Models for Train Passenger Forecasting of Java and Sumatra

Sartono
Finance Education and Training Agency - Ministry of Finance the Republic of Indonesia

Abstract. People tend to take public transportation to avoid high traffic, especially in Java. In Jakarta, the number of railway passengers is over than the capacity of the train at peak time. This is an opportunity as well as a challenge. If it is managed well then the company can get high profit. Otherwise, it may lead to disaster. This article discusses models for the train passengers, hence, finding the reasonable models to make a prediction overtimes. The Box-Jenkins method is occupied to develop a basic model. Then, this model is compared to models obtained using exponential smoothing method and regression method. The result shows that Holt-Winters model is better to predict for one-month, three-month, and six-month ahead for the passenger in Java. In addition, SARIMA(1,1,0)(2,0,0) is more accurate for nine-month and twelve-month oversee. On the other hand, for Sumatra passenger forecasting, SARIMA(1,1,1)(0,0,2) gives a better approximation for one-month ahead, and ARIMA model is best for three-month ahead prediction. The rest, Trend Seasonal and Liner Model has the least of RMSE to forecast for six-month, nine-month, and twelve-month ahead.

1. Introduction
The train is one of the most preferred transports in Indonesia, especially on Eid day. It is a period for people to travel from city to village to see their relatives as a part of local culture in Java and most of the regions in the country. Ministry of Transportation the Republic of Indonesia releases news that during Eid 2016 the number of passengers using a bus (and other similar public transportation) decreases 5.59% from 4,697,945 last year to 4,416,119 this year. On the other hand, the number of passenger for the railway is 4,080,319, raised 3.78% from 3,931,712 its last year [1]. It seems that people likely prefer to take the train for some reasons. In Jakarta, passengers often need to wait for the next train since it has been full at peak time. This indicates a high demand that offers high profit for train provider.

The opportunity should be managed well to optimize revenue, otherwise, the overwhelming customers can lead to disaster. The manager may buy a new set of a train or build a new track to meet the customers need, but the cost would be a lot. Therefore, it will be useful to know how the number of passengers behaves over time. Kusakabe [2] develop a model to estimate the passengers of train choice in Japan using smart card transaction data. Changes, modifications, innovations and creative activities occur by chance and unsystematically in one of the train company in Iran [3], however, managing train capacity allocation can increase the Avenue from 2.6 to 29.3 per cent as well as 8.4 to 29 per cent in passengers carried in India [4]. Besides, modeling can help to improve the manually planned allocation of train passengers in Dutch [5]. In Indonesia, Sartono [6] shows that ARIMA+GARCH model serves a better prediction for goods delivery by train in Java than the other developed models. These results offer motivation to uncover the behavior of the passengers of the train.
Understanding the pattern of train passengers' data may provide a reasonable suggestion for the decision maker to create an accurate plan in controlling and improving the service. This paper will discuss some models which can be occupied to forecast the number of train passenger for the area of Java and Sumatra. These two regions are the first and the second dense island in Indonesia [7]. Now, business on the railway is still concentrated in those areas. This article is a preliminary study which may be advantageous as one of the references for further research on train modeling.

2. Models Building
To construct an appropriate model, a monthly data set of train passengers from January 2006 up to December 2015 is utilized. The set consists of data for Java and Sumatra region. For each region, the data is divided into two parts. One part contains the observed values from 2006 to 2014 which is used to develop some models. The rest of the data is exploited for model comparison to determine the most suitable model among methods for passengers of train data in each area.

2.1. Basic Method
There are no plenty of study which exactly talks about passengers of train prediction, but there are a huge number of researches in forecasting. Many of them use the Box-Jenkins method to build the model. Therefore, it is chosen to be the base of the method applied in this project. It will be compared to other methods to get the desired model.

Commonly, the Box-Jenkins method is known as ARIMA method. It can be written as ARIMA($p$, $d$, $q$), in which $p$, $d$, and $q$ represent the order of auto-regressive, differencing, and moving-average respectively (see [8] page 33-86 to figure out the background of the model). In general, if a series $\{X_t\}$ is an ARMA($p$, $q$), it can be expressed in

$$X_t = \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_p X_{t-p} + \omega_t + \delta_1 \omega_{t-1} + \delta_2 \omega_{t-2} + \ldots + \delta_q \omega_{t-q}$$

where $\omega_t$ is a white noise.

Plotting the data set could be the first step to utilizing this method. It will serve a brief view of the behavior of the data, the possibility of any required transformation, and the existence of any anomaly. ACF plot helps to inspect whether differencing is needed [9]. Next step is evaluating whether the series is stationary. The plot can be used to know it, but a formal test may be more accurate. Augmented Dickey-Fuller test [10] is one of the tools to capture whether the data is non-stationary. When a series is non-stationary, the differencing process may change it becomes a stationary series. If a series is stationary after differencing once, then the order of differencing in ARIMA model is one (written as ARIMA($p$, 1, $q$)). It is possible to have some models that fit to data, in this case, the best model can be determined by the Akaike Information Criterion (AIC) [11].

The ARIMA model is based on the assumption that the errors are independent and identically distributed to zero mean and constant variances. It is checked in the model validation step. The ACF and PACF of the residuals exhibit an indication of the independence. However, a result from formal tests such as McLeod-Li test [12] and Ljung-Box test [13] may be more reassuring. Besides, if the variances are not constant, the auto regressive conditional heteroscedasticity (ARCH) or general auto regressive conditional heteroscedasticity (GARCH) will be worthwhile to fix it.

A series may contain a seasonal effect as well as a trend. SARIMA($p$, $d$, $q$)($P$, $D$, $Q$) is the model based on ARIMA model which can capture the seasonal effect. In this model, ($p$, $d$, $q$) is the order of non-seasonal term and ($P$, $D$, $Q$) is the order of seasonal term.

2.2. Compared Method
The model obtained by Box-Jenkins method will be compared to models constructed by smoothing method and regression method. In this paper, they are Holt-Winters model and Trend and Seasonal Linear Model (TSLM) respectively. The Holt-Winters model can be written in

$$a_t = \alpha (X_t - S_{t-q}) + (1 - \alpha)(a_{t-1} + b_{t-1})$$

$$b_t = \beta (a_t - a_{t-1}) + (1 - \beta)b_{t-1}$$

$$s_t = \theta (X_t - a_t) + (1 - \theta)s_{t-q}$$

(2)
where \(a_i\) and \(b_i\) are estimated the level and slope respectively, and \(s_t\) is estimated seasonal effect at the time \(t\), and \(\alpha, \beta\), and \(\theta\) are smoothing parameters. The TSLM is defined as

\[
X_t = m_t + s_t + \varepsilon_t
\]

(3)

where \(m_t\) is the trend term, \(s_t\) is the seasonal term, and \(\varepsilon_t\) is the error term at the time \(t\). If the trend is straight line, then equation (3) can be written as

\[
X_t = \begin{cases} 
  a_1 t + \beta_1 + \varepsilon_t & t = 1, 13, 25, \ldots \\
  a_2 t + \beta_2 + \varepsilon_t & t = 2, 14, 26, \ldots \\
  a_3 t + \beta_3 + \varepsilon_t & t = 3, 15, 27, \ldots \\
  \vdots & \\
  a_j t + \beta_j + \varepsilon_t & t = 12, 24, 36, \ldots 
\end{cases}
\]

(4)

here, \(s_t = \beta_j\) if \(t\) is in the \(j\)-th season. The parameter in equation (4) can be estimated using Ordinary Least Square (OLS) method.

2.3. Comparison Tool
To find out the best model an analyst need to investigate the ability of the models in forecasting. The better model is the closer its prediction value to the real. The goodness of a model can be evaluated using in-sample and out-sample forecast. Define one-step ahead forecasting errors as

\[
e_t(t) = X_t - \hat{X}_t(t-1)
\]

(5)

Consider to equation (5), Mean Squared Error (MSE) is defined as

\[
MSE = \frac{1}{n} \sum_{t=1}^{n} (e_t(t))^2
\]

(6)

The root of MSE is called RMSE. The more accurate a model the less its RMSE.

3. Result
The data set consist of two parts, Java and Sumatra. In each part, the first model is constructed by Box-Jenkins method. Then, the best model from this step is compared to Holt-Winters model and TSLM model. All computations, hypothesis testing, and plotting for figures are done using R-software.

3.1. Passengers of Train in Java
The plot of the original data and its decomposition can be seen at Figure 1. It seems that there is a seasonal effect, so SARIMA and other models involving seasonal term could be fit the data. The first step is to build the ARIMA model. \(P\)-value = 0.9071 on Dickey-Fuller test suggests that the data is not stationary. Differencing once gives \(p\)-value = 0.01 which means the data is stationary now with \(d=1\) in ARIMA model. Thus, the model will be in the form of ARIMA \((p, 1, q)\).
Figure 1. (a) the number of passengers of train in Java 2006-2014, (b) the composition of data in (a)

Table 1. Table of AIC for Java region

| No | ARIMA        | AIC     | No | ARIMA        | AIC     |
|----|--------------|---------|----|--------------|---------|
| 1  | (0, 1, 0)    | 1807.983| 9  | (2, 1, 0)    | 1795.162|
| 2  | (0, 1, 1)    | 1795.169| 10 | (2, 1, 1)    | 1795.759|
| 3  | (0, 1, 2)    | 1793.930| 11 | (2, 1, 2)    | 1785.396|
| 4  | (0, 1, 3)    | 1795.872| 12 | (2, 1, 3)    | 1785.000|
| 5  | (1, 1, 0)    | 1793.803| 13 | (3, 1, 0)    | 1793.785|
| 6  | (1, 1, 1)    | 1795.470| 14 | (3, 1, 1)    | 1795.723|
| 7  | (1, 1, 2)    | 1795.912| 15 | (3, 1, 2)    | **1784.965**|
| 8  | (1, 1, 3)    | 1796.982| 16 | (3, 1, 3)    | 1786.962|

Consider to Table 1, and the best model is ARIMA(3, 1, 2). Next, the residuals of the chosen model are investigated. Figure 2 reveals that the residuals are white noises. Moreover, Box-Ljung test provides $p$-value = 0.9742 which support the previous result. Therefore, no GARCH model is needed to maintain the residuals. Hence, the suitable ARIMA model is ARIMA(3,1,2). Using the AIC criterion too, the best model for SARIMA is SARIMA(1,1,0)(2,0,0) with AIC = 1738.395.

![Figure 2](image1.png)  
(a) ACF of the square of ARIMA(3, 1, 2) residuals, (b) PACF of the square of ARIMA(3, 1, 2) residuals

3.2. Passengers of Train in Sumatra

Figure 3 shows the number of passengers of the train in Sumatra from 2006 to 2014 and its decomposition. It is likely seasonal effect exists. Just as in Java, models which include seasonal term will be useful. SARIMA, Holt-Winters, and TSLM are some of them. They will be compared each other and so will the ARIMA model. To begin with ARIMA model, the series must be analyzed whether it is stationary. Dickey-Fuller test gives $p$-value = 0.4485, and this means that the series is not stationary. By first order differencing the $p$-value now is 0.01, so the process can be continued for the ARIMA($p$, 1, $q$).
Figure 3. (a) the number of passengers of train in Sumatra 2006-2014, (b) The composition of data in (a)

Table 2. Table of AIC for Sumatra region

| No | ARIMA   | AIC       | No | ARIMA   | AIC       |
|----|---------|-----------|----|---------|-----------|
| 1  | (0, 1, 0) | 1255.976  | 9  | (2, 1, 0) | 1198.158  |
| 2  | (0, 1, 1) | 1197.117  | 10 | (2, 1, 1) | 1197.165  |
| 3  | (0, 1, 2) | 1195.574  | 11 | (2, 1, 2) | 1198.253  |
| 4  | (0, 1, 3) | 1197.014  | 12 | (2, 1, 3) | 1191.712  |
| 5  | (1, 1, 0) | 1213.509  | 13 | (3, 1, 0) | 1198.755  |
| 6  | (1, 1, 1) | 1195.383  | 14 | (3, 1, 1) | 1197.873  |
| 7  | (1, 1, 2) | 1197.319  | 15 | (3, 1, 2) | 1199.800  |
| 8  | (1, 1, 3) | 1198.463  | 16 | (3, 1, 3) | 1201.665  |

Based on AIC criterion as shown on Table 2, ARIMA(2, 1, 3) is the best among the others. The residuals are white noises (see fig. 4). Furthermore, p-value for Box-Ljung test is 0.9782 which strengthen this result, hence, GARCH process is not required. On the other hand, by the AIC criterion, the best SARIMA for Sumatra series is SARIMA(1,1,1)(0,0,2) with AIC = 1171.225.

Figure 4. (a) ACF of the square of ARIMA(2, 1, 3) residuals, (b) PACF of the square of ARIMA(2, 1, 3) residuals

3.3. Models Comparison

From the previous work, the ARIMA(3,1,2) and SARIMA(1,1,0)(2,0,0) are the best on their class for modeling passengers of the train in Java. For Sumatra, they are ARIMA(2,1,3) and SARIMA(1,1,1)(0,0,2). In this section, those models are compared to the Holt-Winters model and TSLM model. For this purpose, the observed values in the year of 2015 are exploited. The MAE and RMSE are used in this article to select the best model, thus, the model which provides the least RMSE.

Table 3. Table of RMSE value in forecasting for Java region

| Time Forecast (Month) | ARIMA | SARIMA | Holt-Winters | TSLM |
|-----------------------|-------|--------|--------------|------|
| 1                     | 708.9387 | 685.2047 | **560.6887** | 4,707.7290 |
| 3                     | 1,889.8508 | 1,427.5119 | **1,145.4301** | 5,127.0130 |
| 6                     | 1,788.9301 | 1,181.6453 | **1,133.6436** | 5,649.3640 |
| 9                     | 1,830.3599 | **980.7162** | 1,016.8504 | 5,730.9110 |
| 12                    | 2,213.6562 | **873.8687** | 1,395.6856 | 6,060.6570 |

Table 4. Table of RMSE value in forecasting for Sumatra region

| Time Forecast (Month) | ARIMA | SARIMA | Holt-Winters | TSLM |
|-----------------------|-------|--------|--------------|------|
| 1                     | 4.7707 | **1.7723** | 6.7918 | 9.1319 |
| 3                     | **11.6466** | 24.0475 | 21.1706 | 21.7184 |
| 6                     | 22.6385 | 22.3371 | 19.9976 | **19.2721** |
| 9                     | 41.9038 | 37.9631 | 39.1960 | **28.7422** |
| 12                    | 43.4062 | 34.2374 | 36.3579 | **28.8428** |
Table 3 shows that the Holt-Winters model is more accurate than the others to estimate the number of passengers of the train for one, three, and six-month ahead, while, SARIMA is more suitable for nine-month and twelve-month estimation. On the other hand, SARIMA model only best for one-month ahead prediction for the region of Sumatra, and to predict in three-month ahead the ARIMA model makes the least error. Besides, Trend and Seasonal Linear Model is more promising to forecast for six, nine, and twelve-month in the future (see Table 4).

4. Summary
Some models have been developed and tested according to get the most suitable one for passengers of train data set. There is no model which dominates all area for all time forecast. For Java, the model obtained by Holt-Winters technique is better than ARIMA, SARIMA, and Trend and Seasonal Linear Model to predict in one, three, and six-month ahead. Nevertheless, it is no more accurate than SARIMA model as the best to foresee in nine and twelve-month in the future. However, it does not mean that SARIMA is better for long-time prediction in general. In fact, for the area of Sumatra SARIMA is more appropriate for one-month time forecast, and less accurate when used for three months or more. Besides, ARIMA is good for three-month ahead forecasting. The rest, Trend and Seasonal Linear Model is the fittest model to forecast for next six, nine, and twelve-month than other models for Sumatra region.

Holt-Winters, SARIMA, and TSLM are models which include the seasonal effect on their formula. It seems that the passengers of train follow a trend or seasonal pattern since those models coming up as the best model in almost all time forecast for both region. Further research may be executed to reveal the trend and seasonal specifically. Other methods considering seasonal effect may give the different view or even strengthen the result in this project.

References
[1] http://www.dephub.go.id/welcome/readPost/menhub-jonan--penumpang-angkutan-umum-lebaran-ini-naik-4.30
[2] Kusakabe, T., Iryo, T. and Asakura, Y. 2010 Estimation method for railway passengers' train choice behavior with smart card transaction data. *Transportation*, 37(5), pp 731-749.
[3] Feyzbakhsh, S. A., Sadeghi, R. and Shoraka, S. 2008 A case study of intrapreneurship obstacles: The RAJA passenger train company. *Journal of Small Business and Entrepreneurship*, 21(2), pp 171-180,251.
[4] Gopalakrishnan, R. and Rangaraj, N. 2010 Capacity management on long-distance passenger trains of indian railways. *Interfaces*, 40(4), pp 291-336.
[5] Abbink, E., Berg, B. v. d., Kroon, L. and Salomon, M. 2004 Allocation of railway rolling stock for passenger trains. *Transportation Science*, 38(1), pp 33-41.
[6] Sartono 2016 *Proc. Nat. Conf. Modelling for Goods Delivery by Railway in Java and Sumatra* (Surakarta: Muhammadiyah University of Surakarta, Mar 12, 2016) pp 790–799.
[7] Agency, I. S. C. 2015 *Percentage distribution of population by province*, 2000-2014, Retrieved Aug 5, 2016, from www.dephub.go.id
[8] Wei, W. W. S. 2006 *Time Series Analysis: Univariate and Multivariate Methods*, 2nd ed. (New York: Pearson)
[9] Sumway, R. H. and D. S. Stoffer, *Time Series Analysis and Its Application - with R Examples*, 2nd ed.(New York: Springer) pp 141–142.
[10] Dickey, D. and W. Fuller, 1979 *J. Am. Stat. Assoc* 74, pp 427–431.
[11] Akaikie, H. 1973 *Biometrika* 60, pp 255–265.
[12] McLeod, A. I. and W. K. Li 1983 *Journal of Time Series Analysis* 4, pp 269–273.
[13] Ljung, G. and G. Box 1978 *Biometrika* 65(2), pp 297–303.
[14] Ardebili, A. S. M., M. Hashemi and M. Barough 2015 *European Online Journal of Natural and Social Sciences* Vol 4 No 1 (s), pp 831–844.
[15] Box, G. E. P., G. M. Jenkins and G. C. Reinsel 1994 *Time Series Analysis: Forecasting and Control*, 3rd ed. (New Jersey: Prentice Hall).
[16] Chatfield, C. 2000 *Time-Series Forecasting* (New York: Chapman and Hall).
[17] Cowpertwait, P. and A. Metcalfe 2009 *Introductory Time Series with R* (New York: Springer).
