

IEE 579 – Time Series Analysis and Forecasting

Final Project

Time Series Analysis of Daily Weather



Instructor: Dr. Douglas C Montgomery

May 01, 2020

Arizona State University

Prakash Sudhakar – 1217272901

Ragul Subramanian - 1215127914

Table of Contents

PRELIMINARY ANALYSIS:.....	3
CHECKING FOR MISSING VALUES:	3
DECOMPOSITION OF DATA:.....	3
SMOOTHING MODELS:	4
SIMPLE MOVING AVERAGE (SMA):.....	4
SIMPLE EXPONENTIAL SMOOTHING (SES):.....	6
DOUBLE(BROWN) EXPONENTIAL SMOOTHING:.....	7
LINEAR (HOLT) EXPONENTIAL SMOOTHING:	9
WINTERS METHOD (ADDITIVE):.....	10
DAMPED TREND LINEAR EXPONENTIAL SMOOTHING:.....	11
SEASONAL EXPONENTIAL SMOOTHING (12, ZERO TO ONE):	12
ARIMA Models	13
FITTING THE ARIMA MODEL	14
MODEL 1 - ARIMA (2,1,2).....	14
MODEL 2 - SEASONAL ARIMA (2,1,2) (0,0,1) 12	15
MODEL 3 - SEASONAL ARIMA (2,1,2) (0,0,2) 12	16
MODELLING WITH PROPHET IN R:.....	17
CONCLUSION.....	18

PRELIMINARY ANALYSIS:

CHECKING FOR MISSING VALUES:

Data Cleaning is an important process for examining the existing data to check for potential errors, missing values and inconsistencies present in the data. The Date column is sorted in ascending order to check for missing values in the data and then proceed with analysing the time series model.

Explore Missing Values

Commands

Missing Value Report

Number of missing values for each column

Missing Value Clustering

Hierarchical clustering of rows and columns missingness

Missing Value Snapshot

Patterns of missing values with graphical map

Multivariate Normal Imputation

Least squares prediction from the nonmissing variables in each row

Multivariate SVD Imputation

Imputation for wide problems using a singular value decomposition with the power-method adapted for missing values

Automated Data Imputation

Automatically selects best dimension for low-rank approximation based on the data and has streaming imputation capabilities

Automated Data Imputation Controls**Missing Columns**☐ Show only columns with missing

Close

Select columns and choose an action.

Select Rows

Color Cells

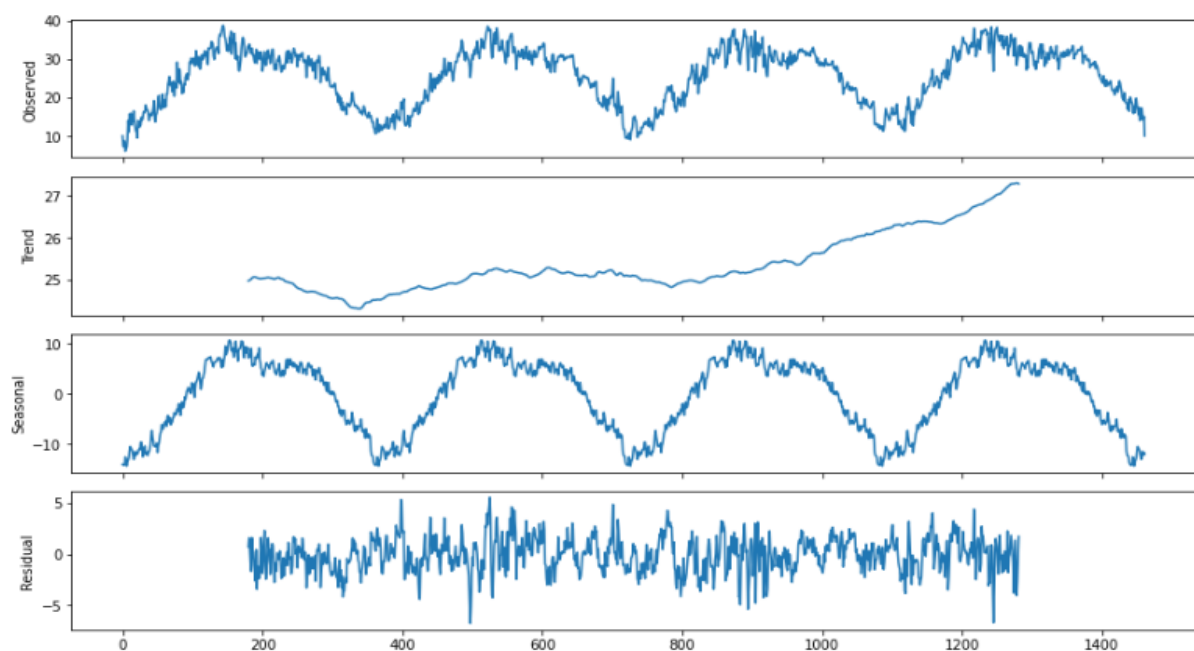
Exclude Rows

Color Rows

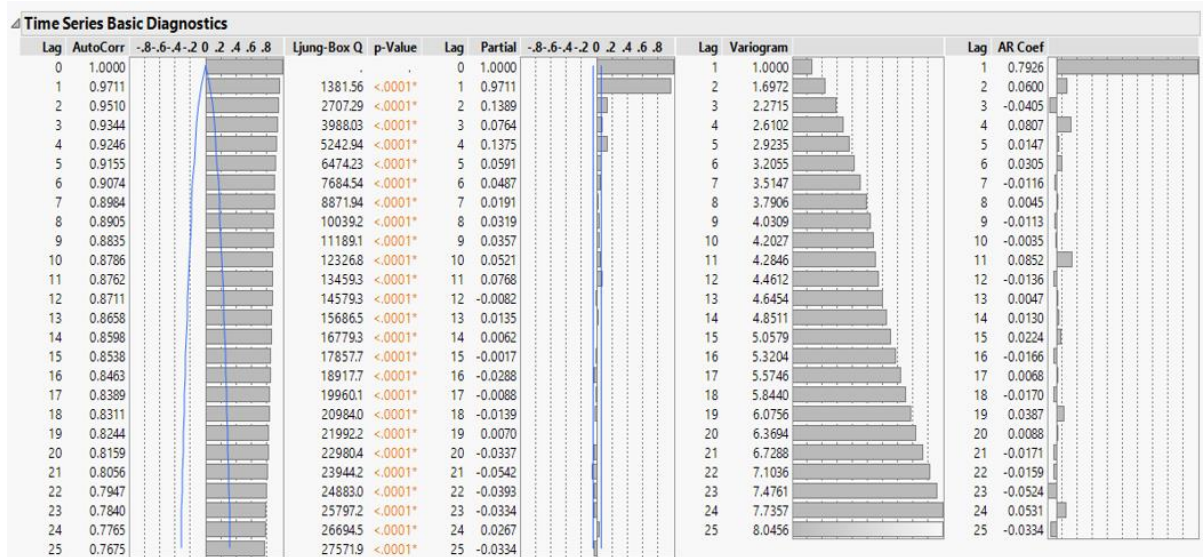
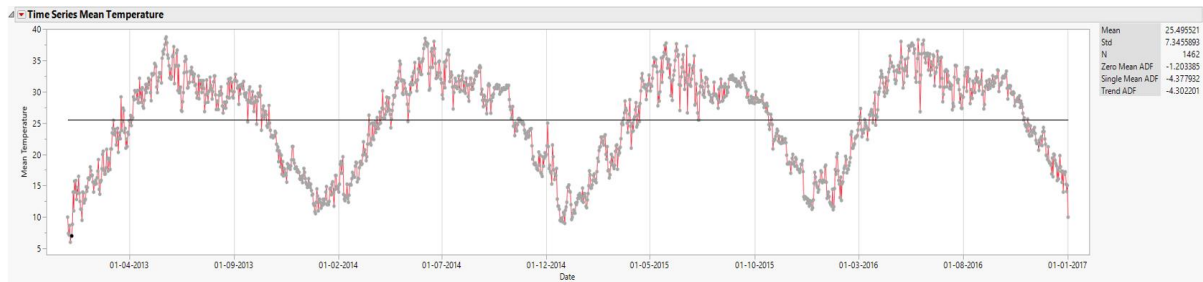
DECOMPOSITION OF DATA:

The below diagram represents an additive model decomposition of a time series into overall:

- [1] Trend
- [2] Seasonality
- [3] Residual/ Noise



From the decomposed data, it can be inferred that there is a slight linear trend with a positive slope in the data along with a seasonal component which is repeated after every 12 periods.



The ACF and PACF plots for the mean temperature are shown above. In the ACF plot, the time series is non-stationary since it doesn't dampen with 25 lags. The above-mentioned statement is also supported by the Variogram as the decay is very slow. From the PACF plot, it can be a potential AR (2) model as there is a cut off at lag 2.

SMOOTHING MODELS:

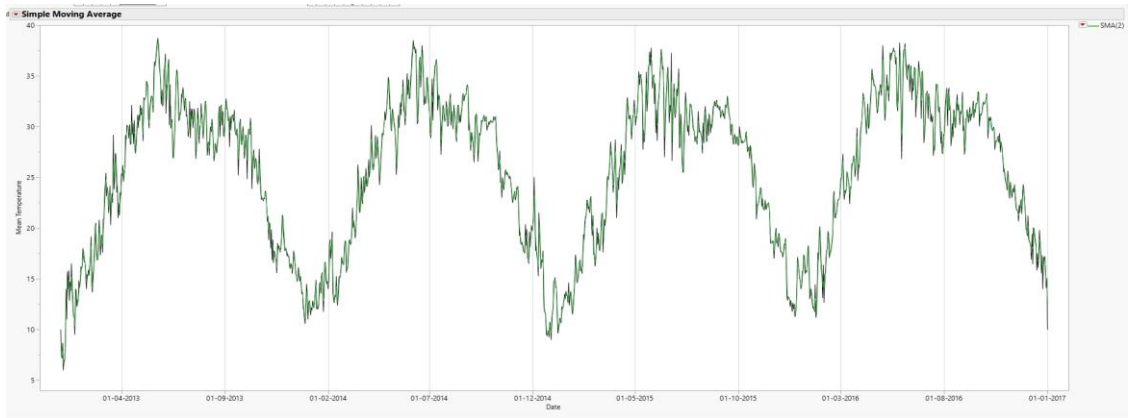
SIMPLE MOVING AVERAGE (SMA):

Simple moving average (SMA) is the base model used to calculate the moving average where we assume the future values of mean temperature depends on the average of its K previous values. It is used to smooth the trend in the data to get meaningful insights from the data.

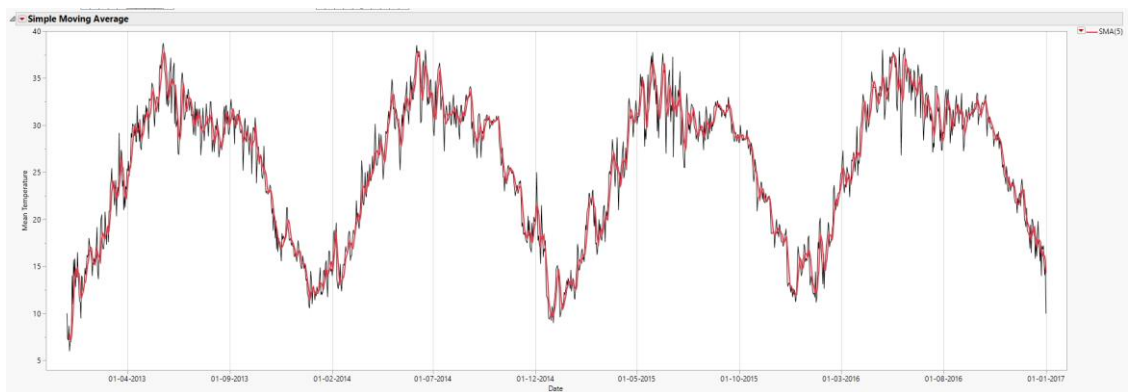
$$\hat{y}_t = \frac{1}{k} \sum_{n=1}^k y_{t-n}$$

The wider the window, the smoother the trend. We use three SMA models:

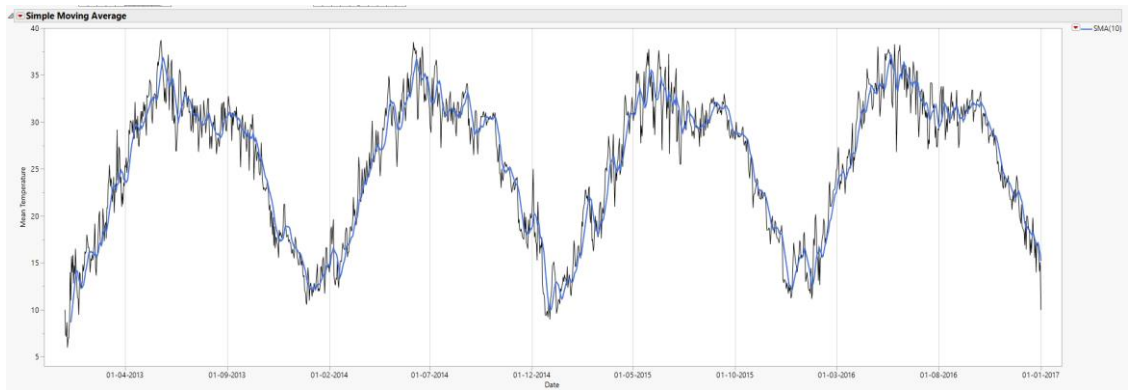
- [1] SMA of Order 2 – SMA [2]
- [2] SMA of Order 5 – SMA [5]
- [3] SMA of Order 10 – SMA [10]



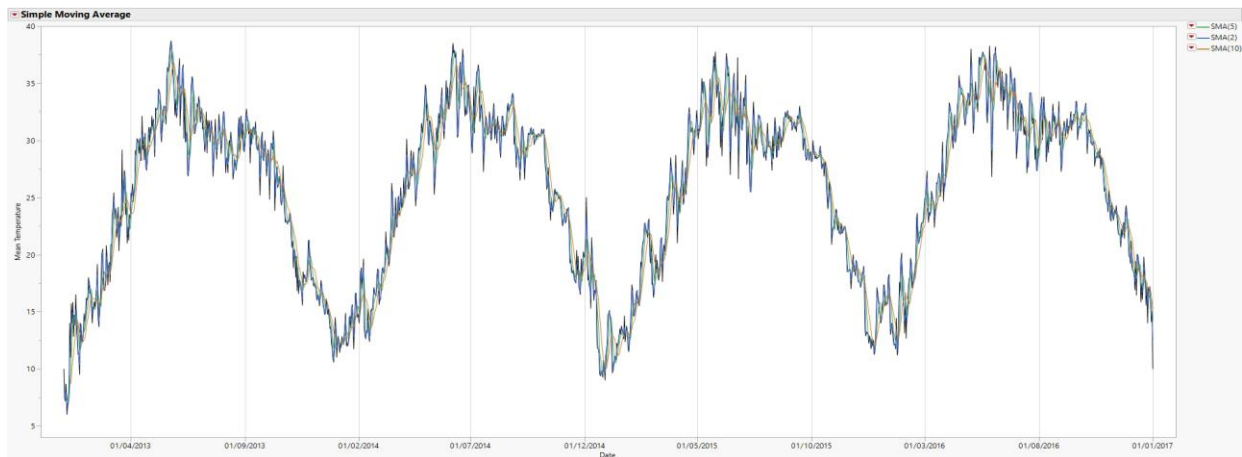
SMA of Order 2 – SMA [2]



SMA of Order 5 – SMA [5]



SMA of Order 10 – SMA [10]



Combined SMA Model

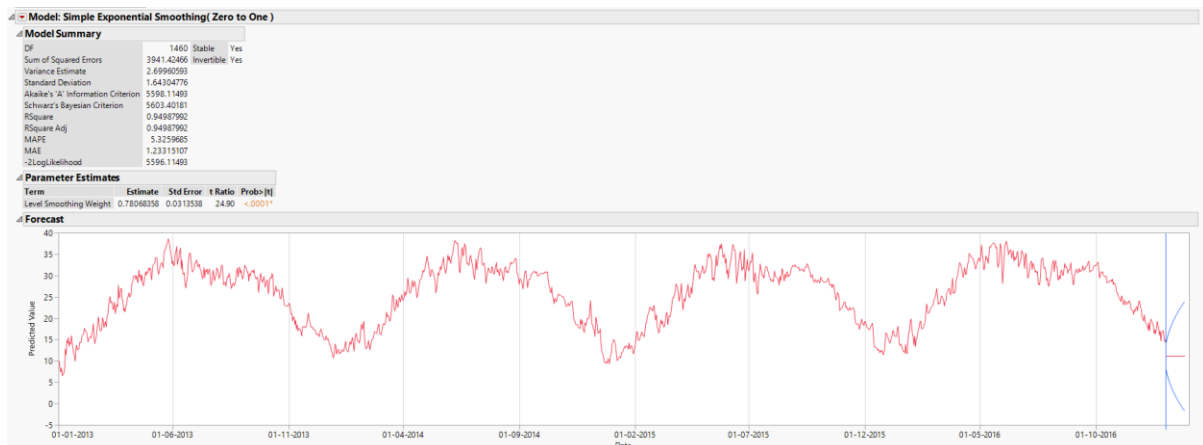
Smaller the values of N in MA, more variance and the model can easily adapt to changes in the graph. Whereas larger values tend to uneven smooth the model and cannot be subjected to an immediate change in the model.

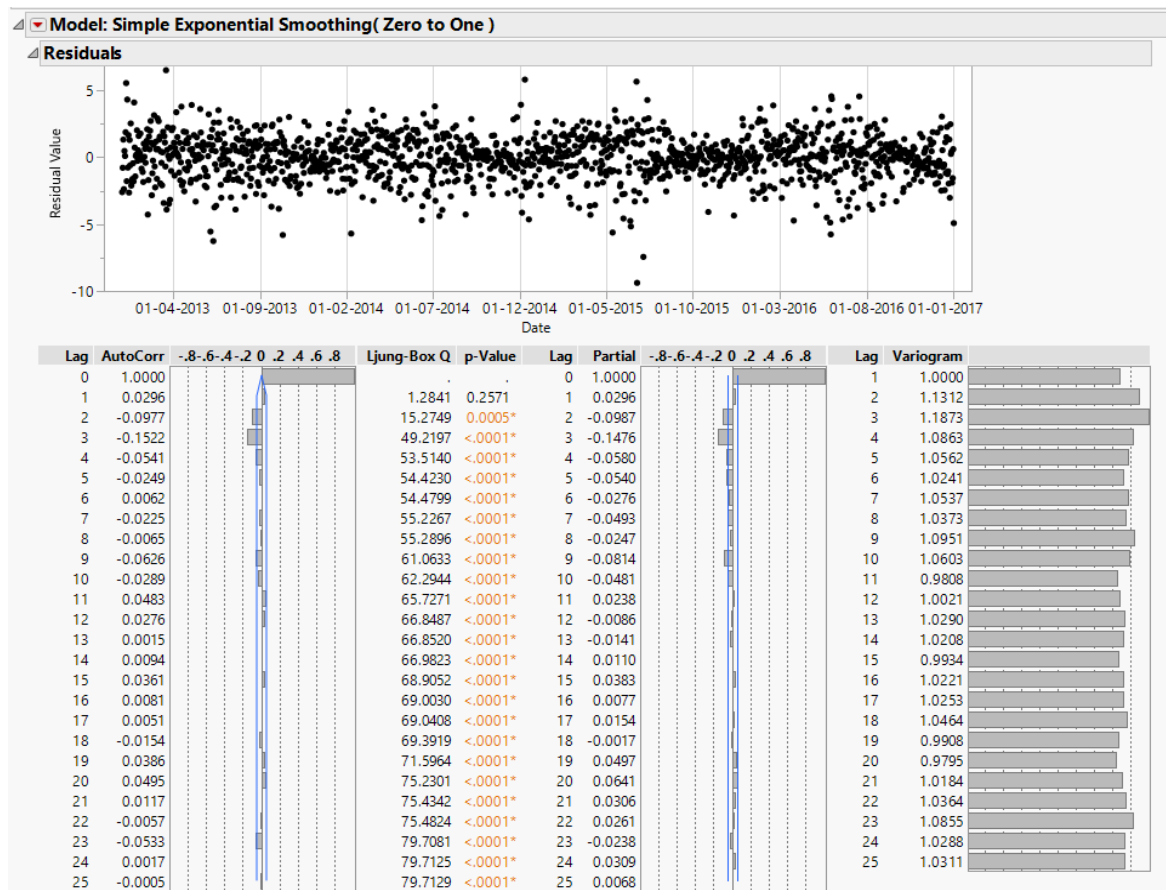
SIMPLE EXPONENTIAL SMOOTHING (SES):

Simple Exponential Smoothing is a technique where we start weighing all available observations with an exponential decreasing weight as we move away from time. The prediction will be a weighted linear sum of the recent past observations.

$$\hat{y}_t = \alpha \cdot y_t + (1 - \alpha) \cdot \hat{y}_{t-1}$$

Here, α refers to the smoothing factor which tells us how quickly we forget our past data. Small values of α reflects more weightage given to last k observations.





From the above figure, lags 1 and 3 are significant in both the ACF and PACF plots resulting in a stationary process.

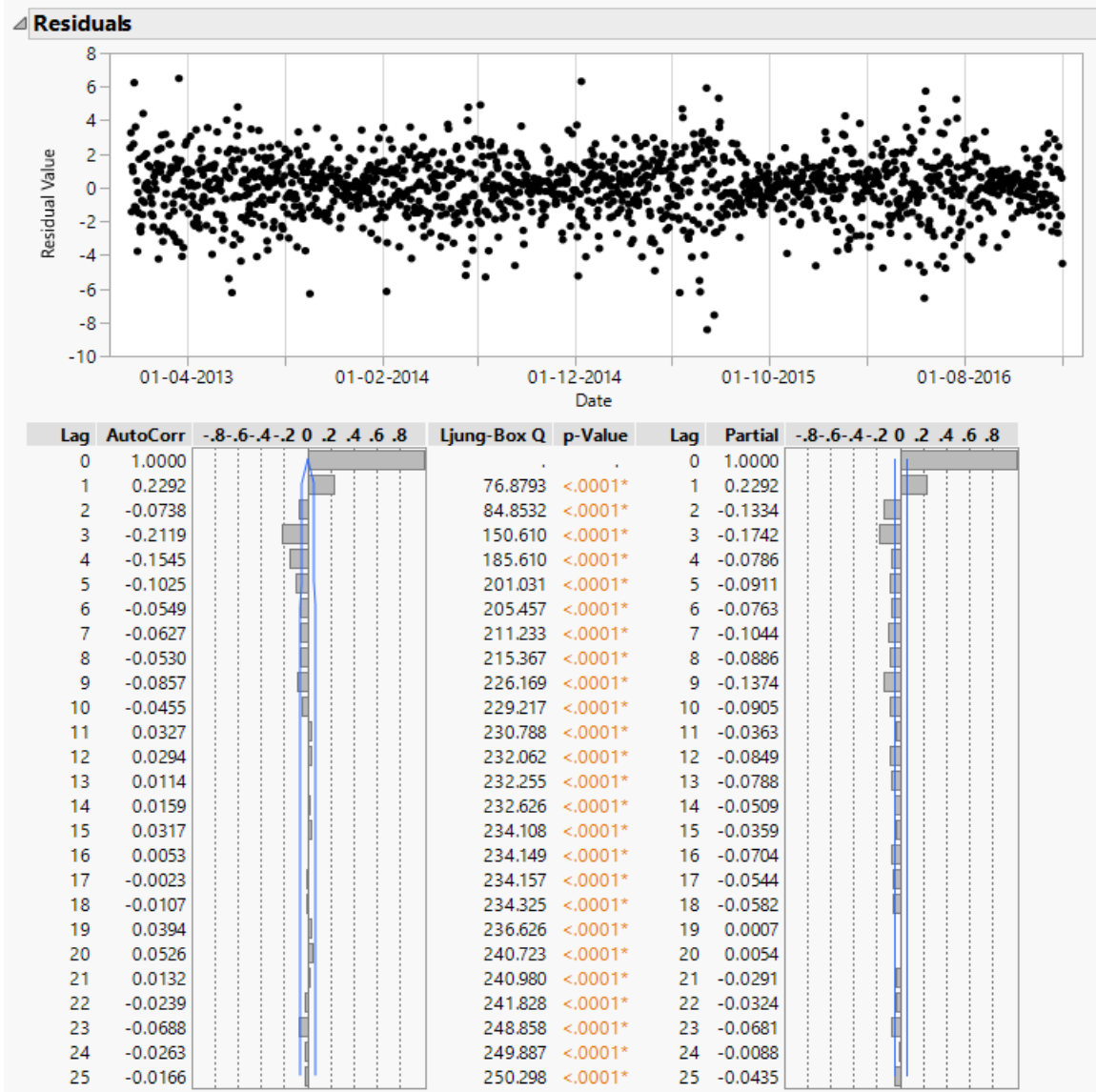
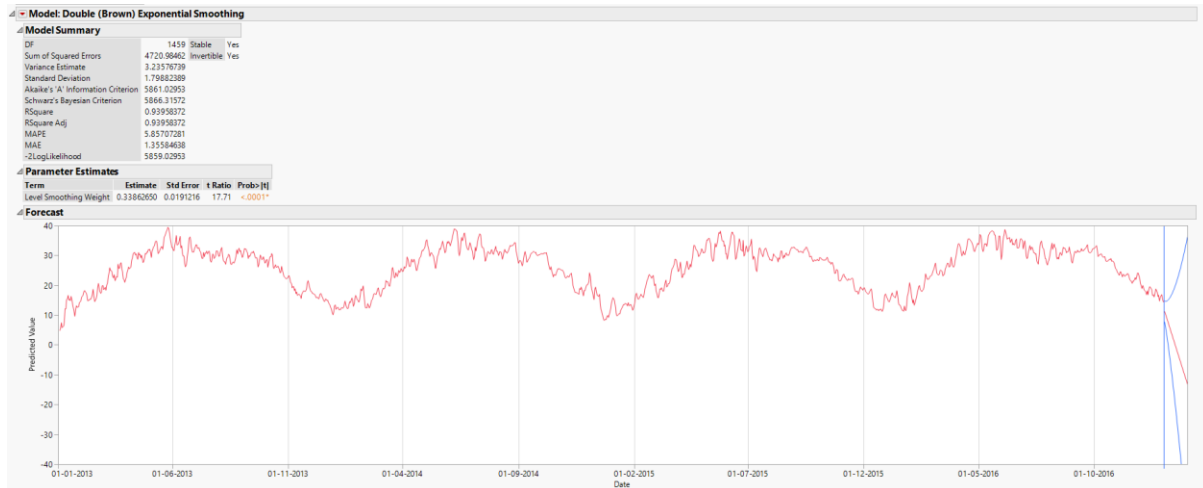
DOUBLE(BROWN) EXPONENTIAL SMOOTHING:

Brown's Exponential Smoothing is a similar concept where the model performs two simple exponential smoothing forecasts when there is a linear trend over time. This is executed without adding any additional parameters to the model.

$$S_t = \alpha y_t + (1 - \alpha) (S_{t-1} + b_{t-1})$$

$$b_t = \gamma (S_t - S_{t-1}) + (1 - \gamma) b_{t-1}$$

The first term represents the intercept whereas the second term gives us the previous values of level and trend. The level smoothing weight for Brown's model is assigned to be 0.3386.



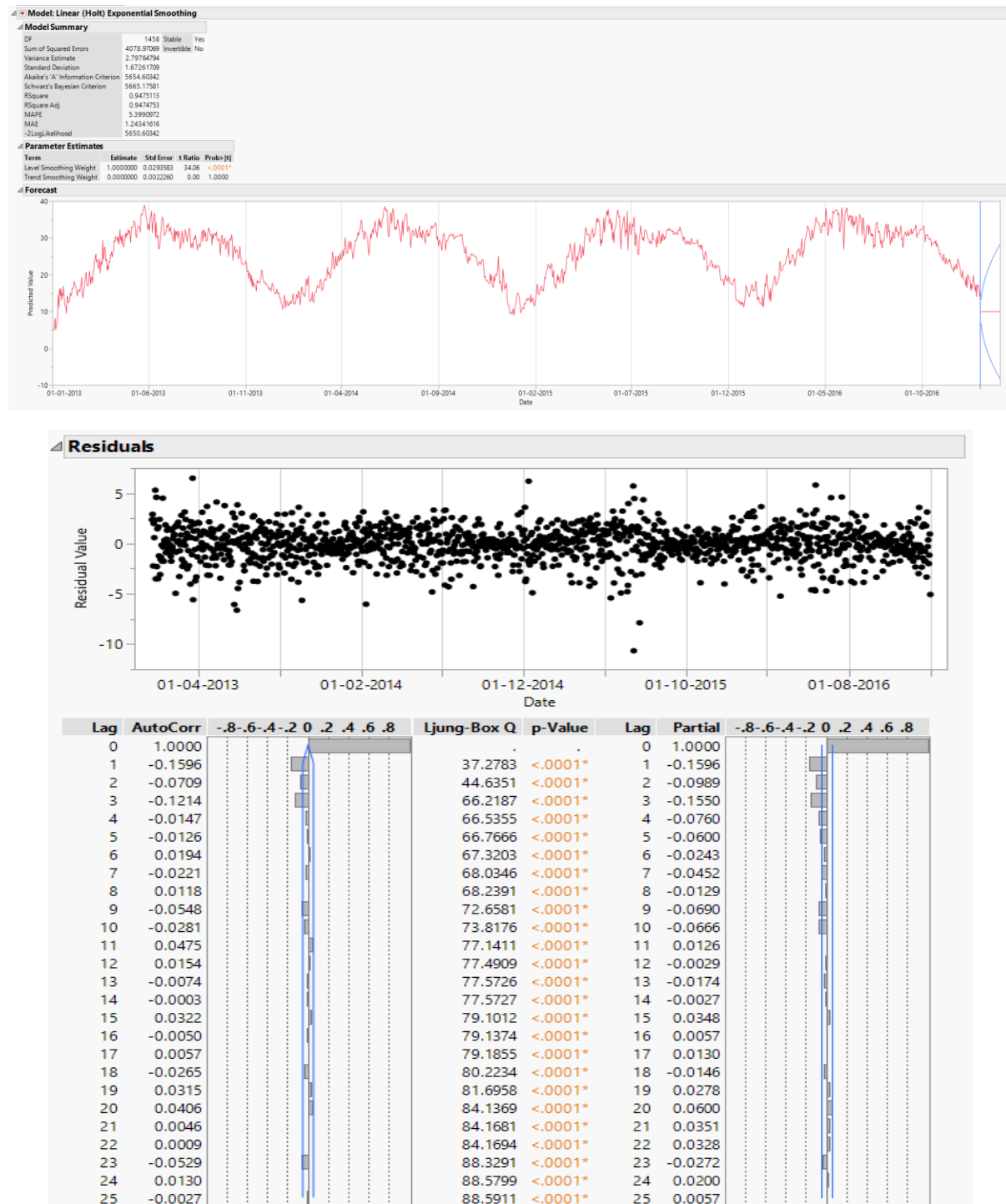
Here again from the above figure lags 1 and 3 are significant both in ACF and PACF plots resulting a stationary process.

LINEAR (HOLT) EXPONENTIAL SMOOTHING:

Holt's linear is used to forecast a model with a linear trend. Linear method uses level and seasonal factor to model the time series. The default parameter estimates selected by JMP is given below:

- [1] Level factor – 1
- [2] Seasonal factor – 0

The seasonal parameter is set is zero as it only considers the linear trend present in the data.

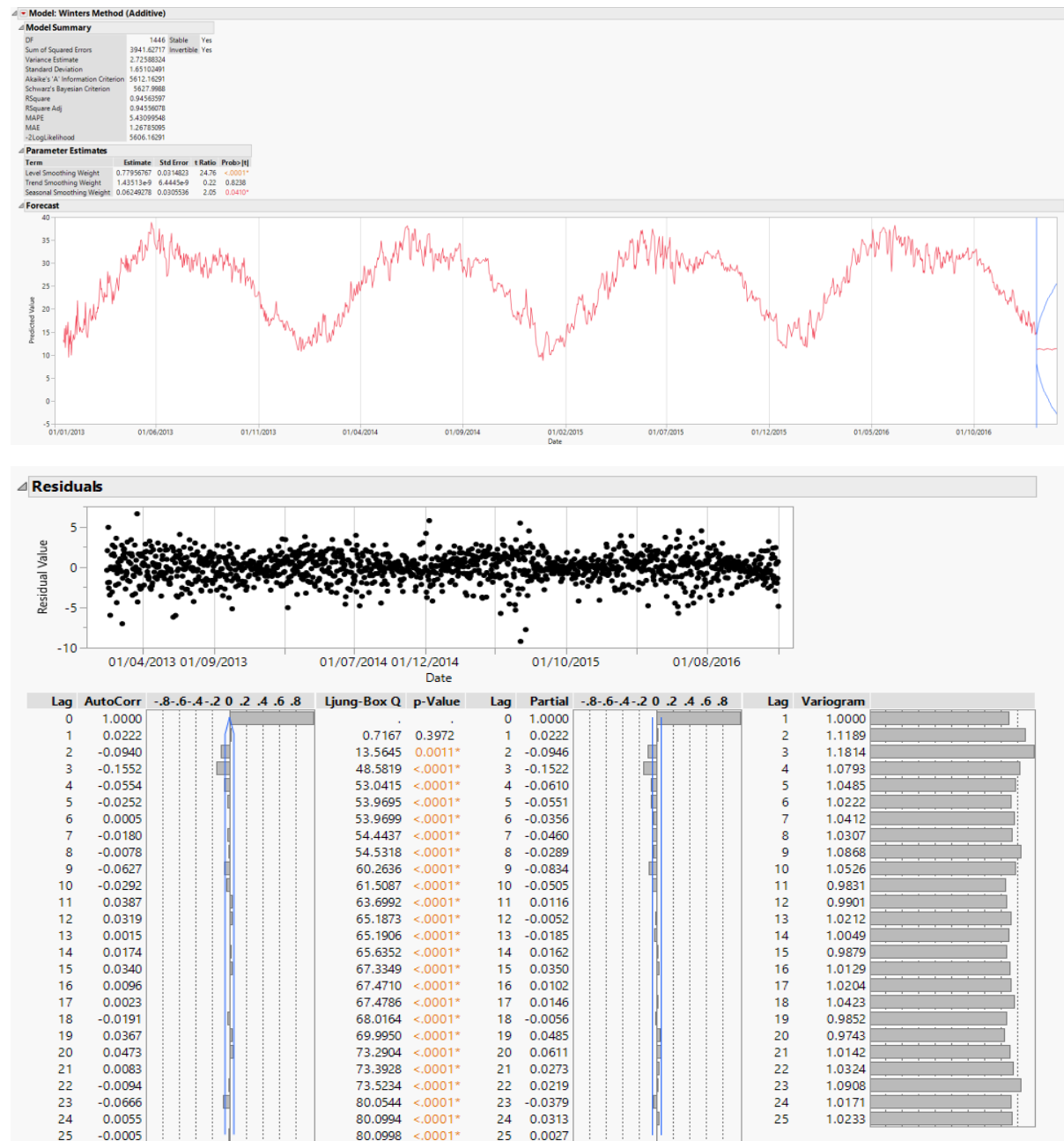


WINTERS METHOD (ADDITIVE):

Holt Winters are used to make short-term forecasts in an additive or multiplicative model with an increasing or decreasing trend. The magnitude of the seasonal variation remains constant, so an additive model is ideal. By default, JMP selects the best exponential smoothing parameters to model the data.

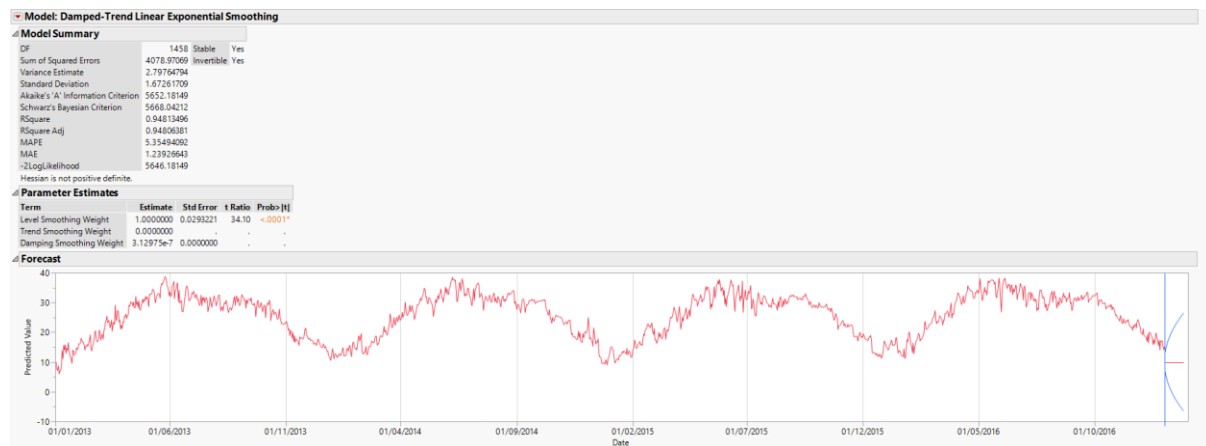
There are three exponential smoothing parameters being used here:

- [1] Smoothing of level factor – 0.779
- [2] Smoothing of trend factor – 1.435
- [3] Smoothing of seasonal index – 0.062

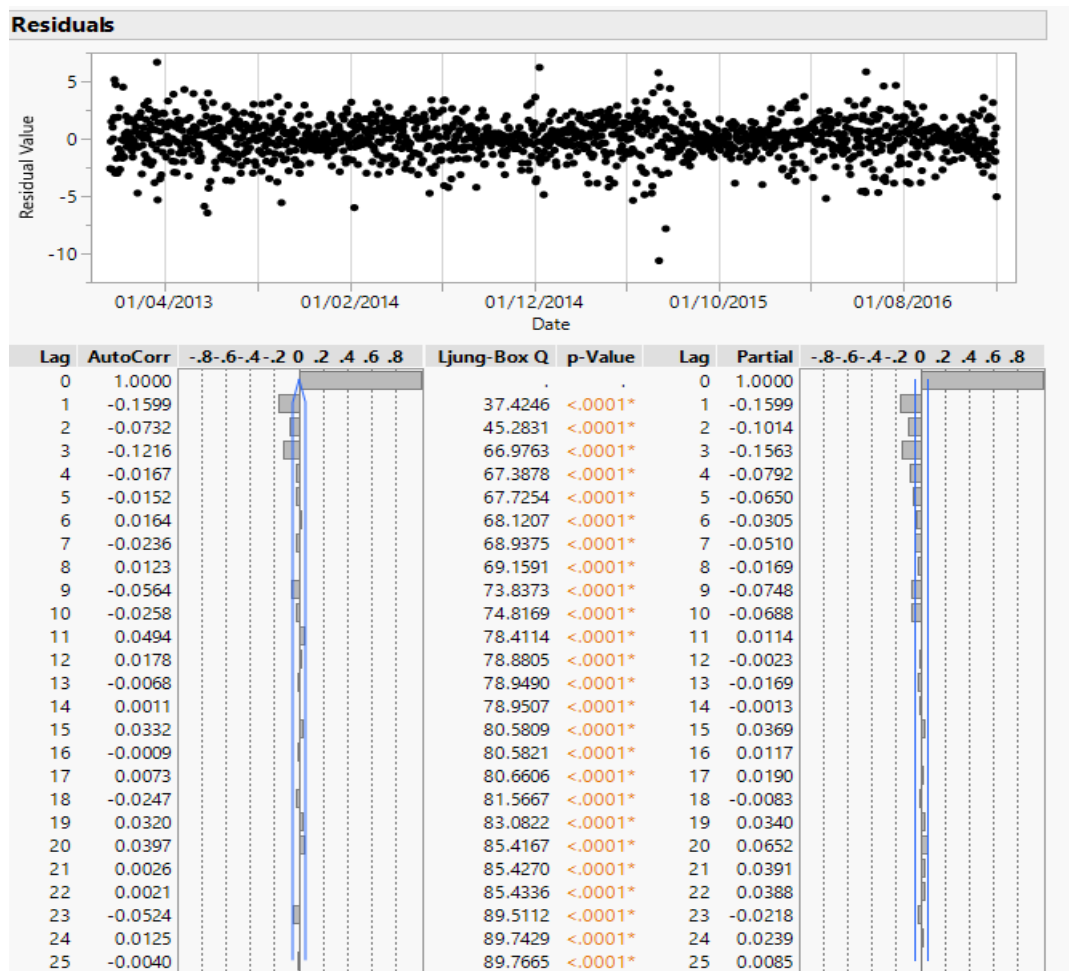


Except for a couple of Lags, every lag seems to lie within the confidence band. The model seems to be adequate.

DAMPED TREND LINEAR EXPONENTIAL SMOOTHING:

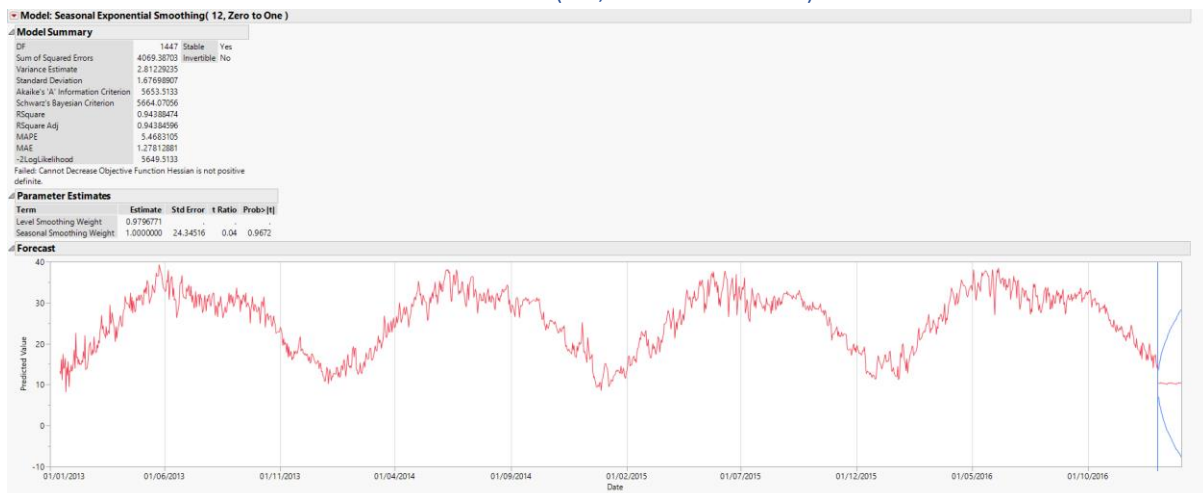


Damped Trend model takes only the Level term into consideration just like Linear(Holt) exponential smoothing. The level parameter estimate are significant which is evident from the above graph. The forecast produced by this model does not seem to be convincing where it doesn't follow any seasonal trend like the past data.

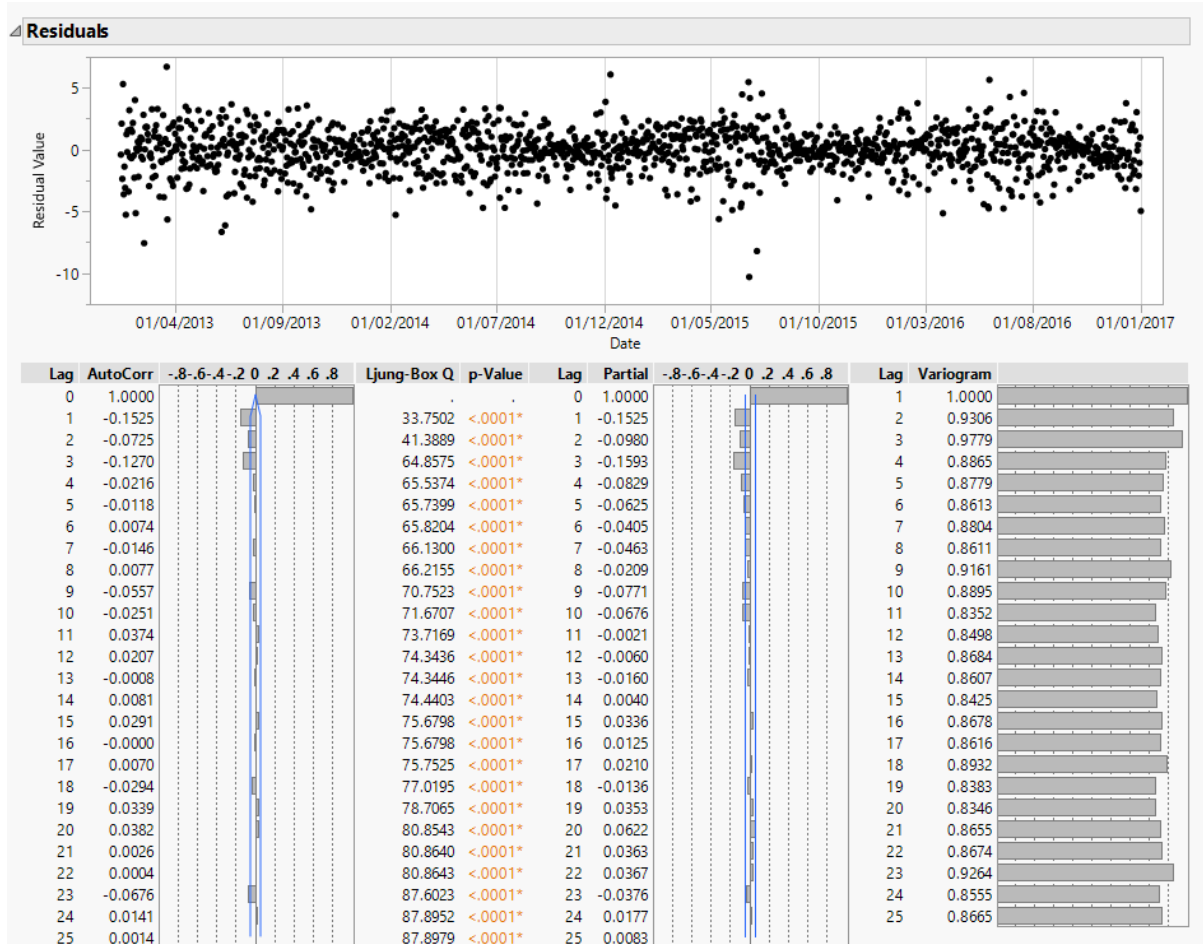


The lags cut off at 1 both in ACF and PACF plots which is again shows a slightly stationary process. All the auto correlation values seems to be statistically significant. There are some lags at which the model exceeds the reference stationary line.

SEASONAL EXPONENTIAL SMOOTHING (12, ZERO TO ONE):









In Seasonal Exponential smoothing, both the level and seasonal terms are being taken into account to model the data. It can also be seen from the above modelling that the seasonal smoothing weight is not statistically significant. Seasonal term is being given more weightage which can be interpreted that more weights are given to the last observations.



The lags cut off at 1 both in ACF and PACF plots which is again shows a stationary process. There are some lags at which the auto correlation values are above the significance border line while still the variogram depicts stationarity.

The below model comparison chart shows us the results of the different models used so far. The MAPE and MAE values are pretty low which indicates these models have good forecast accuracy.

Model Comparison

Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2	.4	.6	.8	MAPE	MAE
<input checked="" type="checkbox"/>	<input type="checkbox"/>	 Simple Exponential Smoothing(Zero to One)	1460	2.6996059	5598.1149	5603.4018	0.950	5596.1149	0.999111					5.325968	1.233151
<input checked="" type="checkbox"/>	<input type="checkbox"/>	 Winters Method (Additive)	1446	2.7258832	5612.1629	5627.9988	0.946	5606.1629	0.000889					5.430995	1.267851
<input checked="" type="checkbox"/>	<input type="checkbox"/>	 Damped-Trend Linear Exponential Smoothing	1458	2.7976479	5652.1815	5668.0421	0.948	5646.1815	0.000000					5.354941	1.239266
<input checked="" type="checkbox"/>	<input type="checkbox"/>	 Seasonal Exponential Smoothing(12, Zero to One)	1447	2.8122923	5653.5133	5664.0706	0.944	5649.5133	0.000000					5.468310	1.278129
<input checked="" type="checkbox"/>	<input type="checkbox"/>	 Linear (Holt) Exponential Smoothing	1458	2.7976479	5654.6034	5665.1758	0.948	5650.6034	0.000000					5.399097	1.243416
<input checked="" type="checkbox"/>	<input type="checkbox"/>	 Double (Brown) Exponential Smoothing	1459	3.2357674	5861.0295	5866.3157	0.940	5859.0295	0.000000					5.857073	1.355846

ARIMA Models

An ARIMA model is given by:

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t$$

Where,

$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$ (Autoregressive parameter)

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$ (Moving average parameter)

ε_t = white noise or error term

d = differencing term

B = Backshift operator i.e. $B^a Y_t = Y_{t-a}$

ARIMA models are denoted with a notation ARIMA(p,d,q). These three parameters are responsible for seasonality, trend and noise generated in the data.

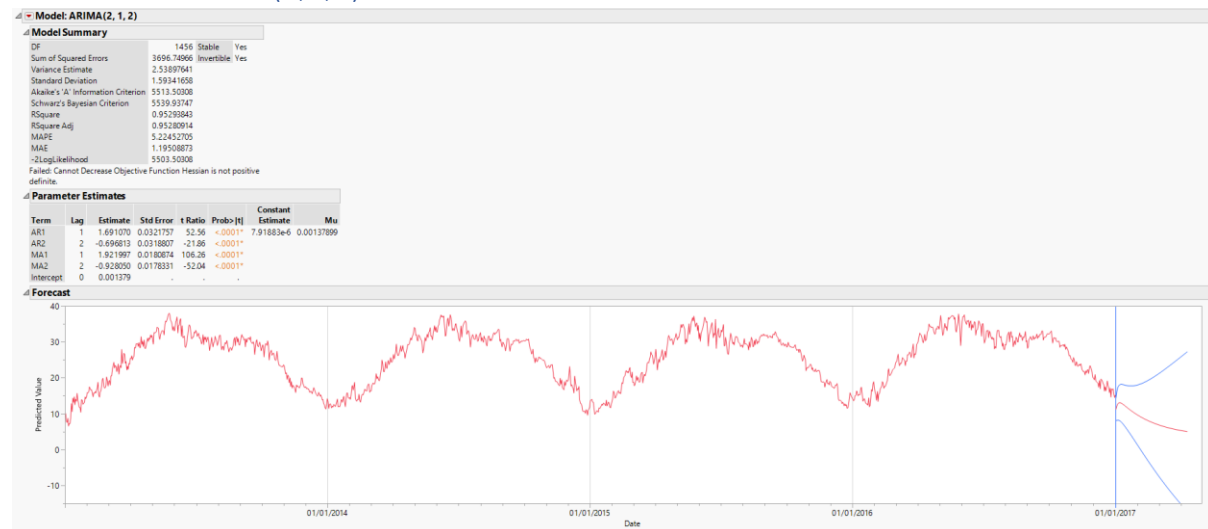
Model Comparison															
Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2	.4	.6	.8	MAPE	MAE
<input checked="" type="checkbox"/>	<input type="checkbox"/>	ARIMA(2, 1, 2)	1456	2.5389764	5513.5031	5539.9375	0.953	5503.5031	0.576433					5.224527	1.195089
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(2, 1, 2)(0, 0, 1)12	1455	2.5400754	5515.1410	5546.8623	0.953	5503.141	0.254141					5.227159	1.195368
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(2, 1, 2)(0, 0, 2)12	1454	2.5416601	5517.0495	5554.0576	0.953	5503.0495	0.097873					5.224867	1.194847
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(2, 1, 2)(2, 0, 1)12	1453	2.5431113	5518.8797	5561.1748	0.953	5502.8797	0.039194					5.223173	1.194491
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(2, 1, 2)(2, 0, 2)12	1452	2.5442367	5520.5334	5568.1153	0.953	5502.5334	0.017145					5.227520	1.195305
<input checked="" type="checkbox"/>	<input type="checkbox"/>	ARIMA(1, 2, 2)	1456	2.5558998	5522.7918	5543.9366	0.952	5514.7918	0.005543					5.265588	1.198448
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(1, 2, 2)(0, 0, 1)12	1455	2.5571588	5524.5085	5550.9395	0.952	5514.5085	0.002349					5.267999	1.198905
<input checked="" type="checkbox"/>	<input type="checkbox"/>	ARIMA(2, 2, 2)	1455	2.5572332	5524.5565	5550.9874	0.952	5514.5565	0.002294					5.261552	1.198145
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(2, 2, 2)(1, 0, 0)12	1454	2.5583782	5526.2077	5557.9249	0.952	5514.2077	0.001005					5.263598	1.198586
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(1, 2, 2)(0, 0, 2)12	1454	2.5587278	5526.4040	5558.1211	0.952	5514.404	0.000911					5.265453	1.198274
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(1, 2, 2)(1, 0, 1)12	1454	2.5587848	5526.4309	5558.1481	0.952	5514.4309	0.000898					5.269070	1.199248
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(2, 2, 2)(0, 0, 2)12	1453	2.5599024	5528.0772	5565.0805	0.952	5514.0772	0.000394					5.260711	1.197891
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(1, 2, 2)(2, 0, 1)12	1453	2.5600798	5528.1777	5565.1810	0.952	5514.1777	0.000375					5.265849	1.198281
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(1, 2, 2)(1, 0, 2)12	1453	2.5600799	5528.1777	5565.1811	0.952	5514.1777	0.000375					5.265846	1.198281
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(2, 2, 2)(1, 0, 1)12	1453	2.5602077	5528.2468	5565.2502	0.952	5514.2468	0.000362					5.264040	1.198699
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(2, 1, 2)(2, 1, 1)12	1441	2.5441456	5529.7306	5571.9596	0.949	5513.7306	0.000173					5.342970	1.235159

Here, we tend to perform parameter selection using the JMP in built ARIMA models group function where based on the inputs, the model is being fit with all possible combinations to determine the best set of parameters that yields the best performance for our model.

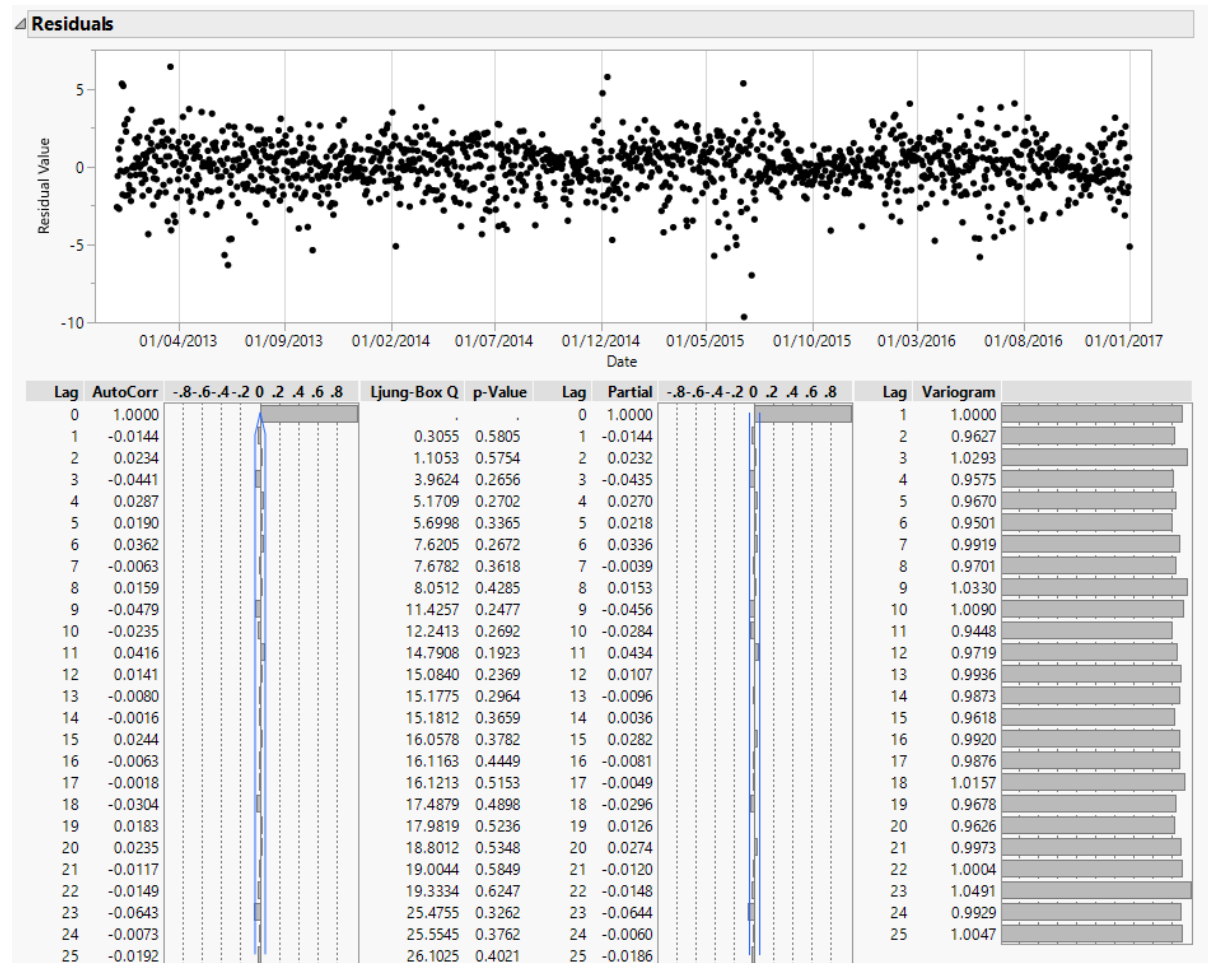
The model is being run for 729 iterations as we choose our parameter search range values from 0 to 2 for seasonal (P,D,Q) and non- seasonal (p,d,q) values. The models are chosen to be the best based on the JMP's default 'Minimum AIC' criterion. As we can see from the model comparison window, the top 3 models are chosen to inspect for further comparison as they have the least AIC and MAPE values.

FITTING THE ARIMA MODEL

MODEL 1 - ARIMA (2,1,2)

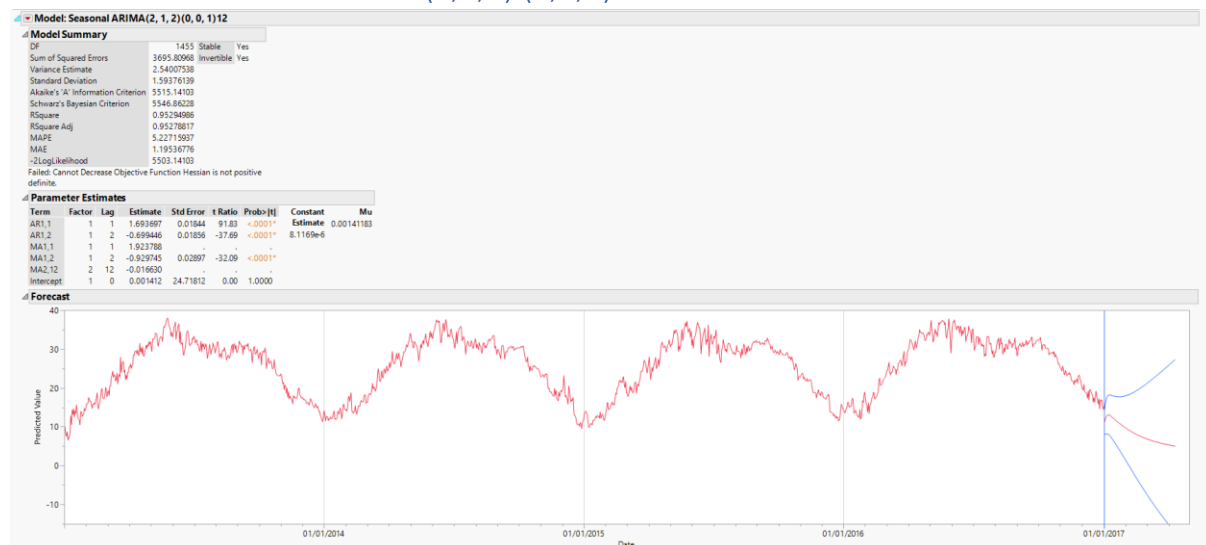


From the above figure, it is clear that all the terms, AR[1], AR[2], MA[1] and MA[2] are statistically significant. The performance metrics (i.e. MAPE and MAE) for the above model are 5.221 and 1.195 respectively which is the minimum of all the parameter search range.

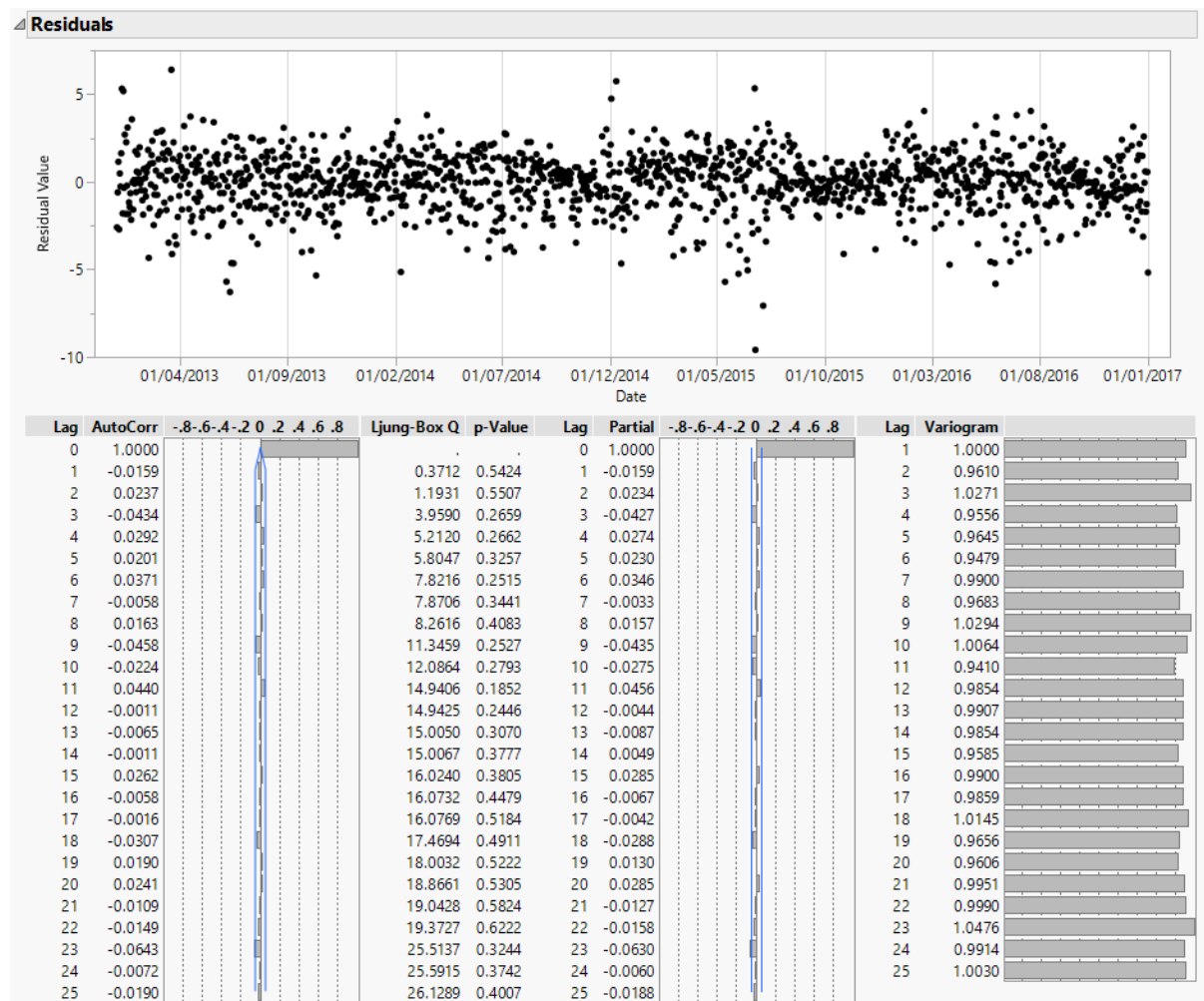


Well, it's clear that the residuals are stationary, there are no apparent autocorrelations in the model which is evident from the ACF and PACF plots which cuts off at lag 0.

MODEL 2 - SEASONAL ARIMA (2,1,2) (0,0,1) 12

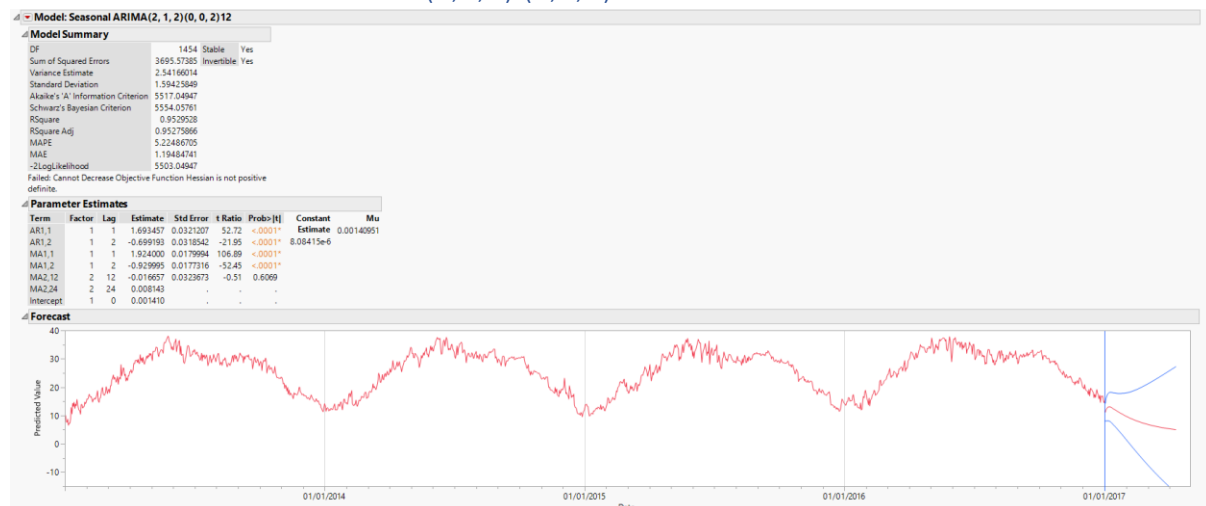


The MAPE and MAE metrics for this model are 5.227 and 1.195 respectively. We can see that all the terms are statistically significant and the model provides us with good level of forecast which can further be compared with the test data to determine its accuracy based on our data.

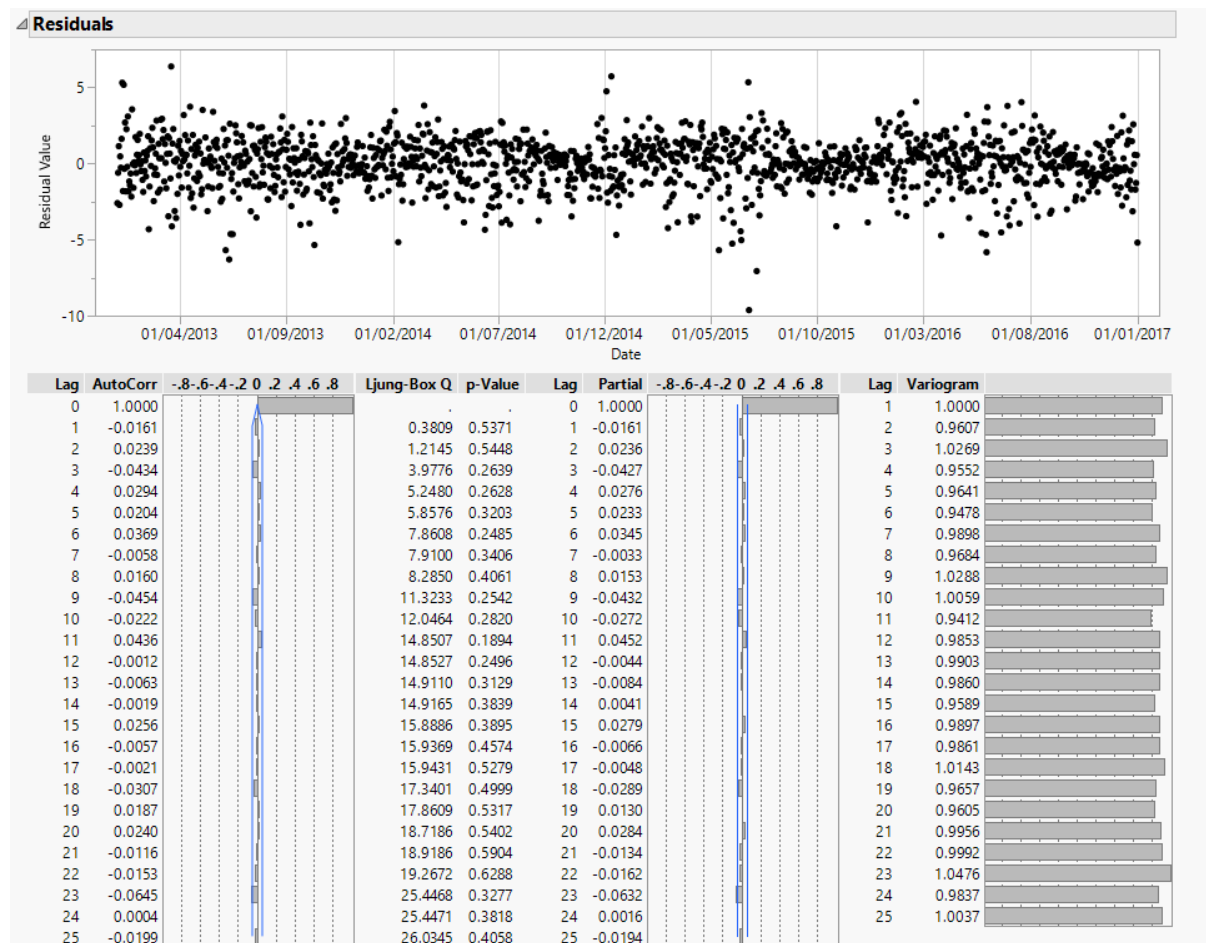


From the residuals, it is can be inferred that this model is adequate. The above figure shows the model is stationary with lags cut-off at 0 for both ACF and PACF plots.

MODEL 3 - SEASONAL ARIMA (2,1,2) (0,0,2) 12



In this model, it is evident that all the models are statistically significant except the MA[2] term. The MAPE and MAE values for this model are 5.224 and 1.194 respectively. The AIC for this model is 5517.04.



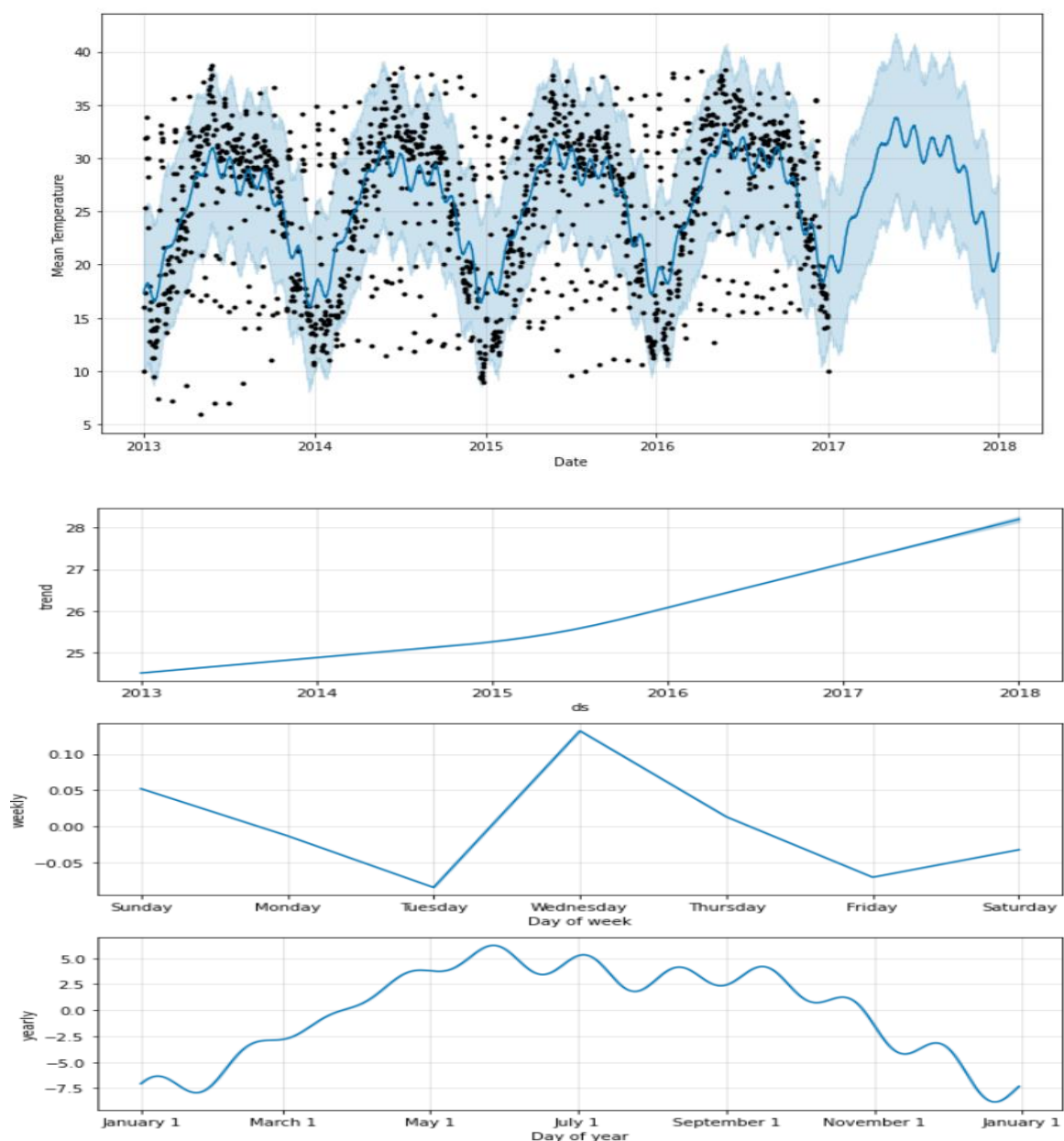
The above figure shows the model is stationary with lags cut-off at 0 for both ACF and PACF plots. There are a few lags where the auto correlation have border values in both the plots. The variogram is clear and depicts the stationary of the model.

MODELLING WITH PROPHET IN R:

Prophet is designed for analysing time series with daily observations that display patterns on different time scales. It has the advanced capabilities for modelling the effects of holidays on time series.

```
> library(prophet)
df <- read.csv('DailyDelhiClimateTrain.csv')
m <- prophet(df)
future <- make_future_dataframe(m, periods = 365)
forecast <- predict(m, future)
plot(m, forecast)
prophet_plot_components(m, forecast)
```

The generated data file contains the estimated mean temperature for the next year. We can also visualize the predictions with the help of the prophet plot function. The black dots represent the actual values, while the blue lines indicate the forecasted values and the light blue shaded regions corresponds to the prediction interval which is supposed to be 95%.



Looking at the above graph the trend clearly seems to be increasing. As we can infer that there is a seasonal pattern within the year where the mean temperature will be high during the period May – August and declines slowly at the end of summer.

CONCLUSION

We are now familiar with different time series analysis and prediction methods and approaches. Unfortunately, there is no perfect match to solve these kinds of problem. We seek for balance between MAPE and AIC values to arrive at the best model. From the above graphs and tables, it can be inferred that the accuracy of the forecasting of exponential smoothing models is better than the SARIMA model.

Out of the ARIMA models tested, we consider Model 2 – SARIMA(2,1,2)(0,0,1) 12 to be the best model for our data based on all the criteria. We could further conduct transfer analysis with the help of Humidity record present in the data set and check how temperature would affect humidity when additional noise is being added to the data.