Time Series Forecasting of the Number of Malaysia Airlines and AirAsia Passengers

N M Asrah¹, M E Nor², S N A Rahim³ and W K Leng⁴
¹,²,³,⁴ Department of Statistics and Mathematics, Faculty of Science, Technology & Human Development, Universiti Tun Hussein Onn Malaysia, 86400 Johor Malaysia

Abstract. The standard practice in forecasting process involved by fitting a model and further analysis on the residuals. If we know the distributional behaviour of the time series data, it can help us to directly analyse the model identification, parameter estimation, and model checking. In this paper, we want to compare the distributional behaviour data from the number of Malaysia Airlines (MAS) and AirAsia passenger’s. From the previous research, the AirAsia passengers are governed by geometric Brownian motion (GBM). The data were normally distributed, stationary and independent. Then, GBM was used to forecast the number of AirAsia passenger’s. The same methods were applied to MAS data and the results then were compared. Unfortunately, the MAS data were not governed by GBM. Then, the standard approach in time series forecasting will be applied to MAS data. From this comparison, we can conclude that the number of AirAsia passengers are always in peak season rather than MAS passengers.

1. Introduction

The current practice in time series forecasting begin with the fitting a model, the analysis then carried out based on the residuals. The behaviour of the time series data are very important. The process in analysis for model identification, parameter estimation and model checking can be easier since we know the behaviour of the data [1]. Time series forecasting are widely used in many fields such as financial [2], big data [3, 4], economy [5], hydrology [6], medical [7, 8] and many more.

While in stock markets, many mathematical models used in forecasting or predicting share prices compare to the time series method. Some of the methods are hidden Markov model, high-order fuzzy time-series model, moving average autoregressive exogenous with combination of Grey system theory and rough set, Markov Fourier Grey model, clustering-genetic fuzzy system and many more. A very simple mathematical model namely geometric Brownian motion (GBM) was applied to forecast the future share price for short period. The results based on GBM was found to be more accurate and can be measure the accuracy by using the MAPE value [9]. GBM was used by [10,11,12] in modelling the forecasting stock prices in the Buenos Aires Stock Exchange, developing the modified version of GBM model by using different data partitioning with and without jumps and found that it GBM is highly accurate for predicting the short term investment. Besides stock markets and finance, GBM also applied
in predicting the coal price [13], river flows [14], electrical consumptions [15], manufacturing industry [16], quality control [17] and many more.

In this study, we want to compare the distributional behaviour of AirAsia and MAS passengers. Based on the previous study [1], AirAsia was govern by GBM process. The distributional of the number of passengers were normal, independent and stationary. This data fulfilled the assumptions of GBM process. To further the study, we want to compare this data with MAS data. Both companies were the Malaysia airline companies. If MAS data govern by GBM, then the number of passengers of MAS can be forecast by using GBM process like in the previous study [1]. If not govern by GBM, the forecasting can be proceed with the suitable method.

2. Literature Review

Air transport is one of the world largest industries in transportation. It is one of the most crucial transports in the global transportation system. While the airline industry is the company that providing long distance travel over the world in a short time. The airline industry has achieved the high rate of growth [18]. Other than that, airline industry also play an important role in economic development. Air transport improves business operations by proving quick access to input supplies, stimulates innovative activities by facilitating face-to-face meeting and represents an essential input for certain industries where there is certainly anecdotal evidence suggestion [19]. Therefore, the demand for air travel has increase.

Malaysia Airline (MAS) is the holding company for the national carrier of Malaysia and one of Asia’s fastest growing airline. MAS was established in 1972. During 1980, the economics have been in a good situation. It helped spur growth at MAS. By the end of decade, 47 overseas destinations have been flying by MAS. In 1991, nearly 5.5 million travellers visited Malaysia. Besides, getting into the package tour business also helped Malaysian airline encourages increased passenger traffic [20]. Today, MAS has been serving for about 40,000 guests on 330 flights, which are more than 50 destinations daily [21].

The developments in technology and global economic crisis have affected the commercial airline industry [22]. Some research investigates the time series analysis of the airline industries such as [22, 23, 24, 25, and 26]. In this study, we want to check the distributional behaviour of MAS passengers. Then, the results will lead us to forecast the number of passengers by using the appropriate method.

3. Research Methodologies

The data used in this study have been collected from Malaysia Airport Holdings Berhad (MAHB) in Sepang, Selangor. Data involved were the number of MAS passenger’s departure from Kuala Lumpur International Airport (KLIA) to the international destinations daily. These data have the same date with AirAsia’s passengers in previous study by [1]. This is very important since we want to compare the
behaviour of both MAS and AirAsia passengers. The two data sets were from January 2009 until March 2009 and from January 2012 until May 2012.

3.1. Data Analysis

The previous study mentioned that AirAsia passenger’s data can be represented by geometric Brownian motion (GBM) process. This tell us that AirAsia passengers’ data were normally distributed, stationary and independent. These criteria must be fulfilled in order to confirm that it follows the GBM process. The normality of the log ratios of the data and the independence from the previous data are the most important test mentioned by [5] to test the GBM assumptions. For normality test, there are two important procedure for checking the distributions assumptions. Those procedures are empirical and statistical procedures [6]. The empirical procedures are more to graphical and visualized properties to check and validate the distribution assumptions. While statistical procedures are quantifiable and more reliable compare to empirical procedures. The Goodness of Fit test is one of the reliable test in this statistical procedure.

Based on the empirical procedures, QQ plot is used to check the normality of the data. The graphical information from the plot can tell us whether the data is normally distributed or not. The data is normal when the points of the plot lie close to a straight line. Then, the Anderson-Darling (AD) test is used to confirm the result. AD test is one of the Goodness of Fit test and suitable for data with small and large samples [6]. The formula for AD can be refer to the equation (1) [28] below:

\[
AD = n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1) \left[ \ln F(x_i) + \ln(1 - F(X_{n-i+1})) \right]
\]

(1)

where \( n \) is the sample size and \( F(x) \) is the cumulative distribution function.

The independence of the data can be checked by studying the relationship of the previous data. Scatter plot is used to visualize the relationship between the previous data. The data is independent if the data scattered and do not show any pattern. Again, it can be confirmed by looking at the correlation coefficient, \( r \). The smaller value and nearer to zero of \( r \) will confirmed the independency of the data. The formula for correlation coefficient \( r \) can be refer to the equation (2) below:

\[
r = -\frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}
\]

(2)
where \( X \) and \( Y \) are the variable represent for current and previous data respectively and \(-1 < r < 1\). Other than correlation coefficient, \( T \) test for independent can be used to test the hypothesis whether the data are independent or not [1]. The \( T \) statistics formulas can be refer to the equation (3) below:

\[
T = r \frac{\sqrt{n-2}}{\sqrt{1-r^2}}
\]

(3)

where \( T \) has a \( t \) distribution with \( n - 2 > 0 \) degree of freedom.

Stationary is also important to check the GBM process. It is crucial for time series data to be stationary to proceed with the next steps in developing the models and to do the forecasting [7]. This explained the process of finite variance which the mean and variance are constant in time. The lag dependent will represent the correlation between observations from different points in time.

3.2. Geometric Brownian Motion (GBM)

The geometric Brownian motion (GBM) can be referred as a valid model for the growth in the price of stock over time [27]. GBM is a mathematical approach for stock market modeling and its stochastic process have the assumptions that returns, profits or losses on the stock are independent and normally distributed [13].

The GBM model according to [27] as in equation (4) below

\[
L(n) = a + L(n-1) + e(n)
\]

(4)

where \( L(n) = \log(Sd(n)), e(n), n \geq 1 \) is a sequence of independent and identically distributed with mean zero (0) and variance \( \sigma^2 / n \), and \( a \) is equal to \( \mu / N \). In GBM, \( \mu \) is the mean (drift) parameter and \( \sigma \) is the associated volatility parameter. Equation (4) consider fitting a more general equation for \( L(n) \), the line is the associated volatility parameter. Equation (4) consider fitting a more general equation for \( L(n) \), the linear regression equation ar regression equation

\[
L(n) = a + bL(n-1) + e(n)
\]

(5)

where \( b \) is a constant value need to be estimated. Equation (5) is the classical linear regression model and the techniques to estimate the parameters \( a, b \) and \( \sigma \) is known. Equation (5) is called the
autoregressive model of order 1 (AR 1) since the log price at time \( n \) in terms of the log price one time period earlier. The parameter \( a \) and \( b \) of the autoregressive model given in Equation (5) are estimated from historical data.

Let \( L(0), L(1), \ldots, L(r) \) are the logarithms of the end-of-day price for \( r \) successive days. When \( a \) and \( b \) are known, the predicted value of \( L(i) \) based on prior log price is

\[
L(i) = a + bL(i-1)
\]

(6)

The log return from day \( i \) to day \( i+1 \) is follow equation (7) below

\[
R_i = \frac{S_d(n+1)}{S_d(n)} = \frac{L(n+1)}{L(n)}
\]

(7)

3.3. Time Series Analysis

The basic steps in modelling the time series model ARIMA were introduced by [32]. The model used for forecasting and the primary procedures are based on model identification, estimation, diagnosis checking, and consider alternative models if necessary [30]. ARIMA \((p,d,q)\) is the combination of autoregressive model AR \((p)\), degree of first differencing \((d)\) and integrated moving average model MA \((q)\) [31].

It is crucial to make sure that the time series data are stationary in its mean and variance before the ARIMA approach is applied. Logarithmic or power transformations is recommended to achieve stationarity in the variance [31]. The Box-Cox Transformation can be used to determine the parameter for power transformation. Equation (8) below shows the parameter for logarithmic or power transformations based on [32].

\[
y^\lambda = \begin{cases} 
  \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\
  \ln y, & \lambda = 0 
\end{cases}
\]

(8)

where \( \lambda \) is the parameter of Box-Cox transformation and \( y^\lambda \) is the transformation calculation. Moreover, autocorrelation function (ACF) and partial autocorrelation function (PACF) can also be used to test for the differenced series to check the stationary of time series data [33].
The full model of ARIMA model are written as in equation (9) below:

\[ Y_t = c + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} + e_t \]  

(9)

where \( Y_t \) is the differenced series, \( e_t \) is the white noise, \( \phi \) is the estimated coefficient and \( \theta \) is the estimated coefficient [34].

### 3.4. Mean Absolute Percentage Error (MAPE)

The mean absolute percentage error (MAPE) is the method to measure the performance of forecasting time series that widely used. It expresses the forecasting errors into percentage errors on actual observations and it is unit free. The formula for MAPE can be obtained in equation (10) as follows:

\[ \text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - F_i}{|A_i|} \right| \times 100 \]  

(10)

where \( A_i \) is the actual observation time series, \( F_i \) is the fitted observation time series and \( n \) is the number of non-missing data points [35]. The lower the value of MAPE, the forecast can be considered as more accurate. The interpretation of the MAPE values [36] can be referred in Table 1 below:

| MAPE values | Interpretation          |
|------------|-------------------------|
| < 10       | Highly accurate forecasting |
| 10-20      | Good forecasting         |
| 20-50      | Reasonable forecasting   |
| > 50       | Inaccurate forecasting   |
4. Results and Discussions

The results based on the methodologies in the previous sections are discussed. It starts with the test of GBM assumptions on normality, independency and also stationary. Then, based on the results, the analysis of both data sets will continue with the suitable method for forecasting

4.1. Tests of GBM Assumptions

Both data sets in 2009 and 2012 are normally distributed. Both QQ plots in Figure 1 and 2 shows that both data set are normally distributed at 5% significance level. The values for AD test in Table 2 also confirm that both data sets are normal since $p$-value is more than 0.05.

![Figure 1. QQ-plot for MAS passengers in 2009](image-url)
Figure 2. QQ-plot for MAS passengers in 2012.

Table 2. Anderson-Darling (AD) test for normality

| Year    | AD value | p-value |
|---------|----------|---------|
| MAS 2009 | 0.260    | 0.703   |
| MAS 2012 | 0.329    | 0.513   |

Meanwhile, the independency of both data can be seen from Figure 3 and 4. The scatter plot for MAS passengers in 2009 as in Figure 3 show that it scattered and did not show any pattern and the result from correlation coefficient and $T$ test in Table 3 confirm it independent at 5% level of significance since $T$ value is smaller than value for $t$ distribution table. The value in the bracket is from the $t$ distribution table.
Figure 3. Scatter plot for logarithmic return of MAS passengers in 2009

Figure 4. Scatter plot for logarithmic return of MAS passengers in 2012.

However, even though the scatter plot for MAS passengers in 2012 in Figure 4 look like scattered and did not show any pattern, the result for correlation coefficient and \( T \) test show otherwise.
The value for coefficient correlation is higher than in 2009 and $T$ test conclude that data is not independent.

**Table 3. Correlation coefficient test for independency**

| Year / Test                  | MAS 2009 | MAS 2012 |
|------------------------------|----------|----------|
| Pearson correlation          | 0.3114   | 0.3857   |
| $R^2$                        | 0.097    | 0.1488   |
| $T$ statistics and $t$       | 3.0570 (2.2809) | 5.1207 (2.2641) |

Stationary is another assumptions that must be fulfilled in order to follow the GBM process. The stationarity plot for both data sets can be seen in Figure 5 and 6. Based on the Figure 5 and 6, it show that both data sets are constant at zero. It can be conclude from the both logarithmic return plot show the stationary pattern.

![Figure 5. Chart for logarithmic return of MAS passengers in 2009](image-url)
From the results above, we can consider that MAS passenger’s data in 2009 follow the GBM process based on the three assumptions mentioned before. At the same time, MAS passenger’s data in 2012 is not follow the GBM process. It is because this data not fulfilled the independent assumption in GBM process.

4.2. Forecasting the number of passengers based on Box-Jenkin method

Next, the analysis is continue with time series forecasting analysis. From the results, the suitable time series model for data in 2009 is \( \text{SARIMA}(0, 0, 1)(1, 0, 0) \). Meanwhile the suitable time series model for data in 2012 is \( \text{SARIMA}(2, 0, 0)(0, 1, 1) \). Figure 7 and 8 shows the comparison between the forecasting numbers of passengers with the actual number of passengers for April 2009 and June 2012 respectively.
Figure 7. Comparison between the forecast number of MAS passengers and actual number of MAS passengers in April 2009.

Figure 8. Comparison between the forecast number of MAS passengers and actual number of MAS passengers in June 2012.
Table 4. MAPE values

| Year   | MAPE (%) |
|--------|----------|
| MAS 2009 | 9.6      |
| MAS 2012 | 7.7      |

The accuracy of both models can be interpreted based on the MAPE values in Table 4 above. Both values show that the models are highly accurate since the percentages are less than ten percent (10%).

5. Conclusions and Recommendations

The results obtained from this study found that data for MAS passengers are not follow the GBM process. The time frame for data sets in this study were exactly like the time frame for AirAsia passengers in the previous study. Even though the data sets are choose to be the same, it still not produce the same results. Since MAS passenger’s data set are not govern by GBM, Box-Jenkins method was applied to forecast the number of MAS passengers.

There is a different between the numbers of passengers for both companies. The different maybe based on the type or function for those airline companies. AirAsia is the low-cost airline company while MAS is the major airline operating flights from Kuala Lumpur International Airport (KLIA) to destination throughout Asia, Oceania and Europe. Although MAS operation is longer than AirAsia, the number of passengers for AirAsia is rapidly increase and almost rival the number of MAS passengers.

For the future study, it is recommend that to apply the data from different local and international airline companies. It includes the low-cost airline companies and major airline operating companies. Perhaps we can study the behaviour of these airline companies. It is interesting to know whether the low-cost airline companies are govern by GBM process while the major airline companies are not. The time frame can be longer compare to the time frame in this study.

Acknowledgments

We would like to express our thanks to Universiti Tun Hussein Onn Malaysia (UTHM) for providing facilities and financial supports. This research was carried out under UTHM Short Term Grant, Vot U369. The authors also gratefully acknowledge the Research and Planning Department of Malaysia Airport Holdings Berhad (MAHB) for the data provided in this study. Special thanks go to the anonymous referees for their constructive comments and suggestions.
References

[1] Asrah N M and Djauhari M A 2013 AIP Conf. Proc. vol 1557 (Malaysia: Selangor) pp 561-565
[2] Kyoung-jae K 2003 Neurocomputing 55 pp 307 – 319
[3] Kane A 2017 Proc. Int. Conf. on Multimedia Big Data (BigMM) (USA) pp 252-259
[4] Tahmassebpour M 2017 Proc. Int. Conf. on IEEE Access 5 pp 10694-10699
[5] Nathaniel B and Jonathan N K 2011 Annu. Rev. Polit. Sci 14 p 331
[6] Wang W C, Chau K W, Cheng C T and Lin Q 2009 Journal of Hydrology 374 pp 294–306
[7] Law A S, Freer Y, Hunter J, Logie R H, McIntosh N and Quinn J 2005 Journal of clinical monitoring and computing 19 pp 183-194
[8] Keogh E, Lin J, Fu A W and VanHerle H 2006 Proc. Int. Conf. on IEEE Transactions on Information Technology in Biomedicine 10 pp. 429-439
[9] Abidin S N Z and Jaffar M M 2012 World Applied Sciences Journal 17 pp 87-93
[10] Laddle G S and Wu L 2009 Nonlinear Analysis: Theory Methods & Applications 71 pp 1203-1208
[11] Willmoot P 2000 Quantitative Finance (Chichester: John Wiley & Son, Ltd)
[12] Siti N Z A and Maheran M J 2012 Proc. Int. Conf. on Decision Modeling (Malaysia: EDC-UUM) pp 149-151
[13] Shafiee S and Topal E 2007 Proc. Int. Conf. on Australian Mining Technology pp 173-179
[14] Lefebvre M 2002 Hydrological Processes pp 1373-1381
[15] Zolkepley Z and Djauhari M A 2012 Proc. Int. Conf. on Statistics in Science, Business, and Engineering (Malaysia: Langkawi) pp 1-3
[16] Li L S and Djauhari M A 2013 AIP Conf. Proc. vol 1557 (Malaysia: Selangor) pp 556-560
[17] Sagadavan R and Djauhari, M A 2013 AIP Conf. Proc. vol 1557 (Malaysia: Selangor) pp 571-575)
[18] Okulski R R and Heshmati A 2010 Time Series Analysis of Global Airline Passengers Transportation Industry (Seoul National University: Ph.D thesis)
[19] Blonigen B A and Cristea A D 2012 Airports and Urban Growth: Evidence from a Quasi Natural Policy Experiment (University of Oregon)
[20] Vatikiotis and Michael 1992 Far Eastern Economic Review pp 65-68
[21] Malaysiaairlines 2016 Our Story http://www.malaysiaairlines.com/my/en/corporate-
[22] Okulski R R and Heshmati A 2010 Technology management, economics and policy program TEMEP p 65

[23] Suzuki Y 2000 Transportation Research Part E: Logistics and Transportation Review 36 pp 139-154

[24] Brons M, Pels E, Nijkamp P and Rietveld P 2002 Journal of Air Transport Management 8 pp 165-175

[25] Semenick A I M and Sickles R C 2000 International Economic Review 41 pp 203-218

[26] Nam K and Schaefer T 1995 Logistics and Transportation Review 31 pp 239

[27] Dmouj A 2006 Stock Price Modelling: Theory and Practice (Vrije Universiteit: Masters Degree Thesis)

[28] Ross S M 2003 An Elementary Introduction to Mathematical Finance: Option and Other Topics (Cambridge University Press)

[29] McNeese B 2011 Anderson-Darling Test for Normality
   https://www.spcforexcel.com/knowledge/basic-statistics/anderson-darling-test-for-normality

[30] Chatfield C and Prothero D L 1973 Journal of the Royal Statistical Society. Series A (General) 136 pp 295-336

[31] Makridakis S and Hibon M 1997 Journal of Forecasting 16 pp 147-163

[32] Box G E and Cox D R 1964 Research Methods Meeting of the Society pp 211-252

[33] Mahsin M, Akhter Y and Begum M 2012 Journal of Mathematical Modelling and Application 2012 1 pp 67-73

[34] Hyndman R J and Athanasopoulos G 2014 Forecasting (Lexington, Ky : Otexts)

[35] Ren L and Glasure Y 2009 International Atlantic Economic Society 2009 15 pp 409–420

[36] Moreno J J N, Palmer P A, Sese A A and Cajal B B 2013 Psicothema 23 pp 500-506