Title**: ML1819 Research Assignment 1**
Team ID**: 43**
Task ID**: 103**
Paper Title**: Analyzing forecasting models based on different granularity of time series data.**

Student Details:

Student Name	Student ID
Lal Singh Dhaila	18304211
Roman Shaikh	18300989
Aman Ray	18301719

Contributions:

Student Name	Contribution
Lal Singh Dhaila	• Selection of appropriate time series dataset • Seasonality Removal Using three different methods namely log, moving average and difference • Custom code to identify minimum AIC value
Roman Shaikh	• Trend analysis • Testing of Stationarity using Augmented Dicky test • Calculation of MSE and MASE error
Aman Ray	• Data Cleaning • Comparison of ARIMA and SARIMA model • Plotting diagnostics ARIMA Model

Word Count: 992 (Excluding abstract, titles, index terms, labels and References)
Source Code: https://github.com/aman-ray-tcd/ML1819--task-103--team-43
Source Code Activity Link: https://github.com/aman-ray-tcd/ML1819--task-103--team-43/graphs/contributors

Analyzing forecasting models based on different granularity of time series data.

Lal Singh Dhalia
Trinity College Dublin
dhailal@tcd.ie

Roman Shaikh
Trinity College Dublin
shaikhr@tcd.ie

Aman Ray
Trinity College Dublin
raya@tcd.ie

***Abstract*—ARIMA and SARIMA are popular forecasting models used for problems like weather and stock price forecasting. These are based on data which have time series distribution. In this paper we analyses the effect of granularity of data on each of the algorithms and try to find the effectiveness based on training the model over a smaller granularity (monthly) and a large granularity (yearly)**

***Index Terms*—ARIMA, SARIMA, Forecasting model, data granularity, time series**

I. INTRODUCTION

In this paper we are focusing on forecasting problems that are based on time series data. Most of the prediction are based on the granularity of the data i.e. if the data collected is having the weekly, monthly or yearly relationship. In this paper we are analyzing the effectiveness of forecasting if done on data captured on monthly basis and yearly basis. Through this paper we are trying to demonstrate two things. Firstly, if the data is analyzed on low granularity level as an example instead on analyzing the data on yearly basis if prediction is done on monthly basis and all the prediction of twelve months are summed up to calculate the prediction for the year it may through a better result and how effective will it be (Ref. [2]). Secondly, we are checking the behavior of two algorithms that we will be using for prediction and will conclude which algorithm is more effective over the time period.

II. RELATED WORKS

Ming-Chien [6] in his research paper talked about use of multiple granularity level, which discuss more about time-delayed relationship. Another researcher J Beel [2] in his research paper discussed about use of small-time frame to check effectiveness of algorithm. We in this paper dive into the possibility of increasing the effectiveness of algorithm by analyzing data on smaller time series.

III. DATA PREPARATION

To do the forecasting analysis its necessary that data should be in time series format, which includes conversion of data into DateTimeIndex. Once the data is converted into time series its necessary to handle the outliers and missing values, which is being handled by removing the row.

IV. SATINARITY AND MODEL BUIDING

We are using ARIMA and SARIMA algorithms [1,3] to forecast the number of customers that are going to commute through San Francisco airport future. Below are the steps that have been taken care before forecasting:

1. Trend: From *Figure 1* an upward trend
2. Seasonality: In *Figure 1* peaks can be seen easily which signifies that there is seasonality in the data. Before proceeding further and forecasting we must remove the seasonality.

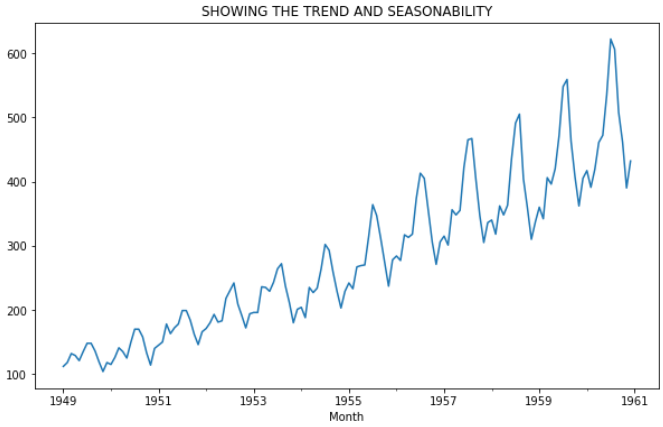

Figure 1: Showing the trend and seasonality in the data.

Data has been modified further to remove the seasonality and stationarity. To achieve the stationarity; moving average, central moving average or weighted moving average can be used. For this, we have used data moving average concept followed by log transformation.

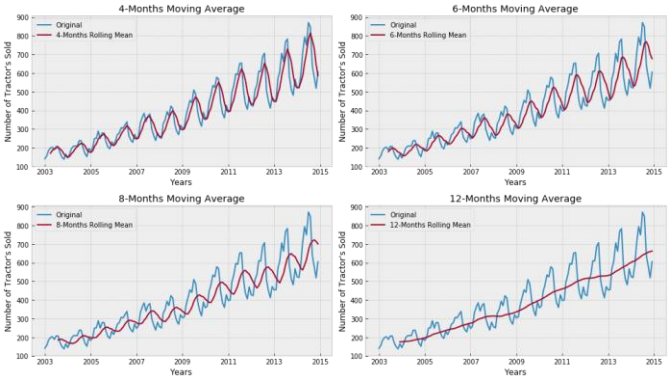

Figure 2: Making data stationary using moving average

To further test the stationarity, we performed the Augmented Dicky test. To check the stationarity, we compared the calculated p value against the p value of 0.05. Once the stationarity is achieved**,** the p and q value**s** are calculated using autocorrelation (ACF) and partial autocorrelation (PACF) function. The value of d is then calculated using the number of differences that have been done to achieve the stationarity. Based on the p, d, r values the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are selected which tells us to what extent the model is reliable. Lesser the value of the AIC**,** more reliable will be the model.

Statespace Model Results

Dep. Variable:	#Passengers	No. Observations:	143
Model:	SARIMAX(2, 1, 0)x(1, 1, 1, 12)	Log Likelihood	214.481
Date:	Sun, 28 Oct 2018	AIC	-418.961
Time:	00:20:36	BIC	-404.623
Sample:	02-01-1949 - 12-01-1960	HQIC	-413.135
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8928	0.082	-10.944	0.000	-1.053	-0.733
ar.L2	-0.3802	0.100	-3.798	0.000	-0.576	-0.184
ar.S.L12	-0.0811	0.168	-0.483	0.629	-0.410	0.248
ma.S.L12	-0.5467	0.189	-2.892	0.004	-0.917	-0.176
sigma2	0.0021	0.000	8.218	0.000	0.002	0.003

Ljung-Box (Q):	63.41	Jarque-Bera (JB):	2.86
Prob(Q):	0.01	Prob(JB):	0.24
Heteroskedasticity (H):	0.48	Skew:	0.11
Prob(H) (two-sided):	0.02	Kurtosis:	3.69

Figure 3: showing the complete State space model for SARIMAX.

V. FORECASTING/PREDICTION

Here we are forecasting the data for each month of the year.The prediction is done based on the training data. The robustness of the model is tested by deriving the below charts:

1. Standardized residual
2. Histogram plus estimated density
3. Normal Q-Q plot
4. Correlogram

Below *Figure 4* shows the plot diagnostics for ARIMA model.

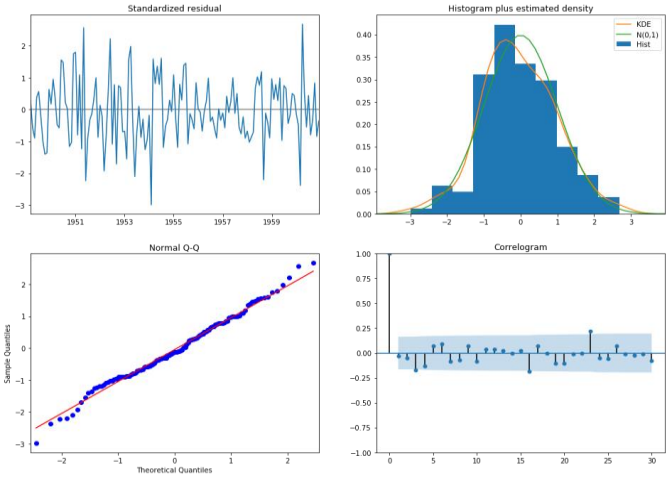

Figure 4: ARIMA model plot diagnostics

We need to ensure that the residuals of our model are uncorrelated and normally distributed with zero-mean. If they are not then it signifies that the model can be further improved, and the process is repeated with the residuals. The

KDE plot of the residuals on the top right is almost similar with the normal distribution. The *qq* plot on the bottom left (*Figure 4*) shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution with *N(0, 1)*. Again, this is a

strong indication that the residuals are normally distributed. The residuals over time (top left plot *Figure 4*) do not display any obvious seasonality and appear to be white noise. This is confirmed by the autocorrelation (i.e. correlogram) plot on the bottom right, which shows that the time series residuals have low correlation with lagged versions of itself.

VI. RESULT

We checked the reliability of this model (ARIMA and SARIMA) by dividing the data into test and train. Knowing the prediction of passengers commuting from the airport in future months, the values forecasted by the models were then plotted with actual predictions to check the reliability of the algorithm. *Figure 5* and *Figure 6* shows the prediction for future using SARIMA and ARIMA models respectively.

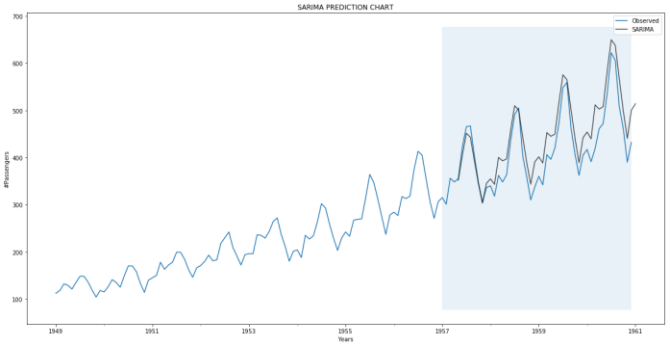

Figure 5: SARIMA future prediction after 1957.

In above *Figure 5* blue line is showing the actual observation and black line is showing the predicted value using SARIMA.

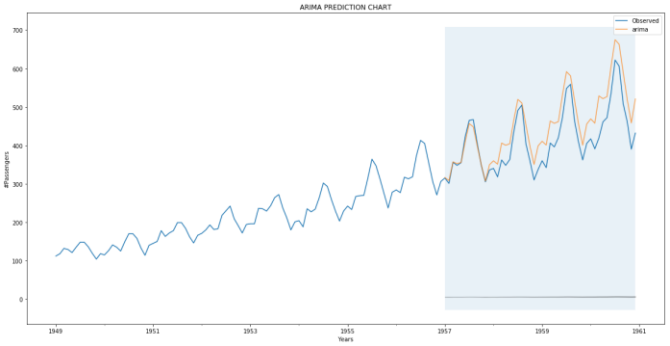

Figure 6: ARIMA future prediction after 1957.

Algorithm	MSE	MASE
SARIMA	0.0076	0.076
ARIMA	0.0113	0.0904

Table 1: MSE and MASE value for SARIMA and ARIMA (Monthly)

To check how well model worked in forecasting the passenger commute volume for the future months, we have first predicted the number of passenger commute in a time interval with known data. To show the accuracy of the two models over time series data the Mean Square Error (MSE) and Mean Absolute Scaled Error (MASE) are calculated as shown in *Table 1*.

Algorithm	MSE	MASE
SARIMA	50.9616	7.1359
ARIMA	50.8746	7.1299

Table 2: MSE and MASE value for SARIMA and ARIMA (Quarterly)

If the prediction is done considering quarterly intervals, we observed that the forecasted value is too low as shown in *Figure 7*. Moreover, calculation of MSE and MASE shows high value in case of quarterly calculation. So, we can conclude that prediction on a low granularity level (monthly) produces a better result over a high granularity data (quarterly). Also, the error rate is comparatively low while using low granularity data.

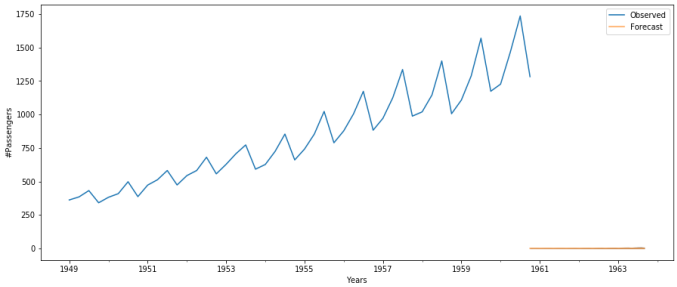

Figure 6: ARIMA and SARIMA showing same trend (quarterly)

VII. LIMITATIONS & OUTLOOK

Missing data is still a big challenge in this filed as replacing with mean average doesn't land on good forecasting. Moreover, if the data is analyzed based on more granularity level i.e. hourly, weekly would give us more a more robust forecasting model. Although SARIMA model is more effective but finding the prefect *pdq* value is still a challenge.

REFERENCES

[1] Peter J. Brockwell and Richard A. Davis "Introduction to Time Series and Forecasting,"

[2] Joeran Beel "It's Time to Consider "Time" when Evaluating Recommender-System Algorithms [Proposal]"

[3] Chris Chatfield, "The Analysis of Time Series: An Introduction, Sixth Edition," Ref. for SARIMA

[4] Roopam Upadhyay (2015) Forecasting & Time Series Analysis – Manufacturing Case Study Example (Blog Post).Retrieved from https://goo.gl/essiej

[5] Robert Nau (2018) ARIMA models for time series forecasting(Blog Post). Retrived from https://people.duke.edu/~rnau/411home.htm

[6] Mehmet Sayal and Ming-Chien Shan "Time series data analysis in multiple granularity levels". https://www.inderscienceonline.com/doi/pdf/10.1504/IJGCRSIS.2009.026725