Logo, company name

Description automatically generated

**Application Of SARIMA Models to Sales and Astronomy Forecasting**

MA 641 Time Series Analysis

Professor Hadi Safari Katesari

April 9th, 2023

|  |  |
| --- | --- |
| Joshua Grou | jgrou@stevens.edu |
| Mez Edoka  Theodore Steffens  Mark Niewierowski | tsteffen@stevens.edu |

Table of Contents

**Introduction1**

**Data Sets1**

**Shampoo Sales Data Set and Modeling4**

**Shampoo Sales Model Conclusions6**

**Sunspot Count Data Set and Modeling6**

**Sunspot Count Model Conclusions11**

**References12**

**Introduction**

In this study data exploration, model selection, model analysis, and forecasting are performed across two different datasets exhibiting different properties. By applying these techniques it is possible to develop models that are able to forecast the future with a degree of confidence, by applying these principles many things can be accomplished such ensuring there is enough product for a store to sell on a given day or understanding when a profit opportunity might exist in the stock market. In this study two datasets have been selected, the first is a sales data set from a shampoo company that is non-stationary and non-seasonal. The second data is an astronomy data set counting the monthly number of sunspots over a period of many years. Each data set has its own challenges that will need to be solve before time series models can be applied to them.

**Data**

The first data set analyzed is the monthly sales of shampoo. This data set was sourced from jbrownlee of the machine learning mastery blog website. He has posted many tutorials on topics in machine learning and time series analysis and hosts a large collection of datasets on his github. This data set is short containing 36 data points and is thus simple to work with, serving as a great introduction to SARIMA model selection, analysis, and forecasting. The data set contains no missing or obviously outlying values, thus no null or outlier handling will need to be considered. A plot of the data set is shown below.

Chart, line chart

Description automatically generated

The above plot gives the intuition that the data set is non-stationary as there is a clear upward trend in the data, with potentially changing variability. This non-stationarity will need to be treated before fitting a time series model. There is not strong evidence of any seasonal components, but this will need to be investigated more thoroughly.

The second data set was also sourced from jbrownlee, this data set is from the astronomy domain and records a monthly count of sunspots. This data set is more challenging both due to its length and strong evidence of a seasonal component. A plot of the data set is shown below.

Chart

Description automatically generated

From the plot there is decent evidence to support that there may be some seasonal trend. The data set also looks like it may be stationary which might not require treatment. One worrying part of the dataset is that there are over 2600 observations, this number of observations might make it difficult to interpret results from ACF and PACF analysis. If troubles are run into there the first strategy that can be tried is to aggregate the data yearly rather than monthly, this will collapse down the number of observations to a more reasonable amount to work with.

**Shampoo Sales Data Set Analysis**

The first step in the Box-Jenkins approach will be to check if data is stationary and if not to make it stationary through some transformation. We can test for stationarity using the augmented Dickey-Fuller test. After applying the augmented Dickey-Fuller test, the resulting p-value is 1. This indicates very strong evidence that the data is not stationary and thus some transformation will need to be applied to make the data stationary. The first method to explore will be the simple difference as this can be included in the SARIMA model which will make the forecasting simple. The results of this difference are shown below.

Chart, line chart

Description automatically generated

This appears to be a better agreement to stationary and the augmented Dickey-Fuller test provides a p-value of ~1.8x10-10 providing strong evidence that the data is stationary. Now that the data has been transformed to stationary through the first order difference the ACF and PACF can be used to help determine the order of the SARIMA model.

Chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

From the ACF and PACF plots there is not strong evidence for a seasonal component, however we can identify a few possible models to try IMA(1,2), IMA(1,1), and ARIMA(1,1,1), 5 Data points were reserved from the end of the dataset for the purpose of forecasting. There is some noise that is appearing at the end of the ACF and PACF plots that could be indicating some seasonality from the data. To get a better analysis of this more lags should be plotted, however as a limitation to the data set it is only available to calculate the first 16 lags. If more data is collected there could potentially be an emergence of a seasonal pattern in the data.

Table

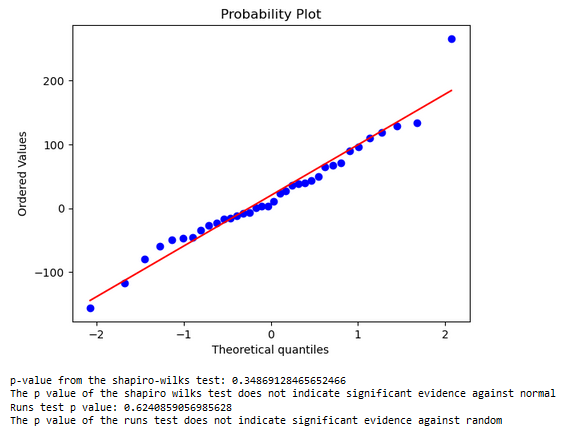
Description automatically generated

Table

Description automatically generated Table

Description automatically generated

From the fitting report the best candidate model in terms of AIC and BIC will be the ARIMA(0,1,2) model. Moving forward with this model a residual analysis is performed to determine if the model is reasonable for this dataset. The residual analysis focuses on checking residuals agreement to normal distribution, runs test for random generation, and Ljung-Box test for uncaptured dependency. The results of the residuals analysis are shown below.

 A picture containing table

Description automatically generated

From the residual analysis it has been determined that the data is coming from normal distribution, is randomly generated, and has no serial autocorrelation as determined by the Ljung-Box test. All of these facts indicate that the model is reasonable for the data and we can procede with forecasting the remaining five months in the dataset.

Chart, line chart

Description automatically generated

We can see that the forecasting does predict the first two datapoints with reasonable accuracy before settling on a constant value with an expanding confidence interval which is as to be expected from a second order non-seasonal model.

**Shampoo Sales Model Conclusions**

After transforming the data set and analyzing the ACF and PACF three models were proposed and fit to the data set. During the fitting a clear best candidate was found in the ARIMA(0,1,2) model. The residuals were analyzed for the candidate model, the results showed that the residuals were coming from normal, however the runs test showed they were coming from random generation and the Ljung-Box test showed that the residuals were uncorrelated indicating that there was no uncaptured dependency in the dataset from the model. Interpreting the results of the residual analysis indicates that the model may be a reasonable choice for the dataset, thus forecasting was performed with the model on the five points that were held back from fitting. The results of this forecasting provided reasonable estimate for the first two unseen datapoints before settling on estimating a constant average with an expanding confidence interval which can be expected from a second order integrated moving average model.

In the context of the problem this model selection will allow the business to predict the next two month of shampoo sales volume with a good degree of certainty, beyond the next two months the degree of certainty decreases as the confidence interval of the prediction grows. This could be used to get an estimate of how much shampoo will need to be produced and how many long lead-time parts may need to be ordered which can reduce any disruption to the supply chain of the business without carrying large amounts of excess inventory. This could allow the business to run its manufacturing process in a leaner fashion delivering some costs savings.

**Sunspot Data Set Analysis**

The first step in the Box-Jenkins approach will be to check if data is stationary and if not to make it stationary through some transformation. We can test for stationarity using the augmented Dickey-Fuller test. After applying the augmented Dickey-Fuller test, the resulting p-value is ~2.33x10-16. This indicates very strong evidence that the data is stationary and thus no treatment will be required before attempting to identify a model for the data. This is where some problems start to arise in the analysis of the data. When viewing the ACF and PACF plots it is very difficult to interpret the results due to the overwhelming number of lags available. The ACF and PACF are shown below.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

The ACF and PACF plots do not yield a strong recommendation for a model. The relatively high number of observations results in the PACF having a small confidence interval. The small confidence interval results in many of the lags appearing to be significant.

To address this issue the data was aggregated yearly rather than monthly. Aggregating the data should allow for more interpretable results from the ACF and PACF which will point towards some better model suggestions. The plot below shows the yearly average count of sunspots between 1749 and 1983.

Graphical user interface, chart, line chart

Description automatically generated

The data set retains its seasonal properties after the yearly averaging. However, it would be wise to check again that the data is stationary using the augmented Dickey-Fuller test. Applying the augmented Dickey-Fuller test to the yearly aggregated data returns a p-value of ~0.15. The p-value does not show strong evidence that the dataset is stationary. To treat this, it would be most simple to start with order one differencing as that can be included in the SARIMA model making forecasting simple. The time series plot of the first difference is shown below.

Chart, line chart

Description automatically generated

Again, checking the results of the augmented Dickey-Fuller test on the differenced data yielded a p-value of ~2.46x10-22 . This p-value gives a strong indication that the data set is now stationary. The differenced data will now need to be viewed in the ACF and PACF to determine the proper order of the SARIMA model with first order differencing. Displayed below are the ACF and PACF plots for the differenced data set.

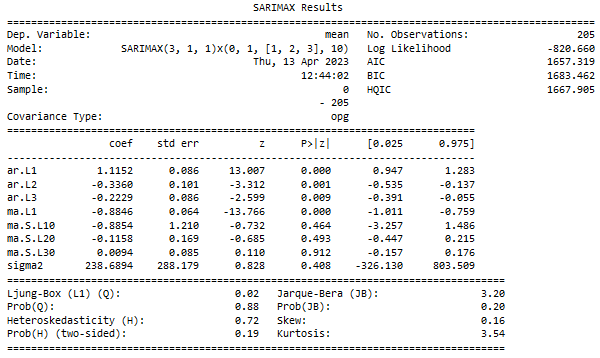
Chart

Description automatically generated

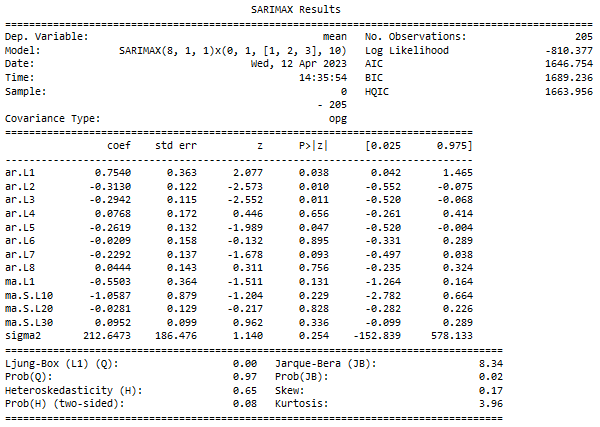
Chart, histogram

Description automatically generated

Initial analysis of the ACF for the aggregated first difference data indicated that a 5-year period was appropriate. Upon further investigation it was concluded that a 10-year period fit the data better. The PACF plot suggests using a trend autoregressive order of either 3 or 8. Using this information, a number of models were proposed including SARIMA(3,1,1)x(0,1,2)10 , SARIMA(3,1,1)x(0,1,3)10, SARIMA(8,1,1)x(0,1,2)10, and SARIMA(8,1,1)x(0,1,3)10. Before fitting these models 30 points were held back from the end for the purpose of evaluating the forecasting potential of the models on un-seen data. In order to evaluate the forecasting potential of the model, the most recent 30 data points were not included in the training data. The fit summary of the 4 models is shown below.

Table

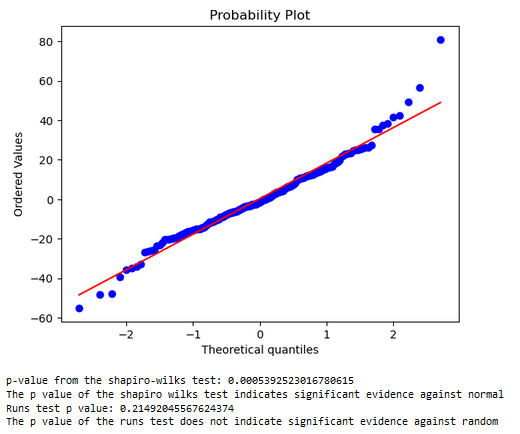
Description automatically generated

Table

Description automatically generated

From these results it can be seen that the non-seasonal component plays a more significant role than the seasonal component as the change in the seasonal autoregressive order was much less impactuful than a change in the trend autoregression order . For the seasonal component, lower AIC and BIC scores are given by using only two moving average components. For the non-seasonal component, lower AIC and BIC scores are given by eight autoregressive. For the two best candidate models, SARIMA(3,1,1)x(0,1,2)10 and SARIMA(8,1,1)x(0,1,2)10, residual analysis is performed. The residual analysis is focused on analysing the distributions normality, the runs test for randomness, and the Ljung-Box test for uncaptured dependancy.

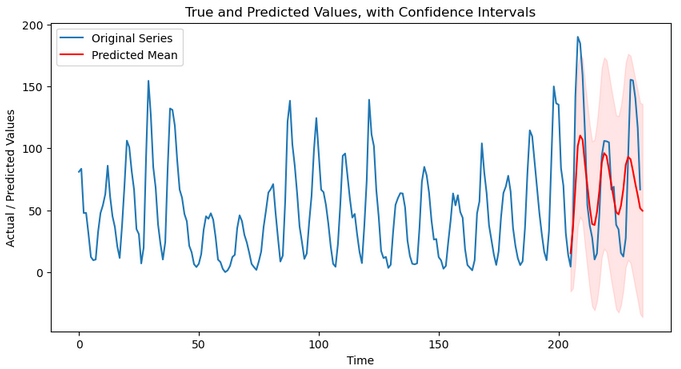
SARIMA(3,1,1)x(0,1,2)10

 Text

Description automatically generated with medium confidenceChart, line chart

Description automatically generated

From the residual analysis it is found that the residuals do not conform to normal however they are randomly generated and do not show any uncaptured dependency in the Ljung-Box test indicating that this could be a reasonable model choice for this dataset. Following the residual analysis, forecasting of the last 30 datapoints that were held back was performed with the results deisplayed below.



As can be seen the model is doing a reasonable job of capturing the peaks and dips in the data and as time goes forward is collapsing toward the mean of the dataset. However, it should be noted that the predicted values cover a range of thirty years.

SARIMA(8,1,1)x(0,1,2)10

Chart, line chart

Description automatically generated Text, table

Description automatically generatedChart, line chart

Description automatically generated

The residual analysis shows similar results as the 3rd order SARIMA model explored previously, one thing of note here is that the p-value of the runs test is much farther from the significance level of 0.05. This provides confidence that the 8th order SARIMA model is a good fit for the data. Again, the model was evaluated by forecasting on the thirty point test data set. The results are displayed below.

Chart, line chart

Description automatically generated

The forecasting in this model has much better agreement with the testing data set and is able to more accurately predict the peaks and valleys in the time series. This indicates that the model may be better than the simpler SARIMA model. The forecasted values of the 8th order SARIMA model is better than that of the 3rd order model. Not only does the 8th order SARIMA model more accurately predict the location of the peaks and troughs of the time series, it also achieves better results in terms of magnitude of those peaks and troughs.

**Sunspot Model Conclusions**

After transforming the data set and analyzing the ACF and PACF, four models were proposed and fit to the data set. After fitting the four models the two best candidates were selected, SARIMA(3,1,1)x(0,1,2)10 and SARIMA(8,1,1)x(0,1,2)10. The SARIMA(3,1,1)x(0,1,2)10 model had a lower BIC but higher AIC than the SARIMA(8,1,1)x(0,1,2)10 model. The residuals were analyzed for both candidate models and the results were similar for both models. These results showed that the residuals were not coming from normal, however the runs test showed they were coming from random generation and the Ljung-Box test showed that the residuals were uncorrelated indicating that there was no uncaptured dependency in the dataset from either model. Interpreting the results of the residual analysis indicates that both of the models may be a reasonable choice for the dataset, thus forecasting was performed with both models on thirty points that were held back from fitting. The results of this forecasting showed that the more complex SARIMA(8,1,1)x(0,1,2)10 had a better fit to the unseen data thus would be the recommended model for forecasting the sunspot count.

Sunspots are dark patches that appear on the surface of the Sun. They are caused by intense magnetic activity in the Sun's outer layer. Sunspots typically appear in pairs, and they can range in size from a few hundred kilometers to tens of thousands of kilometers in diameter.

Sunspots are important to track because they are an indicator of solar activity. The number of sunspots on the Sun's surface varies over a roughly 11-year cycle, known as the solar cycle. During periods of high solar activity, the Sun emits more radiation and energetic particles, which can impact Earth's atmosphere and cause changes in weather patterns and disrupt power grids. The ability to know when to expect high solar activity will be helpful for giving aid, or at least allowing proper planning for things such as power outages or freak storms.

Furthermore, sunspots are associated with the production of solar flares and coronal mass ejections, which can have a significant impact on Earth's magnetic field and lead to geomagnetic storms. These storms can create auroras and interfere with radio communication and satellite navigation systems like GPS. GPS is vital to many, from regular citizens driving to flying aircraft. If a solar flare were to occur and knock out GPS, it could have dire consequences.

With the model selected there is good certainty of the number of sunspots that could be expected to occur for a 30-year period. This could be utilized by a designer of a spacecraft as sunspots emit a large amount of electromagnetic radiation that can damage electronics without proper shielding. Having a sense of the volume of sunspots that would be observed over a long period that the spacecraft could be operational in would allow the designer to determine how much shielding may be required for the craft.

**References**

https://github.com/jbrownlee

<https://www.weather.gov/fsd/sunspots>

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI:

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

<https://spaceplace.nasa.gov/solar-cycles/en/>

<https://stardate.org/astro-guide/sunspots-and-solar-flares>