Soekarno-Hatta Airport International Visitor Arrivals Analysis Using SARIMA and Holt-Winters Method

Introduction

Tourism has become a priority sector in Indonesia's economic development. It is expected to be one of key drivers in accelerating economic growth in Indonesia. To increase the marketing activities, it is required proper planning based on quantitave as well as qualitative information on international tourism performance in the past. This paper gives an analysis of international tourist arrival number in Indonesia over a span of 10 years.

Data set

The data used in this paper is based on monthly report of international visitor arrival at Soekarno-Hata international aiport by the Directoriate General of Immigration of Indonesia. The arrival data span from January 2008 to December 2017. The first six row of the raw data is shown below.

```
##   nama_tahun nama_turunan_tahun data_content nama_item_vertical_variabel
## 1       2017            Januari       185776              Soekarno-Hatta
## 2       2017           Februari       176958              Soekarno-Hatta
## 3       2017              Maret       208663              Soekarno-Hatta
## 4       2017              April       212502              Soekarno-Hatta
## 5       2017                Mei       206169              Soekarno-Hatta
## 6       2017               Juni       162588              Soekarno-Hatta
```

To prepare the data for downstream analysis, we have to clean it according to the method we will use. We change the date format from Indonesian format to standard English format. We also have to arrange the row of our data ascending according to the year and month of arrival. The source data is contains no missing value, so data imputation is not needed. Below is the data after this process has been done.

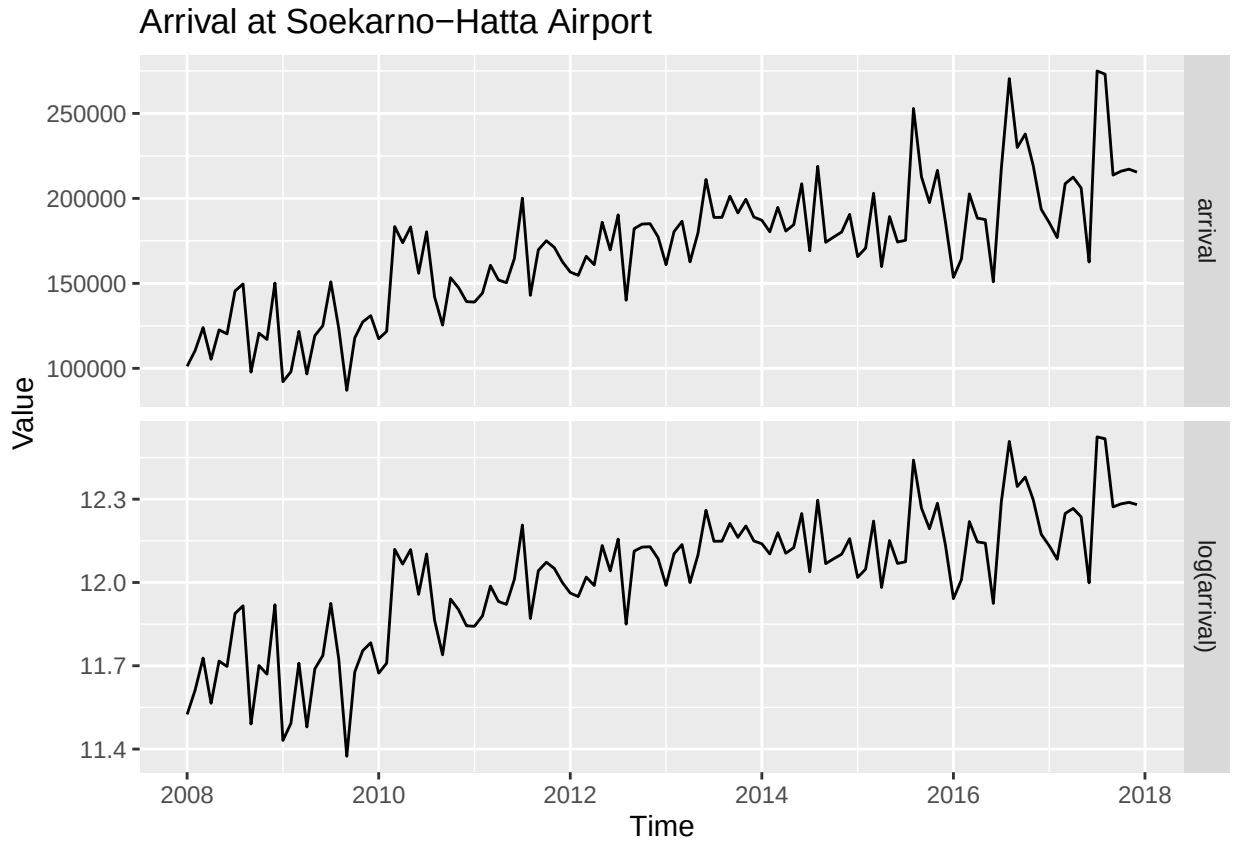
```
##         Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct
## 2008 101190 110477 123981 105338 122627 120270 145535 149635  97764 120683
## 2009  92136  97985 121699  96709 119231 125111 150924 123405  87047 117911
## 2010 117422 121727 183449 173906 183218 155951 180353 142050 125439 153300
## 2011 138987 144299 160650 151989 150407 164689 200180 142974 169777 175068
## 2012 156654 154698 165927 161005 185932 169682 190320 140077 182214 184894
## 2013 160998 180453 186548 162682 179737 211118 188800 188854 201336 191460
## 2014 187123 180362 194720 180787 184534 208624 169135 218903 174169 177274
## 2015 165746 170741 203019 159873 189307 174319 175347 252914 212706 197487
## 2016 153503 164317 202669 188369 187545 150956 217452 270496 229964 237914
## 2017 185776 176958 208663 212502 206169 162588 274974 273112 213721 216010
##         Nov    Dec
## 2008 117008 150209
## 2009 127299 130983
## 2010 147579 139242
## 2011 171215 162787
## 2012 185112 177335
## 2013 199511 189005
## 2014 180208 190598
```

```
## 2015 216517 186299
## 2016 219246 193629
## 2017 217201 215450
```

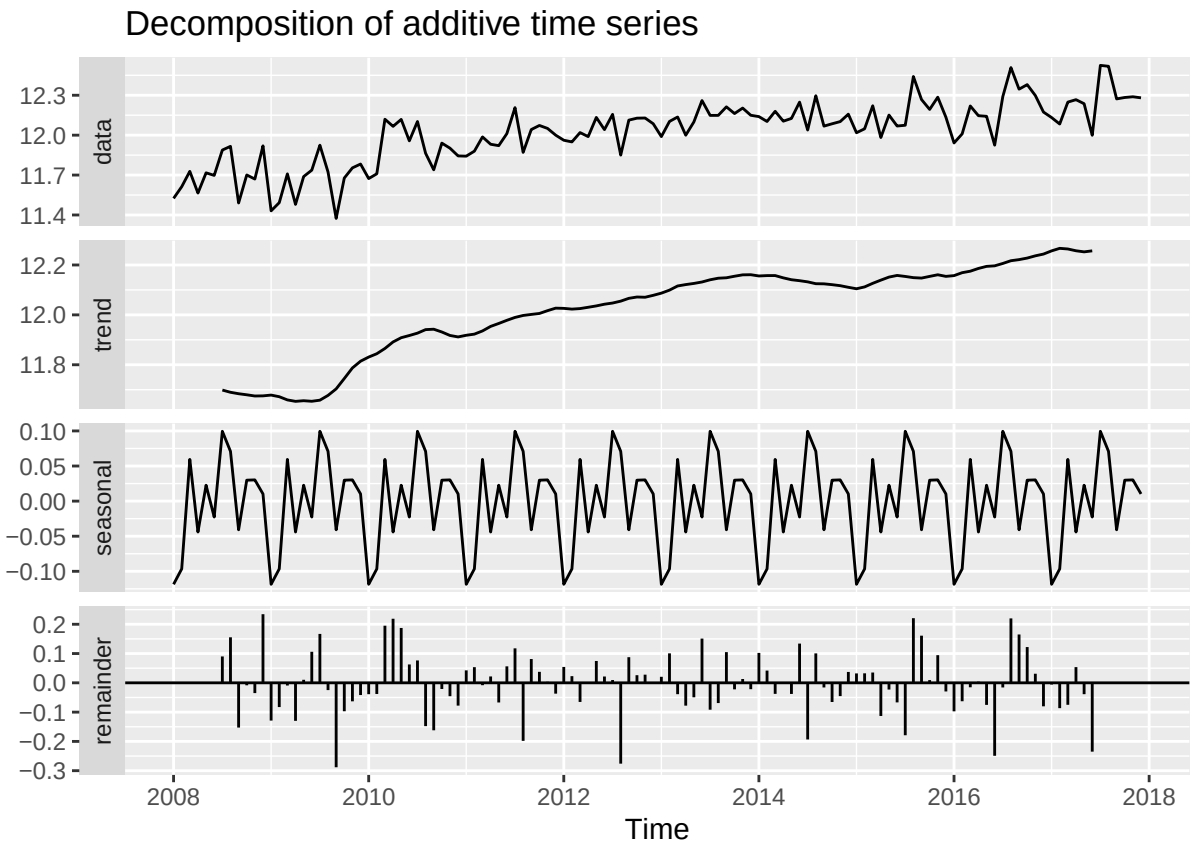
Results

Evaluate pattern of the data

Data from January 2008 to December 2017 are plotted below. From the plot we can see that the time series data is not stationary and strongly have trend and seasonal. There is an increase in the variance of the arrival value over time, so our first step is to take logarithms of data to stabilise the variance.



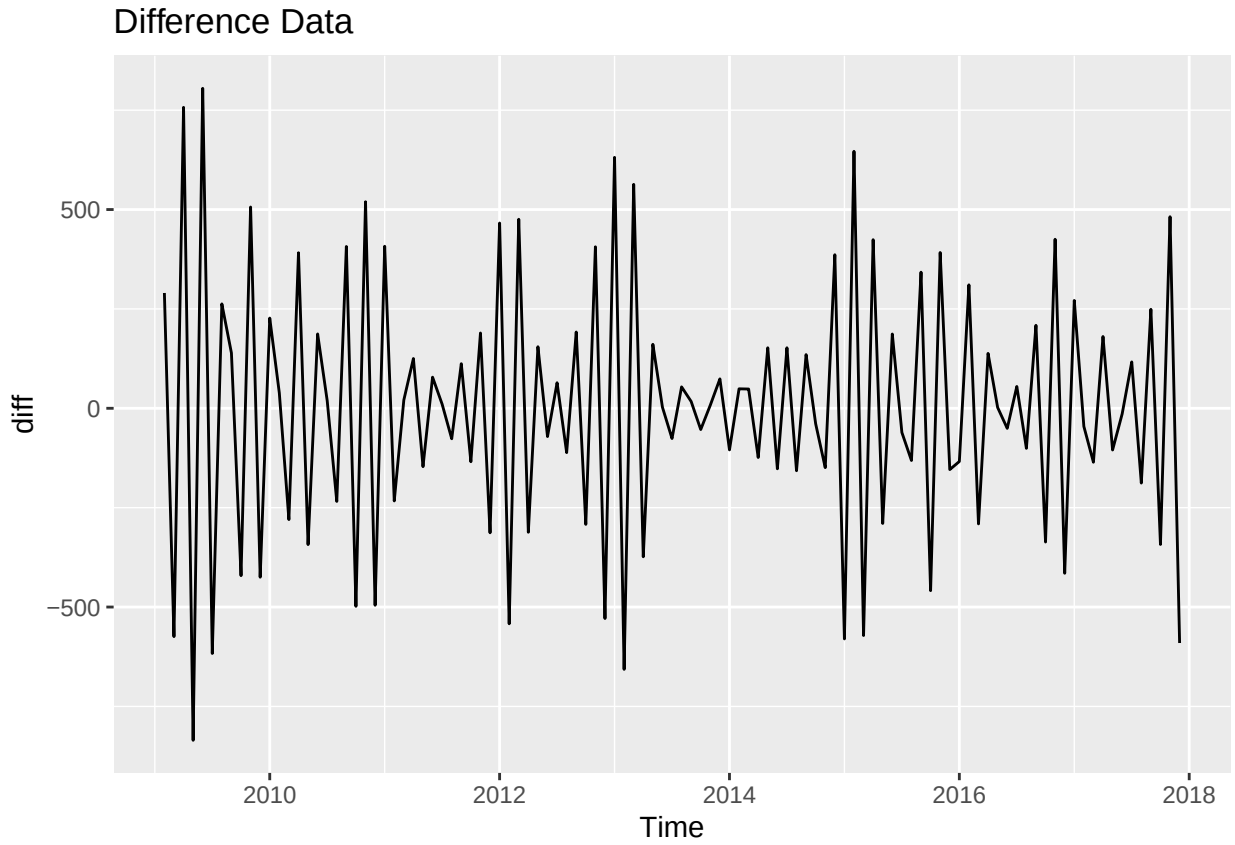
To confirm if there is trend and seasonality, we check the decomposition plot and performed Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests to check for stationarity. The output from KPSS test suggesting that our data are not stationary.




```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: mu with 4 lags.
##
## Value of test-statistic is: 2.1048
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

Transformations

To correct for trend pattern in the time series, we applied a difference of order 1. Furthermore, because the data is monthly, we take the seasonal period as 12 and applied difference of order 12 to the data to correct for seasonality. The output of KPSS test after this process suggesting that the difference data is stationary.



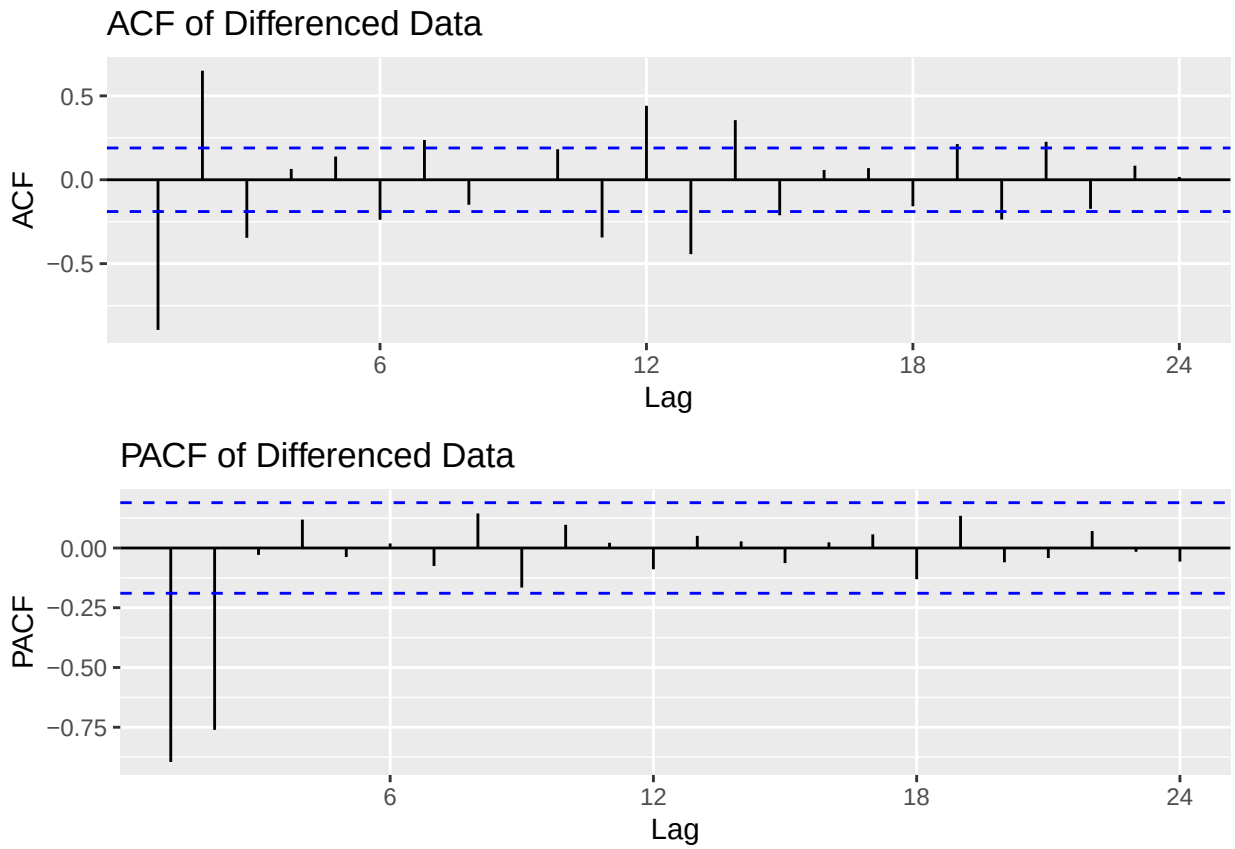

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: mu with 4 lags.
##
## Value of test-statistic is: 0.0599
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

Model selection: ARIMA

We used Seasonal ARIMA model for this time series to get a forecast for the future values from our data. The first approach for this process is to plot ACF and PACF from our differenced data and observing the behaviour of the autocorrelation and partial autocorrelation of it. For the simplicity of our first estimation, we consider a low order model of ARMA.

```
##
## Attaching package: 'gridExtra'
```

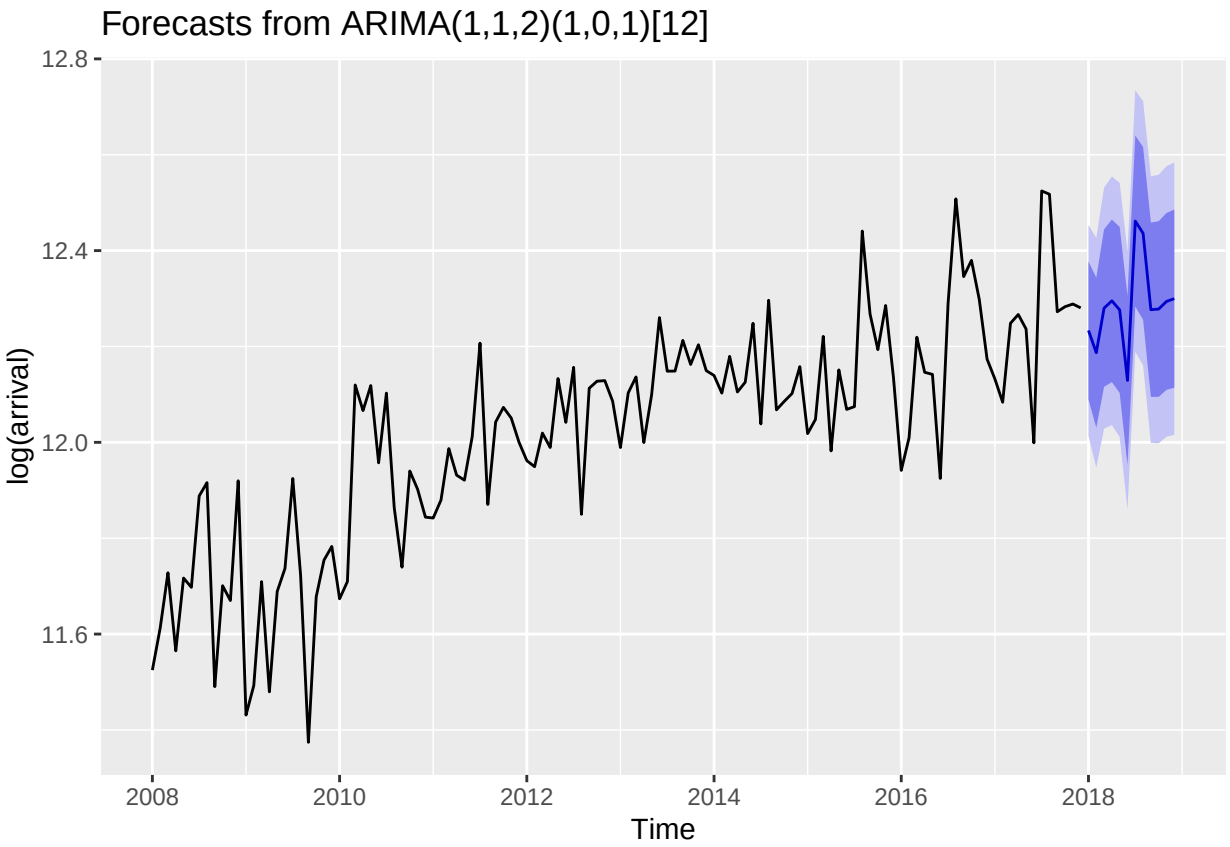
```
## The following object is masked from 'package:dplyr':
##
##     combine
```



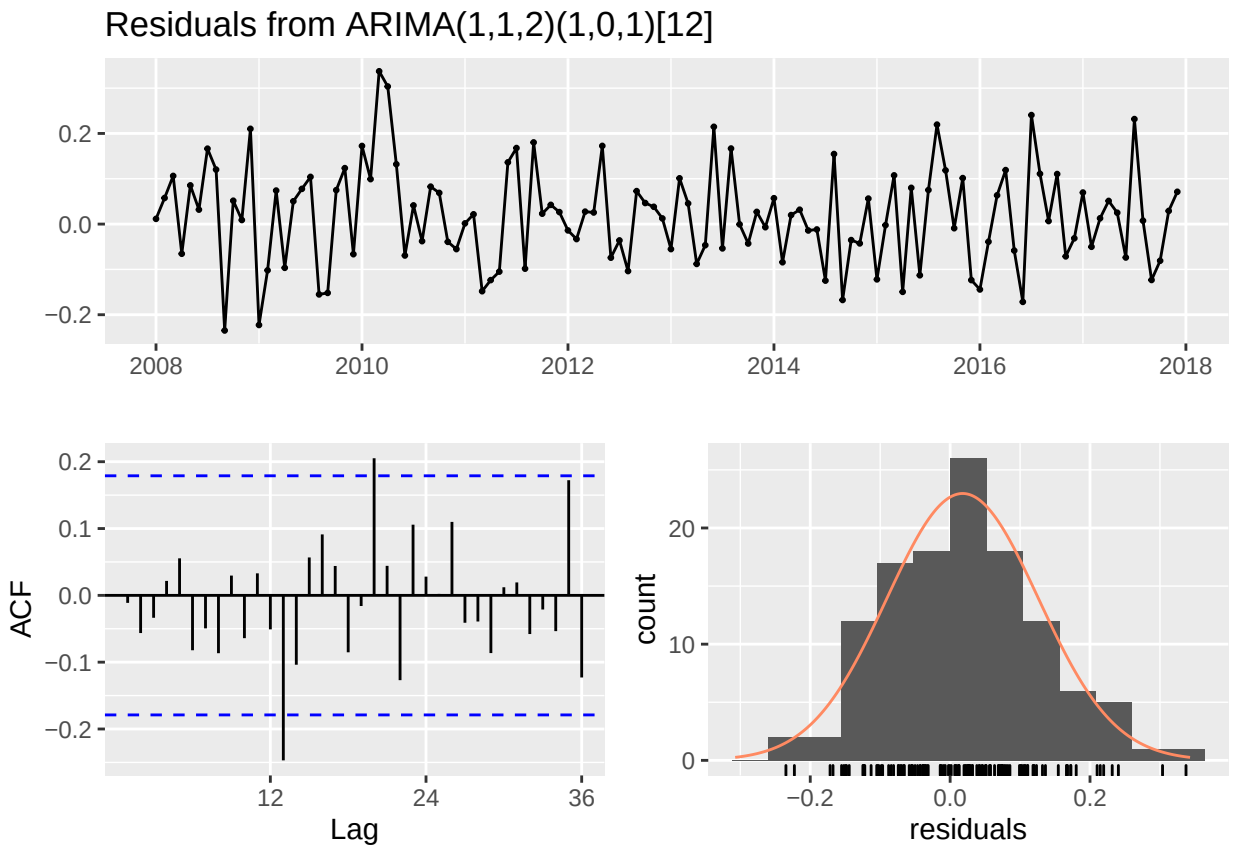
We can see that there is a significant spike at lag 12 in the ACF plot, confirming that the data is seasonal with period equal to 12. The decaying to zero of PACF and the significant spike at the first three lag of ACF suggests a non-seasonal MA(3) component. Looking at the lag 12 and 24, there is signitificant spike in the ACF for lag 12, but nothing at seasonal lags in the PACF. This may suggestive of a seasonal MA(1) component. Consequently, we begin with an $ARIMA(0,1,3)(0,1,1)_{12}$ model. Along with the initial model, we fit and compute AICc value of some variations on it, shown in the following table

```
##                     model_name      AICc
## 1 ARIMA(0, 1, 3)(0, 1, 1)[12] -134.5595
## 2 ARIMA(0, 1, 3)(0, 1, 2)[12] -135.7270
## 3 ARIMA(0, 1, 2)(0, 1, 1)[12] -135.9052
## 4 ARIMA(1, 1, 2)(0, 1, 1)[12] -139.7958
## 5 ARIMA(1, 1, 2)(0, 1, 2)[12] -139.2475
## 6 ARIMA(1, 1, 2)(1, 1, 1)[12] -144.3961
## 7 ARIMA(1, 1, 2)(1, 0, 1)[12] -168.1561
## 8 ARIMA(1, 1, 3)(1, 0, 1)[12] -165.9909
```

Of these modelsm the best is the $ARIMA(1,1,2)(1,0,1)_{12}$ model, since it has the smalles AICc value. The forecast of these model for the future values of next 12 months is given below

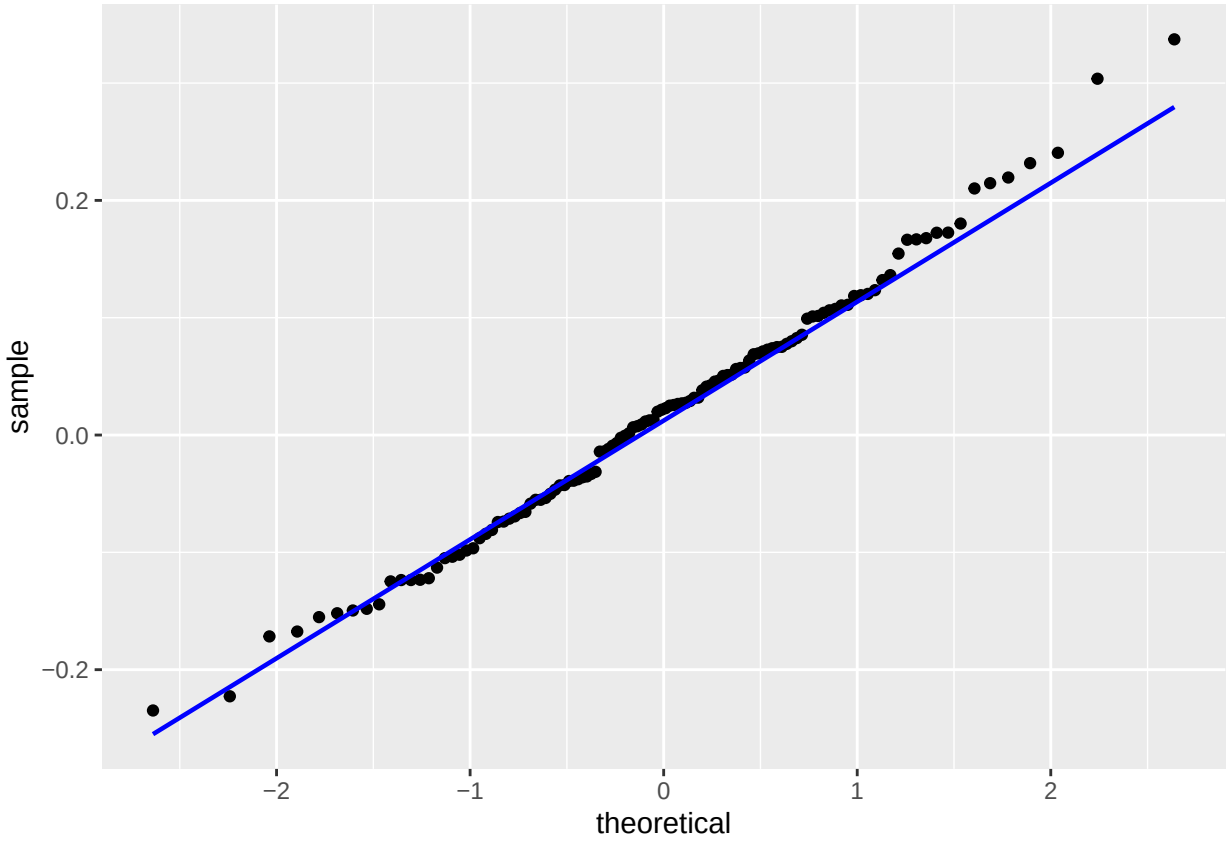

Forecasts from ARIMA(1,1,2)(1,0,1)[12]
12.8
12.4
12.0
11.6
log(arrival)
2008
2010
2012
2014
2016
2018
Time

Model diagnostics




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,2)(1,0,1)[12]
## Q* = 27.793, df = 19, p-value = 0.08749
##
## Model df: 5.   Total lags used: 24
```

The residuals from this model are shown above. There are a few significant spikes in the ACF of the residual. Checking the Ljung-Box test result, gives us a p-value of 0.1829. With this p-value, we accept the null hypothesis of Ljung-Box test that our residual data are independently distributed (i.e. the correlations in the residual data are 0). Moreover, plotting a QQ-plot of the residuals give us the following graph

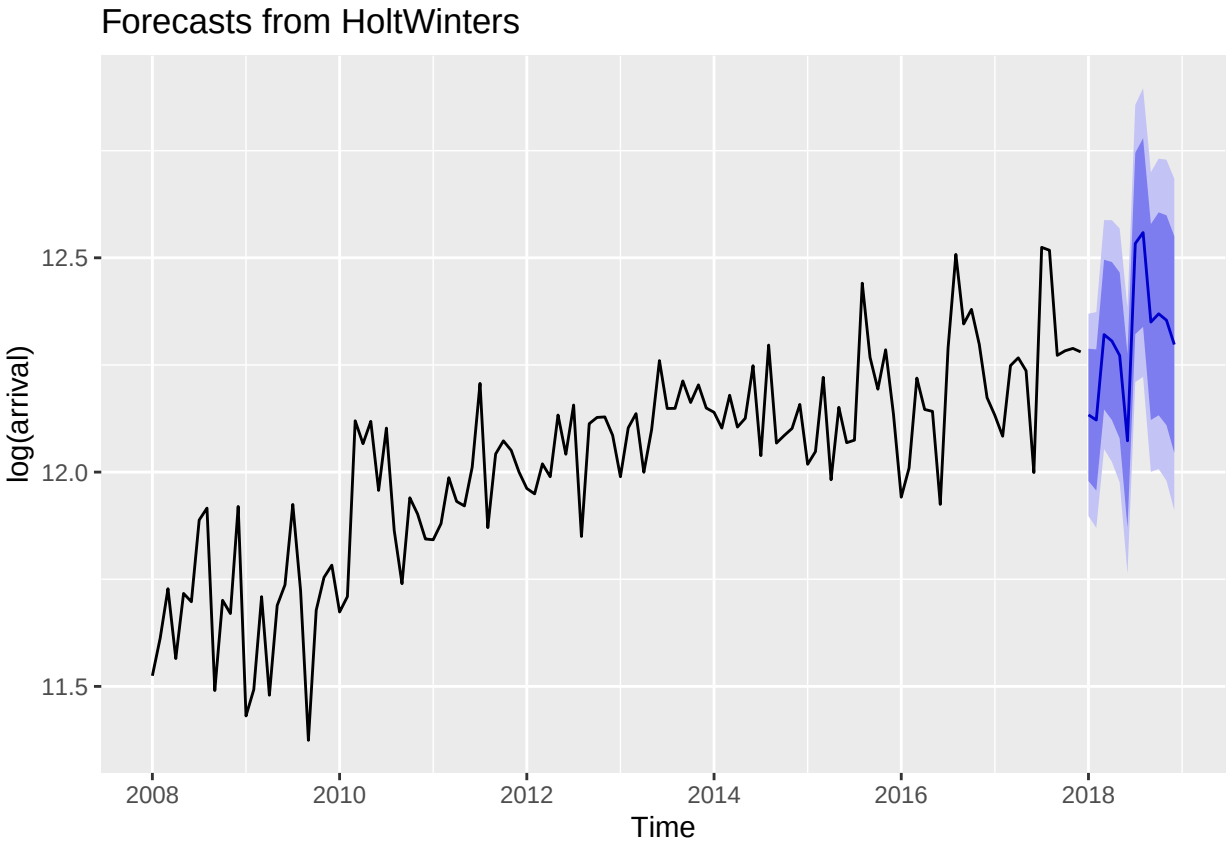


The QQ-plot shows that most of residual points are close to the identity line, confirming that our residuals are distributed as a normal distribution (i.e. white noise).

Smoothing: Holt-Winters

ARIMA is a parametric method to forecast a future value of a time series. Another way to forecast a time series is to use a non-parametric method such as Holt-Winters exponential smoothing. Forecasts produced using exponential smoothing methods are weighted averages of past observations, the more recent th observation the higher the associated weight.

We performed Holt-Winters' additive method to log-transformed data since the log-transformed data have a constant variance. The forecast using this Holt-Winter's method is shown below



Evaluating forecast accuracy

It is important to evaluate forecast accuracy of the two method mentioned above to determine more appropriate method to use for downstream analysis. To get an estimation for both method's accuracy, we separate the available data into two portions, a test set containing a full period of last seasonal data (12 data points), and a training containing the rest of data. We will train both methods on training set and compare the forecast values from both method with true values from test set.

To measure the accuracy, we used Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) as performance measurement metrics. The table below showing the result after running the test. RMSE value from Holt-Winters forecast is slightly better than ARIMA forecast, but for MAE and MAPE, it is evidence that ARIMA model is better than Holt-Winter smoothing for this specific training and test set.

```
##   performance_measurement holt_winters  arima_acc
## 1                    RMSE   0.08411116 0.08762042
## 2                     MAE   0.07459444 0.06618413
## 3                    MAPE   0.60836994 0.53611638
```

Conclusion

This paper has an objective to analyze the data on number of international arrival at Soekarno-Hatta airport, Indonesia from January 2008 to December 2017 and forecast the future values. We used two methods to approach the problem. The first method is a parameteric method, ARIMA modelling. We chose the best ARIMA model by observing the behaviour of ACF and PACF of our data and from several ARIMA model candidates, we chose model with smallest AICc value. The second method is a non-parametric method, Holt-Winter smoothing. Comparison of both method for this data done using a training-test set separation.

The result showed that the seasonal ARIMA model perform better for this specific data set.