high order regressive process $p > 0$ for the augmented version of the test. The results from the AD-F test for our datasets can be found below:

State	NY	CA	FL	TX
p-value	≈ 0.01	≈ 0.01	≈ 0.01	≈ 0.01

* The KPSS test showed that each of the p-values were > 0.05 , meaning we fail to reject the null hypothesis stating that each of the time series are trend stationary, and can conclude that each of our time series are trend stationary after box-cox transformation and differencing.

$$Y_t = \beta(t) + R_t + \epsilon_t$$

where βt is the deterministic trend, $R_t = R_{t-1} + W_t$ is the random walk, ϵ_t is stationary error with zero mean. The results from the KPSS test for our datasets can be found below:

State	NY	CA	FL	TX
p-value	≈ 0.01	≈ 0.01	≈ 0.01	≈ 0.01

* The Shapiro-Wilk test showed that each of the p-values were > 0.05 , meaning that we fail to reject the Shapiro-Wilk null hypothesis that the data is normally distributed, and can conclude that each state's data is normally distributed after box-cox transformation and differencing.

$$W = \frac{\sum_{t=1}^n (a_t y_t)^2}{\sum_{t=1}^n (y_t - \bar{y})^2}$$

State	NY	CA	FL	TX
p-value	≈ 0.689	≈ 0.536	≈ 0.101	≈ 0.95

Now that we have fully pre-processed our time series data and the ADF, KPSS, and Shapiro-Wilk test all pass for each state, we can move on to the model selection stage.

4. Time Series Modeling & Forecasting

Now that our data is stationary, let's now take a look at the ACF and PACF plots (shown for each state during model parameter stage) of each transformed and differenced dataset to determine the potential models and parameters we will use. When selecting the proper potential models for each of these datasets, Box-Jenkins framework will be applied to each state's dataset - we have already pre-processed the data in preparation (Step 1 of Box-Jenkins methodology) for the following steps:

- * Model Selection - examining ACF & PACF plots to select best model to forecast data
- * Parameter Estimation - finding model coefficients for the best fit of the data
- * Model Checking - applying hypothesis tests & analyzing and testing residuals
- * Forecast - compute new forecasted values based on model & compare to actual values

In each of the ACF plots for each state, we see a resemblance to an MA model, weakly stationary (with some spikes slightly outside the 95% CI) and a tail off. The PACF plots show a resemblance to an AR model and also show tail offs. These are both fundamental in the conditions for components of ARMA(p,q). SARIMA was ultimately chosen after determining ARMA with 'I' for the differencing factor and 'S' for the seasonal component we dealt with, as ARMA parameter selection rules apply to SARIMA as well. The model is given:

$$SARIMA(p, d, q)(P, D, Q)_{(S=12)}$$

From the ACF and PACF plots, we can observe the potential parameters for our SARIMA model. To look for P and Q, we take a look at the whole ACF and PACF plots to look at any significant values at the seasonal lags 'l' where there are spikes. To look for p and q (non-seasonal parameters), we take a look at the first set of non-seasonal lags up until lag 12 (or seasonal lag 1) in the ACF and PACF, which will be denoted by 'h'. For d and D (seasonal parameters), they both will equal 1 because there is one seasonal difference and one non-seasonal difference applied to our transformed data. Once we have found potential parameters for the model, we will be using the '*MuMin*' package to produce a search to iterate through models and determine the lowest AICc values for models of each state by running through different combinations of the parameters (method credited to Dr. Coburn in PSTAT 131). Any model that produced a NaN error was eliminated from the final result.

The AICc, or Akaike Information Criterion (and small sample correction) value is calculated as:

$$AICc = -2 \ln L\{\phi_p, \phi_P, \theta_q, \theta_Q, \frac{S(\phi_p, \phi_P, \theta_q, \theta_Q)}{T}\} + \frac{2(M(M+1))}{T-M-1}$$

where the penalty value $M = (p + q + P + Q + 1)$

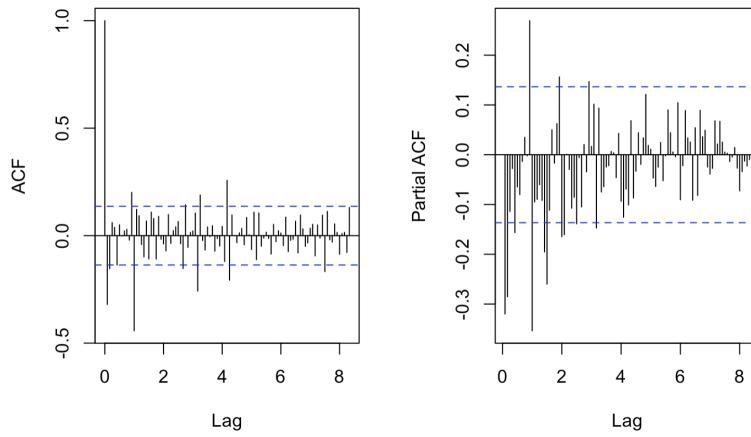
By utilizing Box-Jenkins methodology, a unit circle test will be performed on each of the potential best models to assess stationarity and invertibility. Then, the residuals will be observed for stationarity (or resemblance of white noise as it is an assumption of the SARIMA model) and the Box-Pierce and Ljung-Box tests are applied to see if no significant autocorrelations exist/autocorrelations are independent. The Yule-Walker test fits an AR(p) model based on the PACF of the residuals of the model. If AR(0) is given, we can conclude that the residuals are stationary; if AR(p) where $p > 0$ is given, a unit root test is assessed to see if the AR(p) model of the residuals. If any of these tests fail, the Box-Jenkins approach will allow us to select the next lowest AICc value and redo the process of model testing once again.

Once we have found the 3 best models based on lowest AICc that pass all residual tests, a forecast will be applied at the end and FMSE values are calculated to measure and compare forecasting performance. The FMSE, or the Forecasted Mean Square Error, calculation is shown as:

$$FMSE_h(T) := \frac{1}{h} \sum_{j=1}^h (Y_{T+j} - \hat{Y}_{T+j})^2$$

The first dataset we look at, New York, will provide an understanding of how the Box-Jenkins was applied to the rest of the selection of models.

4.1 New York



Looking at the seasonal ACF, we see a significant peak at seasonal lag $l=1$, so $Q=1$. For the non-seasonal ACF, we see a quick decay at 0 and significant peaks at lags $h=1$ and $h=11$, so $q = 0, 1$, or 11 . Looking at the seasonal PACF, we see significant peaks at seasonal lags $l=1$ and $l=2$. For the non-seasonal PACF, there are significant peaks at $h=1$, $h=2$, and $h=11$, so $p=1$, $p=2$, or $p=11$. $d=1$ because of the seasonal and first differencing applied. Now the SARIMA model with potential parameters is:

$$SARIMA(p=1 \text{ or } 2 \text{ or } 11, d=1, q=0 \text{ or } 1 \text{ or } 11)(P=1 \text{ or } 2, D=1, Q=1)_{(S=12)}$$

We then compute the model using the ‘*purrr*’ package where potential parameters are inputted and the models that produced possible convergence errors were eliminated. Then we take 3 models with the lowest AICc values and fit them using the ‘*astsa*’ package and run more diagnostics on the coefficients and residuals to determine which models of the three are best fit for our forecasting. The code for iterating through the models is included in the appendix and in the R Markdown. Let’s now take a look at our top 3 ranked models for New York:

4.1.1 New York - Model Fit & Forecasting

* **Model i:** $SARIMA(p=1,d=1,q=1)(P=1,D=1,Q=1)_{S=12}$ / $AICc = -577.9282$

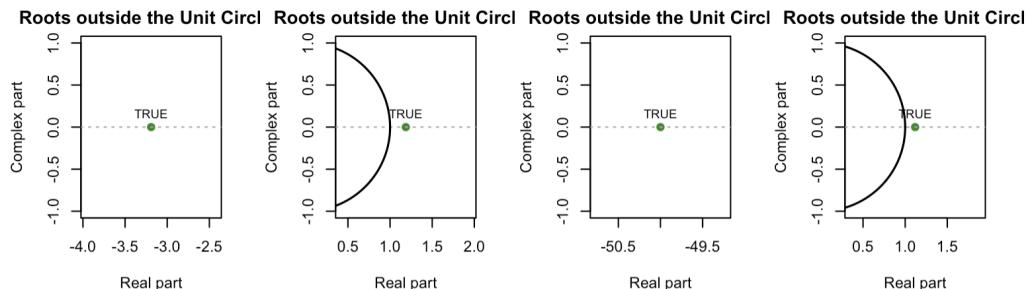
In mathematical equation, this model would translate to:

$$(1 - \phi_1 B)(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t$$

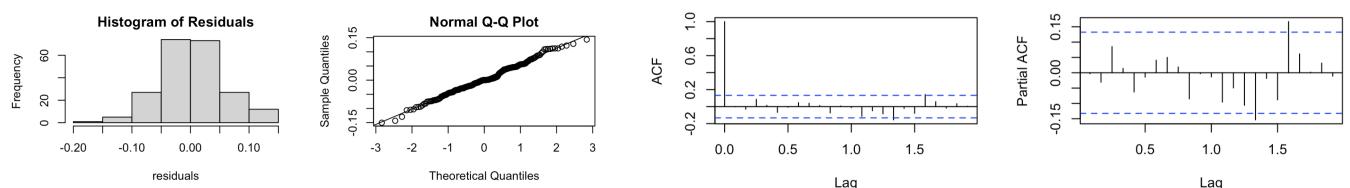
With the estimated coefficients found by the `sarima()` function from the ‘*astsa*’ package, the updated model with the coefficients is:

$$(1 - 0.3134B)(1 - 0.0200B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 - 0.8426B)(1 - 0.8942B^{12})\epsilon_t$$

We then use the package ‘*UnitCircle*’ and the `uc.check()` function to test the roots of our model in order to test for stationarity and invertibility. The roots for each respective AR, MA, SAR, & SMA are outside the unit circle, so we are able to state stationarity and invertibility for Model i:



The residuals of Model i were then observed and tested for white noise as well. The normality and stationary aspects were observed through a histogram, a normality plot and the ACF and PACF plots of the residuals. The histogram and q-q plot show that our residuals are normal and the ACF and PACF plots show a resemblance to white noise, with the exception of a slightly extension beyond the confidence interval at lag 1.6 in the PACF:



To confirm stationarity and exhibiting white noise in the residuals for Model i, Ljung-Box-Pierce and Yule-Walker tests were applied. For the Box-Pierce and Ljung-Box tests, we need to calculate the lag k value when rounded down. $k = \sqrt{220} = 14.83$ so 14. Now, the degrees of freedom for our Box tests will be: $k - (p+q+P+Q) = 14 - (1+1+1+1) = 10$. The formulas for each are given below, and will be the guideline for testing the residuals on the rest of the models:

* The **Box-Pierce Test** examines if the ACF and PACF of residuals are stationary or exhibit white noise, meaning that the autocorrelations are significantly different from zero. The test statistic for the Box-Pierce is:

$$Q_K := T \sum_{k=1}^K (\hat{\rho}\epsilon^2(k)) \sim \chi^2_{K-p-q-P-Q}$$

where T is the length of the sequence after differencing, and p is the autocorrelation, $(K - p - q - P - Q)$ is the degrees of freedom for our Box-Pierce test (as well as the Ljung-Box test).

* The **Ljung-Box Test** is highly similar to the Box-Pierce test in terms of testing if the autocorrelations are independent, but Ljung-Box shows a better approximation of the test statistic (with the same variable assignment applied):

$$Q_K := T(T+2) \sum_{k=1}^K \frac{\hat{\rho}\epsilon^2}{T-k} \sim \chi^2_{K-p-q-P-Q}$$

The results of our tests for the residuals of Model i:

Box-Pierce	$X^2 = 8.3474$, df = 10, p-value = 0.5949
Ljung-Box	$X^2 = 8.782$, df = 10, p-value = 0.5529
Yule-Walker	Order selected 0, σ^2 estimated as 0.002892

Because the p-value for Box-Pierce and Ljung-Box are both > 0.05 , we reject the null hypothesis stating that there are significant autocorrelations in the residuals (stationarity) and the Yule-Walker test produced an AR(0), meaning there is stationarity in the residuals.

A forecast was then performed on our testing set data and evaluated for accuracy, which is shown in the forecast comparison and best model selection for New York's data (4.1.2). We will move on to the rest of the models for NY to select 2 more formidable models for forecasting.

* **Model ii:** $SARIMA(p=2,d=1,q=1)(P=1,D=1,Q=1)_{S=12}$ / $AICc = -575.8084$

In mathematical equation, Model ii would translate to:

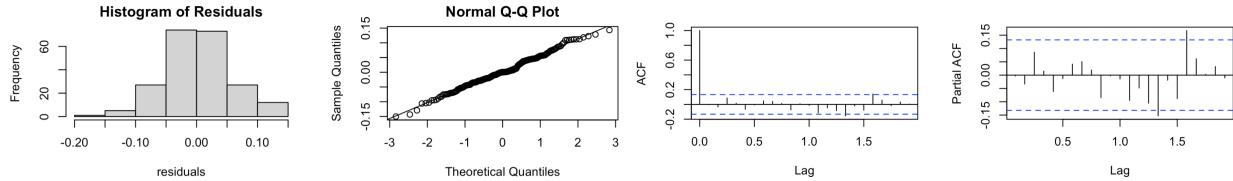
$$(1 + \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t$$

The updated model ii with the coefficients substituted in is:

$$(1 - 0.3164B - 0.0050B^2)(1 - 0.0208B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 - 0.8463B)(1 - 0.8943B^{12})\epsilon_t$$

For the unit root test, the roots for each respective AR, MA, SAR, & SMA are outside the unit circle, so we are able to state stationarity and invertibility for model ii. The plotted results of the unit root test for Model ii are in the appendix.

The residuals for Model ii showed a resemblance to white noise in the ACF and PACF plots and looks normally distributed through the histogram and q-q plots.



To confirm stationarity and exhibiting white noise in the residuals for model ii, Ljung-Box-Pierce and Yule-Walker tests were applied. For the Box-Pierce and Ljung-Box tests, we need to calculate the lag k value when rounded down. $k = \sqrt{220} = 14.83$ so 14. Now, the degrees of freedom for our Box tests will be: $k - (p+q+P+Q) = 14 - (2+1+1+1) = 9$.

Box-Pierce	$X^2 = 8.3706$, df = 9, p-value = 0.4973
Ljung-Box	$X^2 = 8.8029$, df = 9, p-value = 0.4557
Yule-Walker	Order selected 0, σ^2 estimated as 0.002892

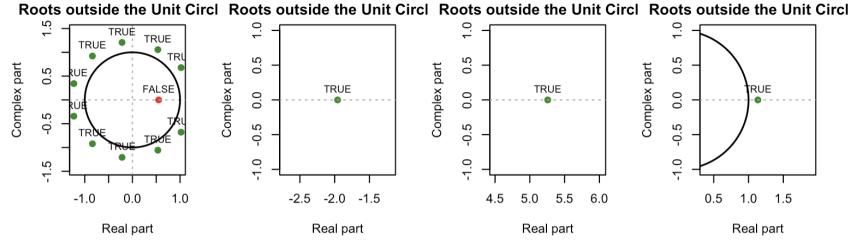
Because the p-value for Box-Pierce and Ljung-Box are both > 0.05 , we reject the null hypothesis stating that there are significant autocorrelations in the residuals (stationarity) and the Yule-Walker test produced an AR(0), meaning there is stationarity in the residuals. Model ii's forecast and performance comparison are found in 4.1.2.

* **Model iii.:** $SARIMA(p=11,d=1,q=1)(P=1,D=1,Q=1)_{S=12}$ / $AICc = -575.8084$

In mathematical equation, this model would translate to:

$$(1 - \phi_1 B - \dots - \phi_{11} B^{11})(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t$$

Before showing the model equation with the corresponding coefficients substituted, it was observed that the respective roots for the polynomials for MA, SAR, and SMA are outside the unit circle. However, the AR component has a root inside the unit circle, indicating that we cannot conclude that the model is stationary and invertible. We will move on to the next best model from our lowest AICc values, as a part of Box-Jenkins methodology.



* **New Model iii.:** $SARIMA(p=2,d=1,q=0)(P=1,D=1,Q=1)_{S=12}$ / $AICc = -566.8084$

In mathematical equation, this model would translate to:

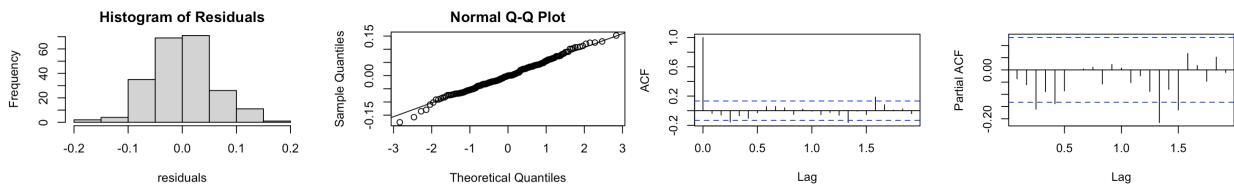
$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 + \Theta_1 B^{12})\epsilon_t$$

The updated new Model iii equation with the coefficients substituted in is:

$$(1 + 0.4313B + 0.3001B^2)(1 - 0.0639B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 - 0.9109B^{12})\epsilon_t$$

For the unit root test, the roots for each respective AR, SAR, & SMA are outside the unit circle, so we are able to state stationarity and invertibility for new Model iii. The plotted results of the unit root test for new Model iii are in the appendix.

The residuals for new Model iii showed a resemblance to white noise in the ACF and PACF plots (with the exception of a few significant lags) and looks normally distributed through the histogram and q-q plots.

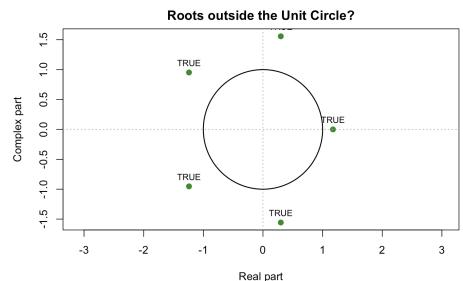


To confirm stationarity exhibiting white noise in the residuals for new model iii, Ljung-Box-Pierce and Yule-Walker tests were applied. For the Box-Pierce and Ljung-Box tests, we need to calculate the lag k value when rounded down. $k = \sqrt{220} = 14.83$ so 14. Now, the degrees of freedom for our tests will be: $k - (p+q+P+Q) = 14 - (2+1+1) = 10$.

Box-Pierce	$X^2 = 13.602$, df = 10, p-value = 0.2202
Ljung-Box	$X^2 = 13.488$, df = 10, p-value = 0.1976
Yule-Walker	Order selected 5, σ^2 estimated as 0.002955

As the Yule-Walker test on the residuals produced an AR(5), the unit root test was applied to see if the roots of the model given were outside the unit circle. As the roots are outside, the model's residuals are stationary:

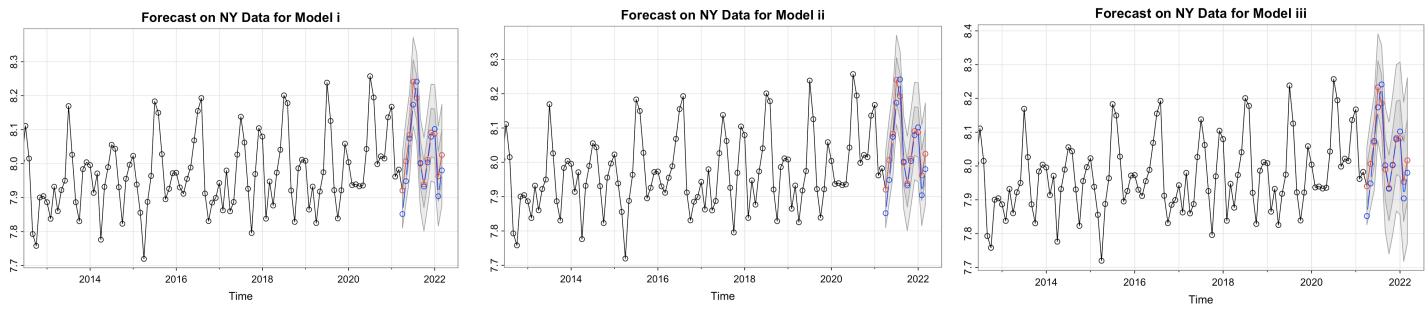
Let's now take a look at the forecasts of each model and compare in the next section.



4.1.2 Forecasting Comparison & Best Model for NY

After forecasting for each of our models using the `sarima.for()` function, it was found that the predictions for each are close to the actual data and the trend and seasonal components were accurately forecasted. The only anomaly found from each model was that each trough was more decreasing than expected from the forecasted values. In addition, the predicted values slightly fall off from the actual values as we travel further through the forecasts. Despite this, peaks and increasing trends were quite accurate. The method of using FMSE values to compare forecast performance is applied to find the best performing model.

- * Model i: SARIMA($p=1, d=1, q=1$) $(P=1, D=1, Q=1)_{S=12}$ / AICc = -577.9282
- * Model ii: SARIMA($p=2, d=1, q=1$) $(P=1, D=1, Q=1)_{S=12}$ / AICc = -575.8084
- * Model iii: SARIMA($p=2, d=1, q=0$) $(P=1, D=1, Q=1)_{S=12}$ / AICc = -566.8084



Red lines represent predicted values, blue lines represent actual values, confidence intervals shaded gray

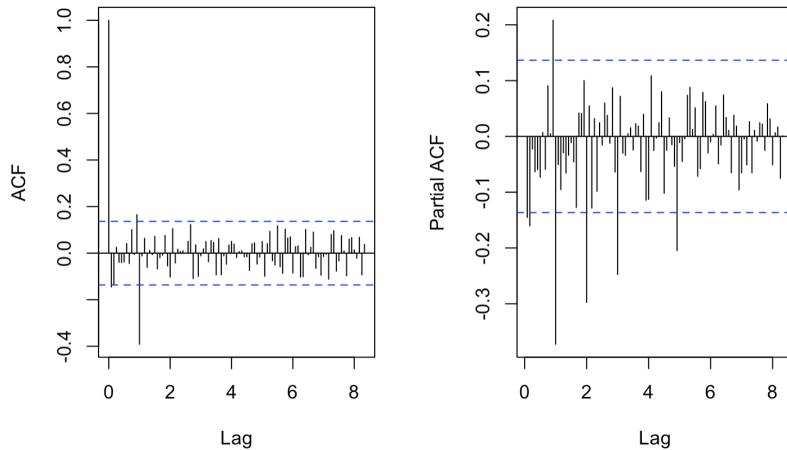
The FMSE values of each forecast performance compared to the actual values are:

Model i	FMSE ≈ 0.001765496
Model ii	FMSE ≈ 0.001735956
Model iii	FMSE ≈ 0.001849743

As a result, Model ii was the best performing model with the lowest FMSE value ≈ 0.001735936 and has the most accurate forecast out of the three models.

$$SARIMA(p=2,d=1,q=1)(P=1,D=1,Q=1)_{S=12}$$

4.2 California



Looking at the seasonal ACF, we see a significant peak only at seasonal lag $l=1$, so $Q=1$. For the non-seasonal ACF, we see a quick decay at 0 and significant peaks at lags $h=1$ and a larger one at $h=11$, so $q = 0, 1, \text{ or } 11$. Looking at the seasonal PACF, we see significant peaks at seasonal lags $l=1, l=2, \text{ and } l=3$, so $P=1, 2, 3$. $P=5$ was considered as well but ultimately resulted to be not as significant as compared to 3 larger and significant peaks. For the non-seasonal PACF, there are significant peaks at $h=1, h=2$ and $h=11$, so $p=1, p=2, \text{ or } p=11$. Again, $d=1$ and $D=1$ because of the seasonal and first differencing applied. Now the SARIMA model with potential parameters is:

$$SARIMA(p=1 \text{ or } 11, d=1, q=0 \text{ or } 1 \text{ or } 11)(P=1 \text{ or } 2 \text{ or } 3, D=1, Q=1)_{(S=12)}$$

After we iterated through the models to assess the AICc values (higher values because of the square root transformation rather than log transformation) of our models, three were chosen:

4.2.1 California - Model Fit & Forecasting

* **Model i:** $SARIMA(p=1,d=1,q=1)(P=1,D=1,Q=1)_{S=12} / AICc = 1103.383$

In mathematical equation, Model i would translate to:

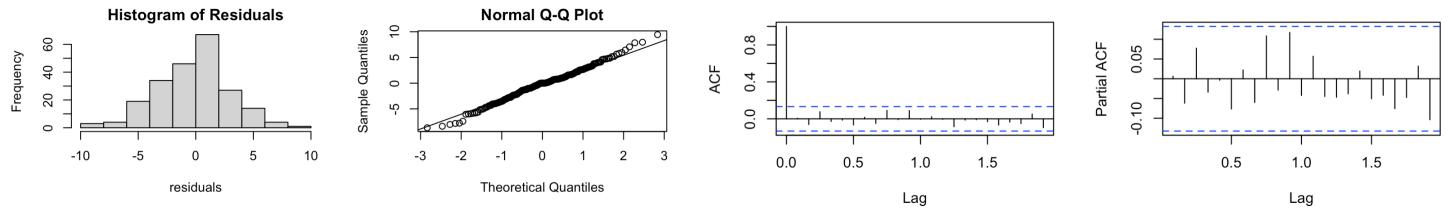
$$(1 - \phi_1 B)(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12})\sqrt{(Y_t)} = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t$$

and with the coefficients substituted in:

$$(1 - 0.7448B)(1 - 0.1276B^{12})(1 - B)(1 - B^{12})\sqrt{(Y_t)} = (1 - 0.8957B)(1 - 0.9177B^{12})\epsilon_t$$

For the unit root test, the roots for each respective AR, MA, SAR, & SMA are outside the unit circle, so we are able to state stationarity and invertibility for Model i. The plotted results of the unit root test for Model i are in the appendix.

The residuals for Model i showed a resemblance to white noise in the ACF and PACF plots and looks normally distributed through the histogram and q-q plots.



To confirm stationarity and exhibiting white noise in the residuals for model ii, Ljung-Box-Pierce and Yule-Walker tests were applied. For the Box-Pierce and Ljung-Box tests, we need to calculate the lag k value when rounded down. The degrees of freedom for our Box tests is: $k-(p+q+P+Q) = 14 - (1+1+1+1) = 10$.

Box-Pierce	$X^2 = 7.5674$, df = 10, p-value = 0.671
Ljung-Box	$X^2 = 7.8954$, df = 10, p-value = 0.6391
Yule-Walker	Order selected 0, σ^2 estimated as 10.08

Because the p-value for Box-Pierce and Ljung-Box are both > 0.05 , we reject the null hypothesis stating that there are significant autocorrelations in the residuals (stationarity) and the Yule-Walker test produced an AR(0), meaning there is white noise in the residuals. Model i's forecast and performance comparison are found in 4.2.2.

$$* \text{Model ii: } SARIMA(p=2,d=1,q=1)(P=1,D=1,Q=1)_{S=12} / AICc = 1104.892$$

In mathematical equation, Model ii would translate to:

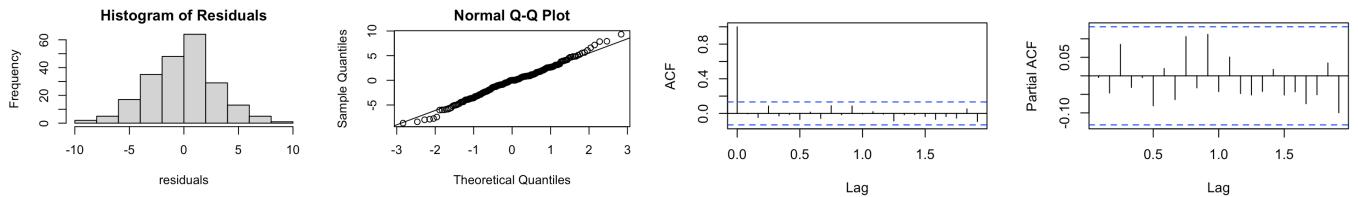
$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12})\sqrt{(Y_t)} = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t$$

and with the coefficients substituted in:

$$(1 - 0.7236B + 0.0319B^2)(1 - 0.1217B^{12})(1 - B)(1 - B^{12})\sqrt{(Y_t)} = (1 - 0.8598B)(1 - 0.9180B^{12})\epsilon_t$$

For the unit root test, the roots for each respective AR, MA, SAR, & SMA are outside the unit circle, so we are able to state stationarity and invertibility for Model ii. The plotted results of the unit root test for Model ii are in the appendix.

The residuals for Model ii showed a resemblance to white noise in the ACF and PACF plots and looks normally distributed through the histogram and q-q plots. They are also highly similar to those in Model i:



To confirm stationarity and exhibiting white noise in the residuals for model ii, Ljung-Box-Pierce and Yule-Walker tests were applied. For the Box-Pierce and Ljung-Box tests, we need to calculate the lag k value when rounded down. The degrees of freedom for our Box tests is: $k - (p+q+P+Q) = 14 - (2+1+1+1) = 9$.

Box-Pierce	$X^2 = 7.6078$, df = 9, p-value = 0.5741
Ljung-Box	$X^2 = 7.9377$, df = 9, p-value = 0.5404
Yule-Walker	Order selected 0, σ^2 estimated as 10.08

Because the p-value for Box-Pierce and Ljung-Box are both > 0.05 , we reject the null hypothesis stating that there are significant autocorrelations in the residuals (stationarity) and the Yule-Walker test produced an AR(0), meaning there is white noise in the residuals. Model ii's forecast and performance comparison are found in 4.2.2.

$$* \text{Model iii: } SARIMA(p=2,d=1,q=0)(P=1,D=1,Q=1)_{S=12} / AICc = 1106.108$$

In mathematical equation, Model iii would translate to:

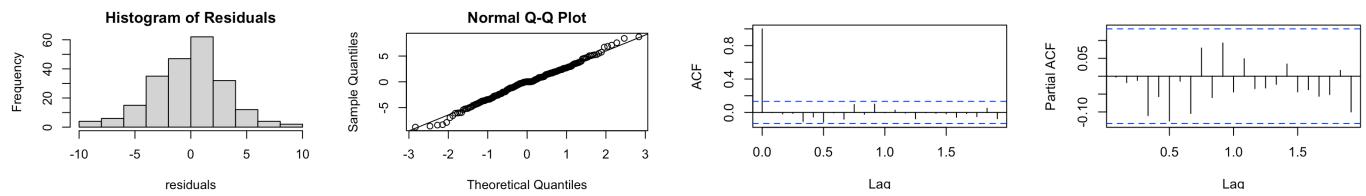
$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12})\sqrt{(Y_t)} = (1 + \Theta_1 B^{12})\epsilon_t$$

and with the coefficients substituted in:

$$(1 + 0.1093B + 0.1535B^2)(1 - 0.1084B^{12})(1 - B)(1 - B^{12})\sqrt{(Y_t)} = (1 - 0.8847B^{12})\epsilon_t$$

For the unit root test, the roots for each respective AR, SAR, & SMA are outside the unit circle, so we are able to state stationarity and invertibility for Model iii. The plotted results of the unit root test for Model iii is in the appendix.

The residuals for Model ii showed a resemblance to white noise in the ACF and PACF plots and looks normally distributed through the histogram and q-q plots:



To confirm stationarity and exhibiting white noise in the residuals for model ii, Ljung-Box-Pierce and Yule-Walker tests were applied. For the Box-Pierce and Ljung-Box tests, we need to calculate the lag k value when rounded down. The degrees of freedom for our Box tests is: $k - (p+q+P+Q) = 14 - (2+1+1) = 10$.

Box-Pierce	$X^2 = 12.384$, df = 10, p-value = 0.2602
Ljung-Box	$X^2 = 12.926$, df = 10, p-value = 0.2279
Yule-Walker	Order selected 0, σ^2 estimated as 10.43

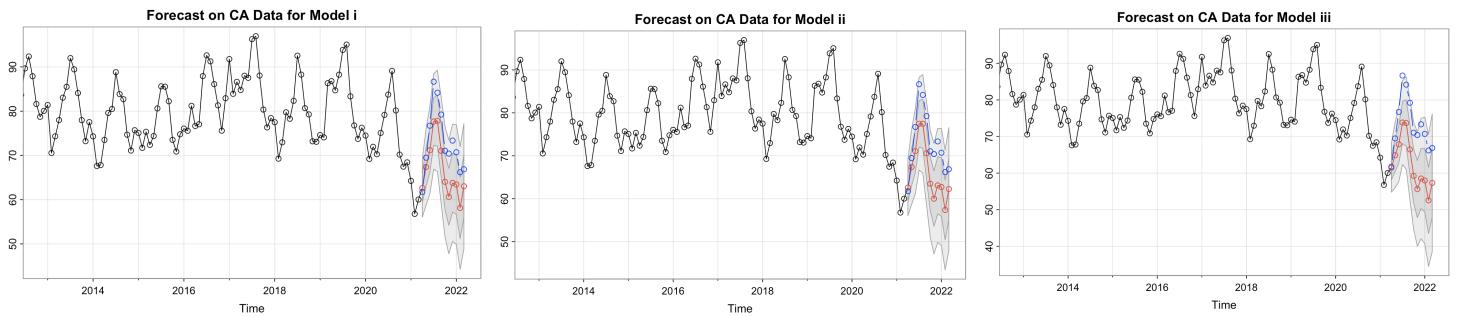
Because the p-value for Box-Pierce and Ljung-Box are both > 0.05 , we reject the null hypothesis stating that there are significant autocorrelations in the residuals (stationarity) and the Yule-Walker test produced an AR(0), meaning there is white noise in the residuals. Model iii's forecast and performance comparison are found with the other models' in the next section.

4.2.2 Forecasting Comparison & Best Model for CA

After forecasting for each of our models, it was found that the predictions are not accurate to the actual values. Starting from the second prediction point (May 2021), every forecast was lower than the

actual value at that month, though within the outer shaded interval. On the other hand, the seasonality aspect was quite accurate, with an increase from March 2021 until July 2021 and a decrease past August 2021. It is interesting to see that the decreasing trend after 2019 was accounted for heavily in the predictions and other factors may have caused this. To explore, more research was conducted on the earlier time interval in California and will be discussed further in the Discussion & Conclusion Section. Again, the method of using FMSE values to compare forecast performance is applied to find the best performing model.

- * Model i: SARIMA($p=1, d=1, q=1$) $(P=1, D=1, Q=1)_{S=12}$ / AICc = 1103.383
- * Model ii: SARIMA($p=2, d=1, q=1$) $(P=1, D=1, Q=1)_{S=12}$ / AICc = 1104.892
- * Model iii: SARIMA($p=2, d=1, q=0$) $(P=1, D=1, Q=1)_{S=12}$ / AICc = 1106.108



Red lines represent predicted values, blue lines represent actual values, confidence intervals shaded gray

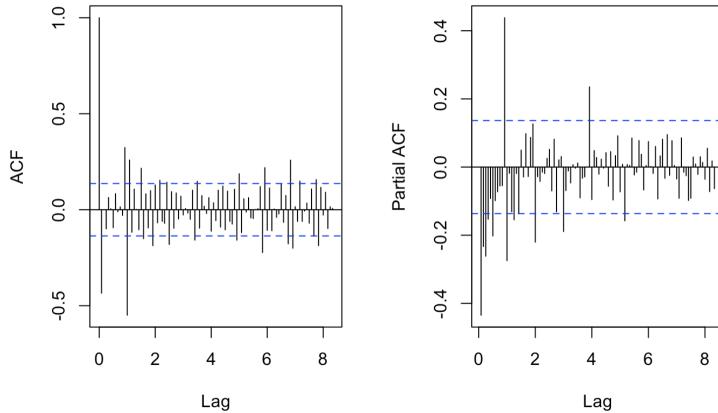
The FMSE values of each forecast performance compared to the actual values are:

Model i	FMSE ≈ 49.44868
Model ii	FMSE ≈ 56.08161
Model iii	FMSE ≈ 129.2983

As a result, Model i was the best performing model with the lowest FMSE value ≈ 49.44868 and is the closest forecast out of the three models, but is overall not useful in terms of accuracy as the prediction is far from the actual values.

$$SARIMA(p=1, d=1, q=1)(P=1, D=1, Q=1)_{S=12}$$

4.3 Florida



Looking at the seasonal ACF, we see a significant peak at seasonal lag $l=1$, so $Q=1$. For the non-seasonal ACF, we see a quick decay at 0 and significant peaks at lags $h=1$ and $h=11$, so $q = 0, 1$, or 11 . Looking at the seasonal PACF, we see significant peaks at seasonal lags $l=1$, $l=2$, and $l=3$. For the non-seasonal PACF, there are significant peaks at lags $h=1$, $h=2$, $h=3$, and $h=11$, so $p=1$, $p=2$, $p=3$, or $p=11$. $d=1$ and $D=1$ because of the seasonal and first differencing applied. Now the SARIMA model with potential parameters looks like this:

$$SARIMA(p=1 \text{ or } 2 \text{ or } 3 \text{ or } 11, d=1, q=0 \text{ or } 1 \text{ or } 11)(P=1 \text{ or } 2 \text{ or } 3, D=1, Q=1)_{(S=12)}$$

After we iterated through the models to assess the AICc values of our models, three were chosen through Box-Jenkins methodology as shown in the next section:

4.3.1 Florida - Model Fit & Forecasting

$$* \text{Model i: } SARIMA(p=1,d=1,q=1)(P=1,D=1,Q=1)_{S=12} / AICc = -718.2448$$

In mathematical equation, Model i would translate to:

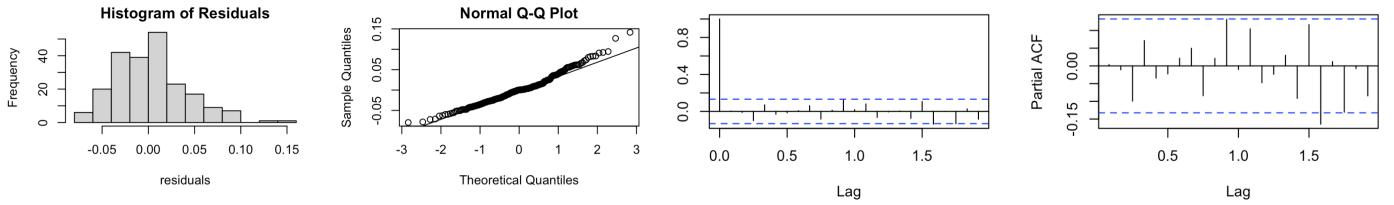
$$(1 - \phi_1 B)(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t$$

and with the coefficients substituted in:

$$(1 - 0.2611B)(1 + 0.0378B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 - 0.9086B)(1 - 0.8951B^{12})\epsilon_t$$

For the unit root test, the roots for each respective AR, MA, SAR, & SMA are outside the unit circle, so we are able to state stationarity and invertibility for Model i. The plotted results of the unit root test for Model i are in the appendix.

The residuals for Model i showed a resemblance to white noise in the ACF and PACF plots and looks normally distributed through the histogram (with a few outlier exceptions that didn't affect the forecast) and q-q plots.



To confirm stationarity and exhibiting white noise in the residuals for Model i, Ljung-Box-Pierce and Yule-Walker tests were applied. For the Box-Pierce and Ljung-Box tests, we need to calculate the lag k value when rounded down. The degrees of freedom for our Box tests is: $k-(p+q+P+Q) = 14 - (1+1+1+1) = 10$.

Box-Pierce	$X^2 = 11.555$, df = 10, p-value = 0.3159
Ljung-Box	$X^2 = 12.148$, df = 10, p-value = 0.2753
Yule-Walker	Order selected 0, σ^2 estimated as 0.001451

Because the p-value for Box-Pierce and Ljung-Box are both > 0.05 , we reject the null hypothesis stating that there are significant autocorrelations in the residuals (stationarity) and the Yule-Walker test produced an AR(0), meaning there is white noise in the residuals. Model i's forecast and performance comparison are found in 4.3.2.

* **Model ii:** $SARIMA(p=1,d=1,q=1)(P=2,D=1,Q=1)_{S=12} / AICc = -716.5408$

In mathematical equation, Model ii would translate to:

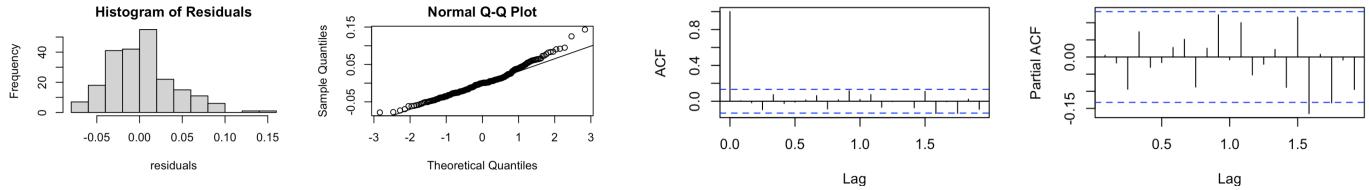
$$(1 - \phi_1 B)(1 - \Phi_1 B^{12} - \Phi_2 B^{24})(1 - B)(1 - B^{12}) \log(Y_t) = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t$$

and with the coefficients substituted in:

$$(1 - 0.2622B)(1 + 0.0152B^{12} - 0.0569B^{24})(1 - B)(1 - B^{12}) \log(Y_t) = (1 - 0.9075B)(1 - 0.9267B^{12})\epsilon_t$$

For the unit root test, the roots for each respective AR, MA, SAR, & SMA are outside the unit circle, so we are able to state stationarity and invertibility for Model ii. The plotted results of the unit root test for Model ii is in the appendix.

The residuals for Model ii showed a resemblance to white noise in the ACF and PACF plots and looks normally distributed through the histogram (with a few outlier exceptions similar to the first model) and q-q plots:



To confirm stationarity and exhibiting white noise in the residuals for Model ii, Ljung-Box-Pierce and Yule-Walker tests were applied. For the Box-Pierce and Ljung-Box tests, we need to calculate the lag k value when rounded down. The degrees of freedom for our Box tests is: $k-(p+q+P+Q) = 14 - (1+1+2+1) = 9$.

Box-Pierce	$X^2 = 11.123$, df = 9, p-value = 0.2674
Ljung-Box	$X^2 = 12.691$, df = 9, p-value = 0.2313
Yule-Walker	Order selected 0, σ^2 estimated as 0.001434

Because the p-value for Box-Pierce and Ljung-Box are both > 0.05 , we reject the null hypothesis stating that there are significant autocorrelations in the residuals (stationarity) and the Yule-Walker test produced an AR(0), meaning there is white noise in the residuals. Model ii's forecast and performance comparison are found in 4.3.2.

* **Model iii:** $SARIMA(p=2,d=1,q=1)(P=1,D=1,Q=1)_{S=12}$ / $AICc = -716.3363$

In mathematical equation, Model ii would translate to:

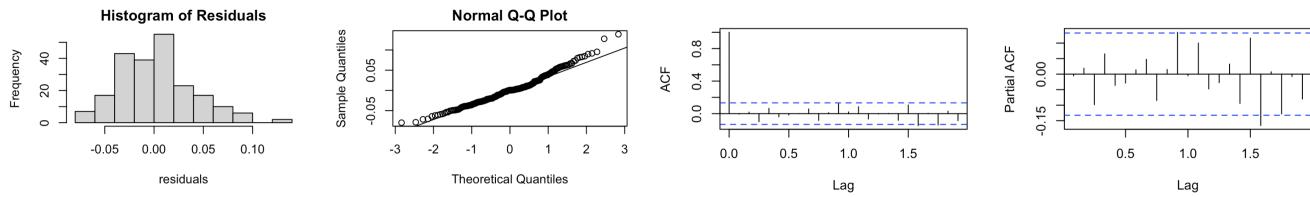
$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t$$

and with the coefficients substituted in:

$$(1 - 0.2622B + 0.0378B^2)(1 + 0.0458B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 - 0.8983B)(1 - 0.8969B^{12})\epsilon_t$$

For the unit root test, the roots for each respective AR, MA, SAR, & SMA are outside the unit circle, so we are able to state stationarity and invertibility for Model iii. The plotted results of the unit root test for Model iii are in the appendix.

The residuals for Model iii showed a resemblance to white noise in the ACF and PACF plots and looks normally distributed through the histogram (with a few outlier exceptions similar to the first and second model) and q-q plots:



To confirm stationarity and exhibiting white noise in the residuals for Model iii, Ljung-Box-Pierce and Yule-Walker tests were applied. For the Box-Pierce and Ljung-Box tests, we need to calculate the lag k value when rounded down. The degrees of freedom for our Box tests is:
 $k-(p+q+P+Q) = 14 - (2+1+1+1) = 9$.

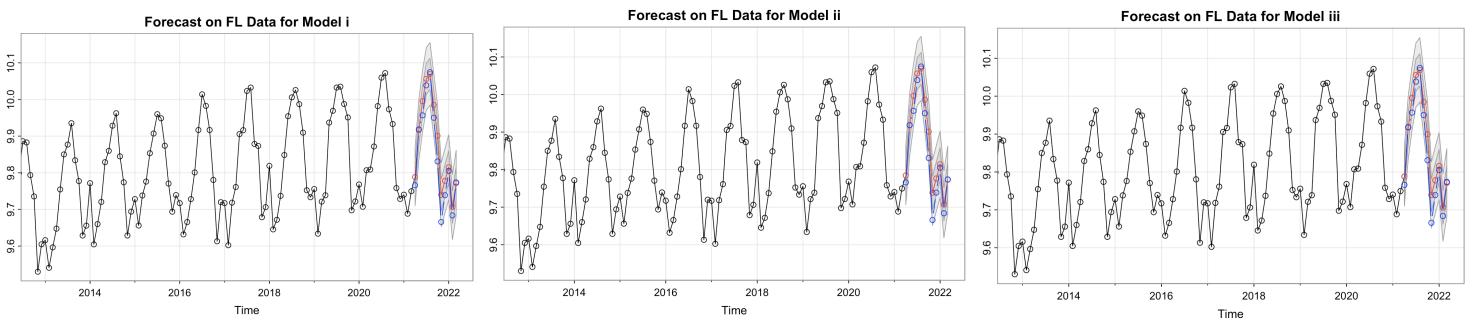
Box-Pierce	$X^2 = 11.555$, df = 9, p-value = 0.2396
Ljung-Box	$X^2 = 12.148$, df = 9, p-value = 0.2051
Yule-Walker	Order selected 0, σ^2 estimated as 0.001451

Because the p-value for Box-Pierce and Ljung-Box are both > 0.05 , we reject the null hypothesis stating that there are significant autocorrelations in the residuals (stationarity) and the Yule-Walker test produced an AR(0), meaning there is white noise in the residuals. Model iii's forecast and performance comparison are found in the next section along with Model i and Model ii.

4.3.2 Forecasting Comparison & Best Model for FL

After forecasting for each of our models, it was found that the predictions for each are very close to the actual data and the trend and seasonal components were closely forecasted. However, in each model, the predicted values from October 2021 until February 2022 are slightly higher than the values of our actual testing set data. Despite this, peaks and trends were quite accurate and the seasonal component is still accounted for in the forecast. The same method of using FMSE values to compare forecast performance is applied to find the best performing model:

- * Model i: SARIMA($p=1,d=1,q=1$)($P=1,D=1,Q=1$) $_{S=12}$ / AICc = -718.2448
- * Model ii: SARIMA($p=1,d=1,q=1$)($P=2,D=1,Q=1$) $_{S=12}$ / AICc = -716.5408
- * Model iii: SARIMA($p=2,d=1,q=1$)($P=1,D=1,Q=1$) $_{S=12}$ / AICc = -716.3363



Red lines represent predicted values, blue lines represent actual values, confidence intervals shaded gray

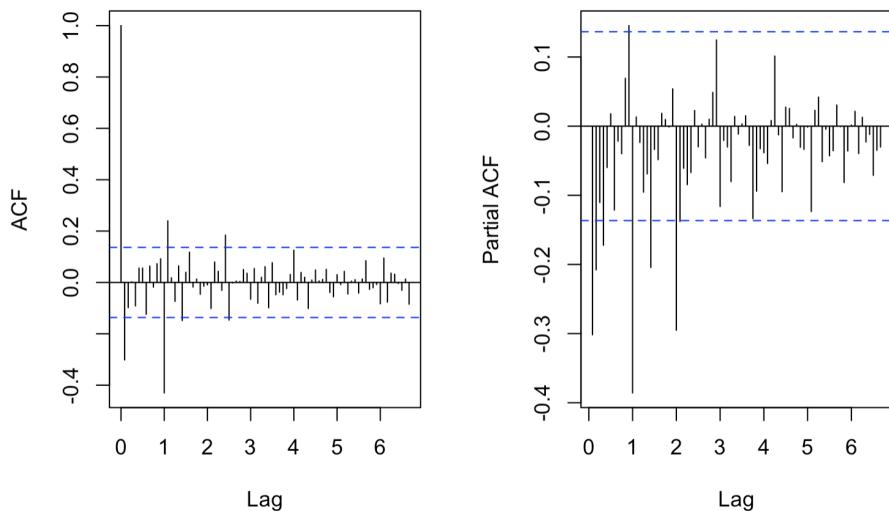
The FMSE values of each forecast performance compared to the actual values are:

Model i	FMSE ≈ 0.001342143
Model ii	FMSE ≈ 0.001317129
Model iii	FMSE ≈ 0.001311017

As a result, Model i was the best performing model with the lowest FMSE value ≈ 0.001311017 and has the most accurate forecast out of the three models.

$$SARIMA(p=2,d=1,q=1)(P=1,D=1,Q=1)_{S=12}$$

4.4 Texas



Looking at the seasonal ACF, we see a significant peak at seasonal lag $l=1$, so $Q=1$. For the non-seasonal ACF, we see a quick decay at 0 and significant peaks at lag $h=1$, so $q=0$ or $q=1$. Looking at the seasonal PACF, we see significant peaks at seasonal lags $l=1$ and $l=2$. For the non-seasonal PACF, there are significant peaks at lags $h=1$, $h=2$, and $h=4$, so $p=1$, $p=2$, or $p=4$. $p=11$ was considered but the same consideration applied from previous model parameter decision making. $d=1$ and $D=1$ because of the seasonal and first differencing applied. Now the SARIMA model with potential parameters looks like so:

$$SARIMA(p=1 \text{ or } 2 \text{ or } 4, d=1, q=0 \text{ or } 1)(P=1 \text{ or } 2, D=1, Q=1)_{(S=12)}$$

After we iterated through the models to assess the AICc values of our models, three were chosen:

4.4.1 Texas - Model Fit & Forecasting

* **Model i:** $SARIMA(p=1, d=1, q=1)(P=2, D=1, Q=1)_{S=12} / AICc = -469.9877$

In mathematical equation, Model i would translate to:

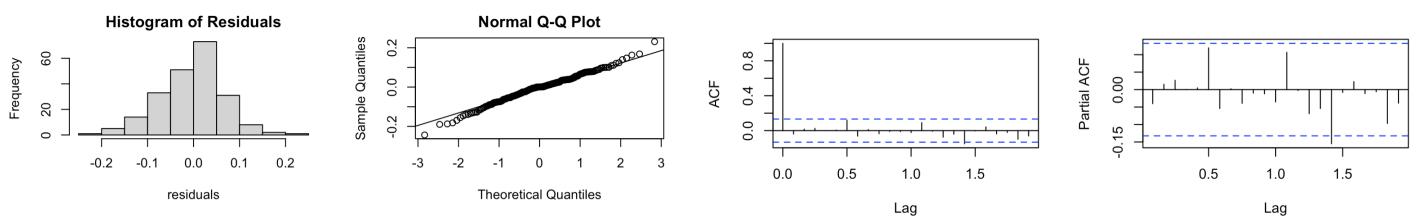
$$(1 - \phi_1 B)(1 - \Phi_1 B^{12} - \Phi_2 B^{24})(1 - B)(1 - B^{12}) \log(Y_t) = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t$$

and with the coefficients substituted in:

$$(1 - 0.5351B)(1 + 0.1794B^{12} + 0.1583B^{24})(1 - B)(1 - B^{12}) \log(Y_t) = (1 - 0.9526B)(1 - 0.5512B^{12})\epsilon_t$$

For the unit root test, the roots for each respective AR, MA, SAR, & SMA are outside the unit circle, so we are able to state stationarity and invertibility for Model i. The plotted results of the unit root test for Model i are in the appendix.

The residuals for Model i showed a resemblance to white noise in the ACF and PACF plots and looks normally distributed through the histogram and q-q plots.



To confirm stationarity and exhibiting white noise in the residuals for Model i, Ljung-Box-Pierce and Yule-Walker tests were applied. For the Box-Pierce and Ljung-Box tests, we need to calculate the lag

k value when rounded down. The degrees of freedom for our Box tests is: $k-(p+q+P+Q) = 14 - (1+1+2+1) = 9$.

Box-Pierce	$X^2 = 6.8585$, df = 9, p-value = 0.6518
Ljung-Box	$X^2 = 7.1829$, df = 9, p-value = 0.6181
Yule-Walker	Order selected 0, σ^2 estimated as 0.005028

Because the p-value for Box-Pierce and Ljung-Box are both > 0.05 , we reject the null hypothesis stating that there are significant autocorrelations in the residuals (stationarity) and the Yule-Walker test produced an AR(0), meaning there is white noise in the residuals. Model ii's forecast and performance comparison are found in 4.4.2.

* **Model ii:** $SARIMA(p=2,d=1,q=1)(P=1,D=1,Q=1)_{S=12}$ / $AICc = -469.7540$

In mathematical equation, Model ii would translate to:

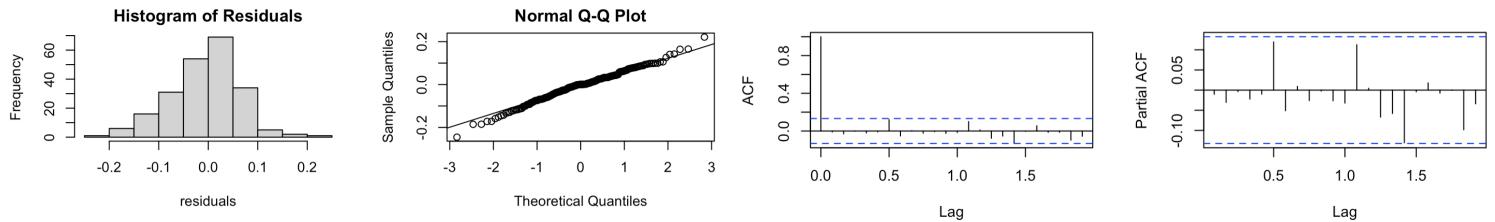
$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t$$

and with the coefficients substituted in:

$$(1 - 0.5171B - 0.0690B^2)(1 + 0.0486B^{12})(1 - B)(1 - B^{12}) \log(Y_t) = (1 - 0.9644B)(1 - 0.6650B^{12})\epsilon_t$$

For the unit root test, the roots for each respective AR, MA, SAR, & SMA are outside the unit circle, so we are able to state stationarity and invertibility for Model ii. The plotted results of the unit root test for Model ii are in the appendix.

The residuals for Model ii showed a resemblance to white noise in the ACF and PACF plots and looks normally distributed through the histogram and q-q plots.



To confirm stationarity and exhibiting white noise in the residuals for Model ii, Ljung-Box-Pierce and Yule-Walker tests were applied. For the Box-Pierce and Ljung-Box tests, we need to calculate the lag

k value when rounded down. The degrees of freedom for our Box tests is: $k-(p+q+P+Q) = 14 - (2+1+1+1) = 9$.

Box-Pierce	$X^2 = 6.6921$, df = 9, p-value = 0.6691
Ljung-Box	$X^2 = 7.0266$, df = 9, p-value = 0.6343
Yule-Walker	Order selected 0, σ^2 estimated as 0.005067

Because the p-value for Box-Pierce and Ljung-Box are both > 0.05 , we reject the null hypothesis stating that there are significant autocorrelations in the residuals (stationarity) and the Yule-Walker test produced an AR(0), meaning there is white noise in the residuals. Model ii's forecast and performance comparison are found in 4.4.2.

* **Model iii:** $SARIMA(p=2,d=1,q=1)(P=2,D=1,Q=1)_{S=12} / AICc = -468.8533$

In mathematical equation, Model iii would translate to:

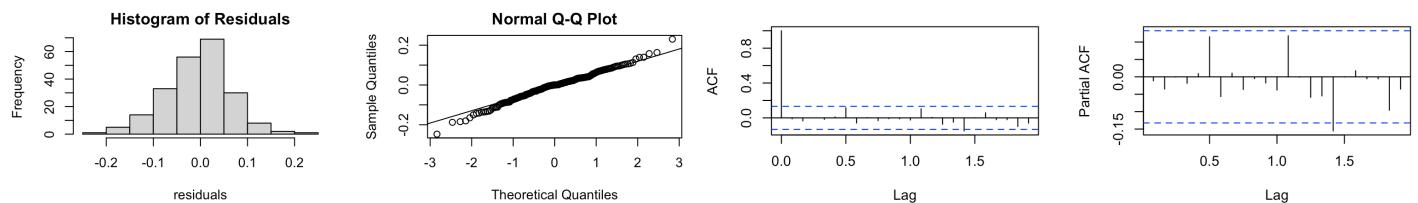
$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12} - \Phi_2 B^{24})(1 - B)(1 - B^{12}) \log(Y_t) = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t$$

and with the coefficients substituted in:

$$(1 - 0.5077B - 0.0779B^2)(1 + 0.1672B^{12} + 0.1642B^{24})(1 - B)(1 - B^{12}) \log(Y_t) = (1 - 0.9652B)(1 - 0.5504B^{12})\epsilon_t$$

For the unit root test, the roots for each respective AR, MA, SAR, & SMA are outside the unit circle, so we are able to state stationarity and invertibility for Model iii. The plotted results of the unit root test for Model iii are in the appendix.

The residuals for Model iii showed a resemblance to white noise in the ACF and PACF plots and looks normally distributed through the histogram and q-q plots:



To confirm stationarity and exhibiting white noise in the residuals for Model iii, Ljung-Box-Pierce and Yule-Walker tests were applied. For the Box-Pierce and Ljung-Box tests, we need to calculate the lag k value when rounded down. The degrees of freedom for our Box tests is: $k-(p+q+P+Q) = 14 - (2+1+2+1) = 8$.

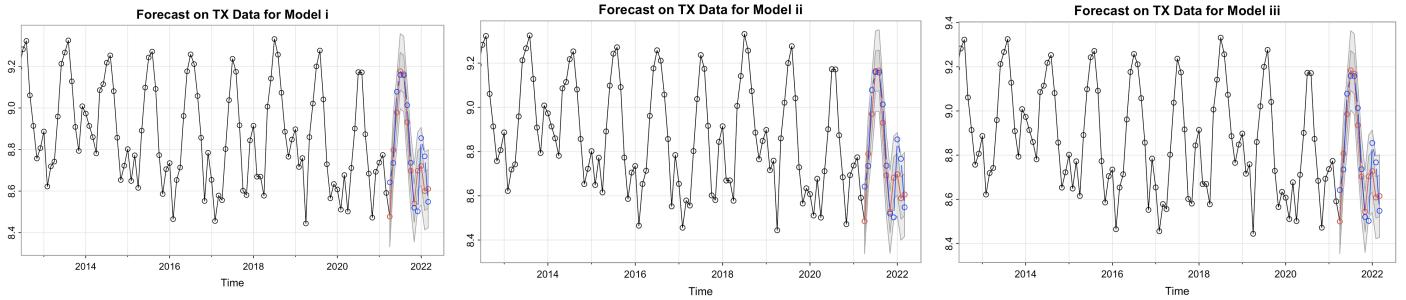
Box-Pierce	$X^2 = 6.9095$, df = 9, p-value = 0.5468
Ljung-Box	$X^2 = 7.256$, df = 9, p-value = 0.5093
Yule-Walker	Order selected 0, σ^2 estimated as 0.004993

Because the p-value for Box-Pierce and Ljung-Box are both > 0.05 , we reject the null hypothesis stating that there are significant autocorrelations in the residuals (stationarity) and the Yule-Walker test produced an AR(0), meaning there is white noise in the residuals. Model iii's forecast and performance comparison are found in the next section along with Model i and Model ii.

4.4.2 Forecasting Comparison & Best Model for TX

After forecasting for each of our models, it was found that the predictions for each are close to the actual data and the trend and seasonal components were accurate in the beginning half of the forecast. As expected and as we have seen previously, the predictions get worse the later in the forecast with the actual values being higher than the predicted values at the peaks in the end of 2021/beginning of 2022. Despite this, seasonality and overall trend were quite accurate in the forecast. Same MSE method applies:

- * Model i: SARIMA(p=1,d=1,q=2)(P=1,D=1,Q=1) $_{S=12}$ / AICc = -469.9877
- * Model ii: SARIMA(p=2,d=1,q=1)(P=1,D=1,Q=1) $_{S=12}$ / AICc= -469.7540
- * Model iii: SARIMA(p=2,d=1,q=1)(P=2,D=1,Q=1) $_{S=12}$ / AICc= -468.8533



Red lines represent predicted values, blue lines represent actual values, confidence intervals shaded gray

The FMSE values of each forecast performance compared to the actual values are:

Model i	FMSE ≈ 0.01150192
Model ii	FMSE ≈ 0.01170401
Model iii	FMSE ≈ 0.01073718

As a result, Model iii was the best performing model with the lowest FMSE value ≈ 0.01073718 and has the most accurate forecast out of the three models:

$$SARIMA(p=2,d=1,q=1)(P=2,D=1,Q=1)_{S=12}$$

5. Discussion & Conclusion

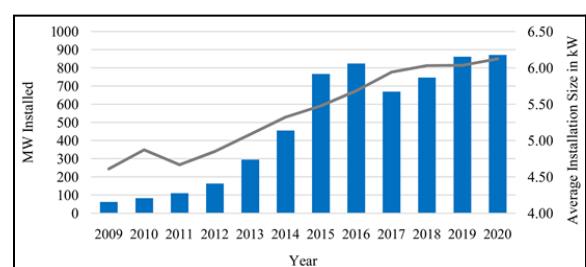
After observing the forecasts of electricity generation for each of the four states, the best models that performed quite accurately for the next year's prediction are:

- * New York - $SARIMA(p=2,d=1,q=1)(P=1,D=1,Q=1)_{S=12}$
- * Florida - $SARIMA(p=2,d=1,q=1)(P=1,D=1,Q=1)_{S=12}$
- * Texas - $SARIMA(p=2,d=1,q=1)(P=2,D=1,Q=1)_{S=12}$

Another note observed was that the first 6 months are more accurate than the latter 6 months for New York, Florida, and Texas, which is expected from a univariate model like ours. On the other hand, California's best performing model was not accurate in forecasting future electricity generation, with the predicted values for the next year interval being far lower than the actual values from April 2021 to March 2022.

- * California - $SARIMA(p=1,d=1,q=1)(P=1,D=1,Q=1)_{S=12}$

With these forecast results, we are able to explore more into answering the questions of interest. Colder winter states such as New York and Texas are more suitable to predict with our models than hotter summer states such as California and Florida - there is a variation in electricity generation between hotter summer and colder winter temperature states in forecasting. In general, consistent seasonal patterns throughout the time frame of the datasets was a determining factor in the accuracy of the forecasts. Florida and Texas had the same consistent seasonal pattern and non-erratic trend from 2003 to 2021 and New York's trend wasn't as erratic as California's through the time period, which we can use to differentiate the accuracy in the forecasts. California, a hotter state, was not as accurate as Florida, another hot state. While exploring the forecast for California, the forecasted values were still able to account for the seasonal component from the data but because of the erratic trend from 2003 to around 2013 being different from the decreasing trend we see after 2013. As previously mentioned, California is aggressively implementing ways to combat climate change such as the heat waves and droughts that are faced on a yearly basis. In 2013, California saw a 200% increase in the installation of solar panels and has grown exponentially since then (Ybarra, Broughton, & Nyer

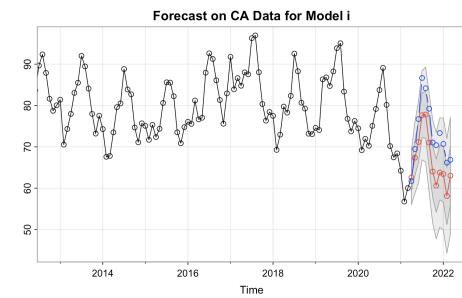


2021). The increase of solar panels has reduced the demand of electricity power from plants and overall generation. The following graph above from the Chapman article shows the trend.

According to the Solar Energy Industries Association in 2021, California was the leading solar capacity user for electricity generation by a large margin compared to the other three states and has continued this trend (SEIA 2021). This factor combined with the unexpectancy of COVID-19 from 2020 to 2022 where lockdowns and social distancing increased electricity consumption rates at home may have factored into the prediction believing it was continuing California's decreasing trend from 2019. Therefore, our models are more suitable to predict for colder winter states, compared to hotter summer states.

Some possible extraneous variables were also explored such as renewable energy policies, power grid and plant infrastructure and notable time periods where the weather was extreme. As previously mentioned, California and New York are two states who are adamantly opposed to an increase of non-renewable energy, whereas Florida and Texas are not as determined. States with more non-renewable energy sources could have a correlation to a more accurate prediction. Notable time periods where states experienced above average/extreme summers and winters were recorded as well. California also experienced their hottest summer to date in 2021, which could correlate to the prediction lacking in the actual values for our forecast (Smith 2021). However, Texas also experienced a power crisis in 2021 with three severe storms, but our forecast was able to still accurately predict how the generation would highly increase in the first half of 2021. By exploring the extraneous factors in each state and implementing them in a more sophisticated model, state governments and agencies such as EIA can maintain the large percentage that households account for in electricity generation across all states.

As a result, the models for New York, Florida, and Texas could be useful in predicting the next 6 months but not reliable. However, more variables and complex models to suit them could explain and account for other factors that aren't included that result in a poor forecast, such as in the case of California.

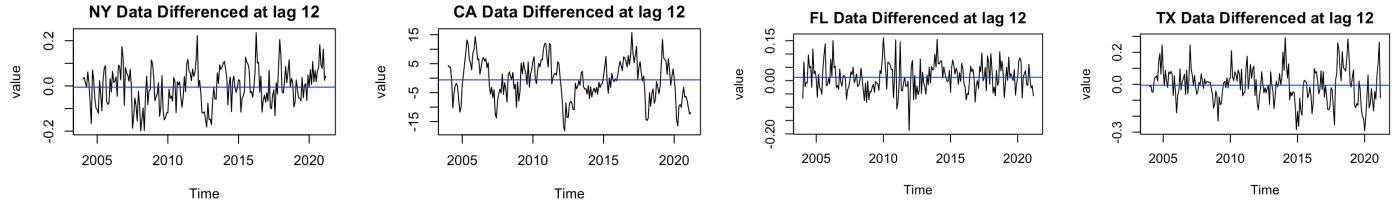


6. Bibliography

- Hyndman, R.J. (2001). *Box-Jenkins modelling*. [online] Available at:
<https://robjhyndman.com/papers/BoxJenkins.pdf>.
- Hyndman, R.J. and Athanasopoulos, G. (2018). *Forecasting : Principles and Practice. 2nd ed.* [online] Heathmont, Vic.: Otexts. Available at: <https://otexts.com/fpp2/>.
- Shah, R. (2022) *Energy Generation Time Series Data*. [online] Kaggle. Available at:
<https://www.kaggle.com/datasets/ravishah1/electricity-generation-time-series>.
- Smith, H. (2021). *California records its hottest summer ever as climate change roils cities*. [online] Los Angeles Times. Available at:
<https://www.latimes.com/california/story/2021-09-09/california-records-hottest-summer-amid-heat-wave-flex-alert>.
- Tsay, R.S. (2013). An Introduction to Analysis of Financial Data with R. Hoboken, N.J.: Wiley.
- U.S. Energy Information Administration (EIA) (n.d.) API Dashboard - Electric Power Operational Data, API Dashboard - U.S. Energy Information Administration (EIA). [online] Available at:
<https://www.eia.gov/opendata/browser/electricity/electric-power-operational-data>.
- U.S. Energy Information Administration (EIA) (n.d.). About EIA - Our work - U.S. Energy Information Administration (EIA). [online] Available at: https://www.eia.gov/about/mission_overview.php.
- U.S. Energy Information Administration (EIA) (2022). U.S. Energy Facts Explained - Consumption and Production - U.S. Energy Information Administration (EIA). [online] Eia.gov. Available at:
<https://www.eia.gov/energyexplained/us-energy-facts/>.
- US EPA, O. (2016). *California Prepares for Increased Wildfire Risk to Air Quality From Climate Change*. [online] US EPA. Available at:
<https://www.epa.gov/arc-x/california-prepares-increased-wildfire-risk-air-quality-climate-change>.
- www.statskingdom.com. (n.d.). Shapiro Wilk Test. [online] Available at:
[https://www.statskingdom.com/doc_shapiro_wilk.html#:~:text=Shapiro%20Wilk%20test%20with%20tables&text=W%3D\(%E2%88%91ni%3D](https://www.statskingdom.com/doc_shapiro_wilk.html#:~:text=Shapiro%20Wilk%20test%20with%20tables&text=W%3D(%E2%88%91ni%3D).
- Ybarra, C.E., Broughton, J.B. and Nyer, P.U. (2021). Trends in the Installation of Residential Solar Panels in California. [online] Low Carbon Economy, pp.63–72. Available at:
<https://www.scirp.org/journal/paperinformation.aspx?paperid=111563>.

7. Appendix

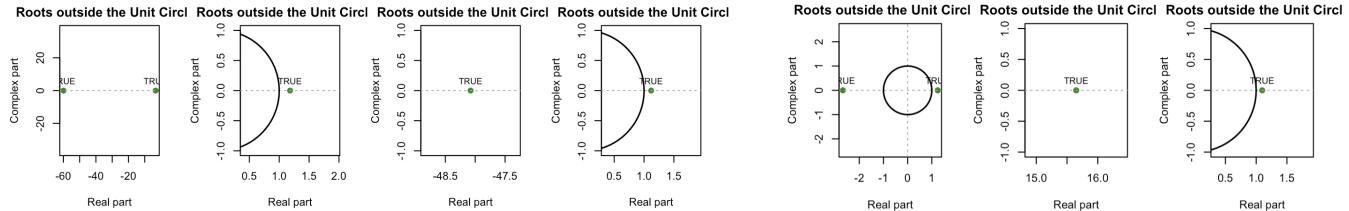
* time series plots of transformed data, differenced at lag 12 for *NY*, *CA*, *FL*, and *TX* (left to right) (pg. 7)



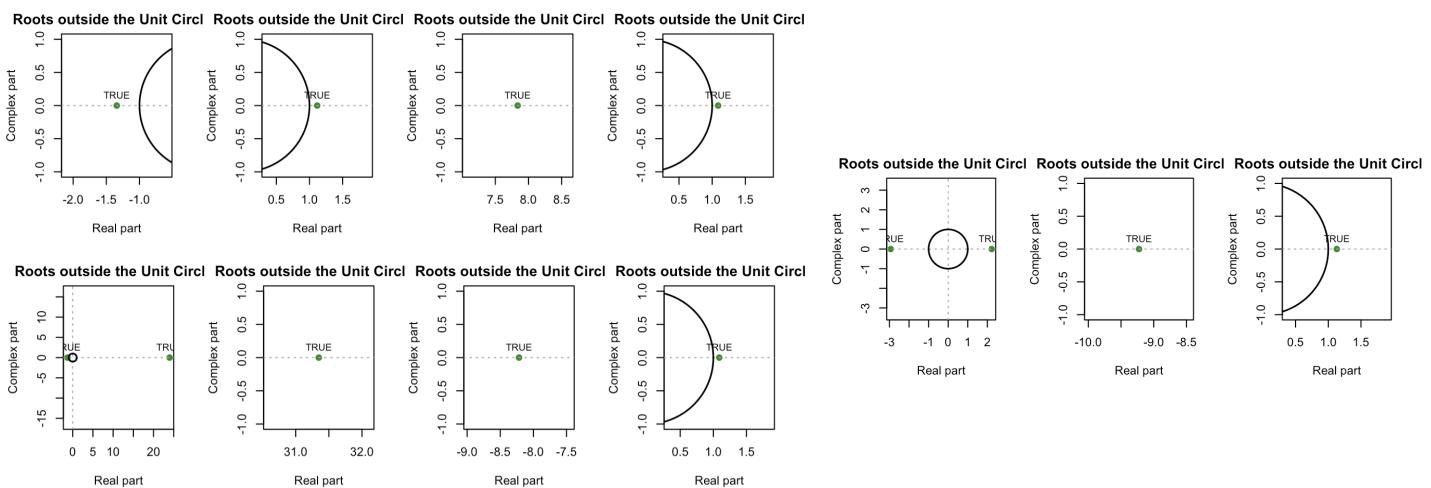
* model search code (ex. for *NY*) - removing errors and getting AICc values (pg. 11)

```
```{r, warning=FALSE, eval=FALSE}
Model Search
NY_model <- expand.grid(p=c(1,2,11), q=c(0,1,11), P=c(1,2), Q=1)
NY_model$AICc <- map_dbl(1:nrow(NY_model), function(i) {
 tryCatch({
 NY_arima.obj <- arima(NY_train_trsf,
 order=c(NY_model$p[i], 1, NY_model$q[i]),
 seasonal=list(order=c(NY_model$P[i], 1, NY_model$Q[i]), period=12),
 method="ML")
 AICC(NY_arima.obj)
 }, error = function(e) NA_real_)
})
NY_model <- NY_model[!is.na(NY_model$AICc),] # remove NaN errors
```

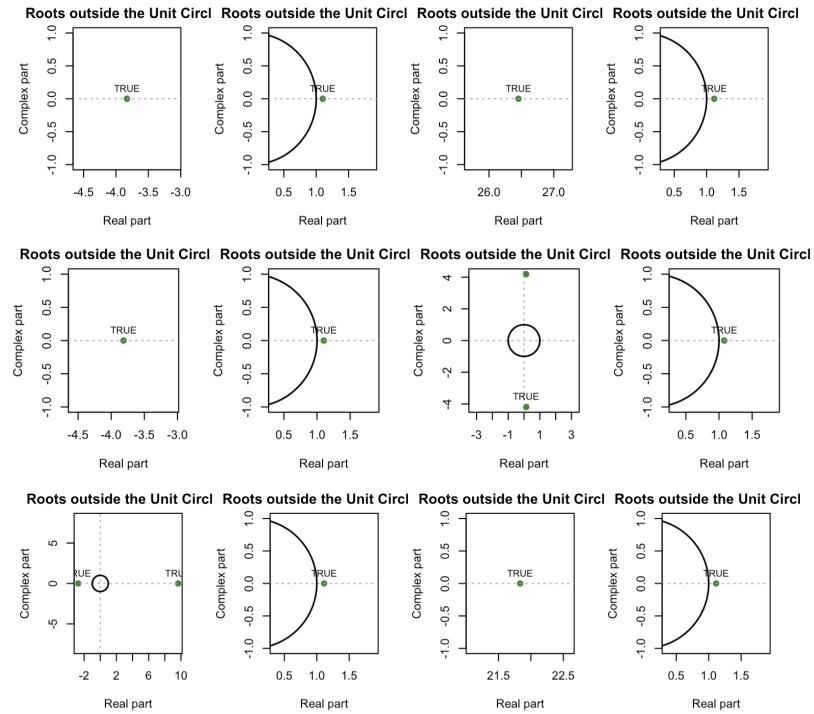
\* unit root test for *NY* Model ii (left) & *NY* new Model iii (right) (pg. 13 & 14)



\* unit root test for *CA* Model i (top), *CA* Model ii (bottom), and *CA* Model iii (right) (pg. 17, 18, & 19)



\* unit root test for FL Model i, FL Model ii, and FL Model iii (in order) (pg. 22, 23, & 24)



\* unit root test for TX Model i, TX Model ii, and TX Model iii (in order) (pg. 26, 27, & 28)

