Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Reliability and Delay Modelling of Chhatrapati Shivaji Maharaj International Airport, Mumbai

Priyanka Paithankar[a]*,Gopal R. Patil[b]

[a]*Graduate Student, Indian Institute of Technology Bombay, Mumbai-400076, India*
[b]*Associate Professor, Indian Institute of Technology Bombay, Mumbai-400076, India*

Abstract

Aviation is an integral part of the global transportation system and is at the vertex of the cyclic system of the global economy. Globalisation has facilitated the speedy evolution of air passengers and air freight flows in India which should be supported by the increased capacity of airports. However, airport capacity cannot keep on increasing as per the rate of air transport demand. Hence airports often suffer capacity-demand imbalance at peak hours resulting in airport congestion and flight delays. Airport efficiency can be increased by optimising the airport schedules or by performing dynamic flight scheduling after identifying the slack time between the consecutive flights beforehand. For successful implementation of such an approach, it is imperative to diagnose the hourly delay patterns at every stage of flight phase. Thus, study focusses in characterising delay patters in arrival activities at Chhatrapati Shivaji Maharaj International Airport, Mumbai, India and modelling the arrival and airborne delays using statistical regression, probability distribution and machine learning methods.

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY.

Keywords: Delay modelling, departure delays, airborne delays, arrival activities.

1. Introduction

Globalisation has facilitated the speedy evolution of air passengers and air freight flows which should be supported by the increased capacity of airports. However, airport capacity cannot keep on increasing as per the rate of air transport demand. Hence it is often seen the capacity-demand imbalance at peak hours which directly result in airport congestion and flight delays. Flight delays have many negative impacts on passengers, airlines, and the airport itself. Due to the uncertainty of the events, people tend to travel hours before their scheduled time of commitment to ensure the early arrival. Airlines on the other side, endure extra operation cost and penalties. Moreover, considering

* Corresponding author. Tel.: 9082467456;
E-mail address: p.v.paithankar2525@gmail.com

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY

sustainability, delays also cause environmental harm due to the fact that it increases fuel use and hence the emission of greenhouse gases which have a huge impact on human health, human welfare and ultimately on the environment. This situation can be resolved either by using space available more effectively or increasing the airport capacity. For successful implementation of such an approach, it is imperative to diagnose the hourly delay patterns at every stage of flight phase. These delays can be modelled to predict probable future delays at the airport.

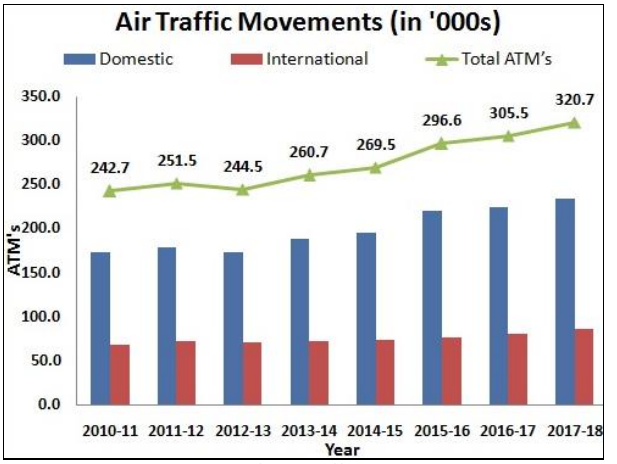
Method of approach to model the delays has purely based the objectives of delay modelling. Some of the popular methods are probabilistic models, statistical models, simulation models and machine learning from which statistical tools are seen to be most common among the researchers One of such study was done by Abdel-Aty et al. (2007) who identified the periodic patterns in arrival delay for domestic flights during 2002-2003 at Orlando International Airport and also studied its cause factors. Mueller and Chatterji (2002) characterised delay distributions and modelled delay vs time probability density functions with normal and Poisson distributions for 21 days of data at selected airports. Wu et al. (2018) studied the extend interrelationship of arrival delays against their departure delay at origin airport using copula function. Tu et al. (2008) presented models which estimate individual flight arrival and airport delays at Orlando International Airport, USA and notify future breakdowns by identifying the patterns in airport delays and arrival delays using logistic regression, neural networks and ANOVA. Yuan (2007) analysed reliability modelling for departure delay distribution and related cost using computer-aided numerical simulation to achieve flight schedule optimisation and to enhance dispatch reliability of Australian Airlines fleet. Jose et al. (2018) presented a network model which considers spatial as well as temporal delay phases as explanatory variables and developed a model which predicted departure delays for next 24 hours using Random Forest Algorithm. Rashmi Vane (2016) analysed on-time performance of airlines schedules at Indira Gandhi International Airport, Delhi, India. Data was collected for the year 2015 and usability of the scheduled performance in reducing the delays were assessed. Tu et al. (2008) focused mainly on departure delays and developed a model for estimation of these delays which is a requirement of any basic prediction models for air traffic congestion. The model analyses seasonal and daily trends with the help of nonparametric methods and adopts mixture distribution to evaluate the residual errors. In "CS229 Final Report : Modeling Flight Delays" (2008), flight information and corresponding weather data of 40 largest airports of USA were used for prediction of flight delays (delay more than 15 min) and tried to capture second-order behaviour of consequent flight due to already delayed previous flights.

The aviation industry in India has grown strongly over recent decades and predicted to continue to grow well in coming future years due to increased travels and tourism contribution to GDP which expected to grow by 12.79% of growth rate compounded annually. India is the ninth largest civil aviation market in the world while it has ranked fourth in domestic passenger volume (80.16 million) as per the financial year 2016("Association of Private Airport Operators" 2008.).Aviation market in India is expected to become the third largest market by 2020 and largest by 2030. Passenger traffic in 2016 was 106.45 million according to reports by World Travel and Tourism Council, Airport Authority of India and is expected to grow up to 421 million by 2020. The civil aviation sector has shown growth of 13.8% in the last 10 years and is predicted to continue well in the future. It was predicted that CSMIA in the year 2015 had handled passenger traffic of 36.6 million and cargo movement of 0.7 million tons. ("Association of Private Airport Operators" 2008.). So, the scope of this study is Chhatrapati Shivaji Maharaj International Airport, Mumbai, India. Objectives of the paper are to assess the on-time arrival, departure performance, identify patterns in flight delays and to model the arriving, departing delay patterns using probabilistic density distributions, logistic regression and machine learning methods. In this study, Mumbai airport is used alternatively for CSMI airport.

2. Study Area

The Chhatrapati Shivaji Maharaj International (CSMI) Airport which is located in Mumbai, Maharashtra, formerly known as Sahar International Airport is one of the largest aviation hubs in country typically handles 34,993,738 passengers per year and ranks second in the list of busiest airports in India, 14th busiest airport in Asia and 29th busiest airport in the world. reports by Airport Authority of India. In this study considers arrivals and departure activities at both T1 and T2 terminals of CSMIA together with only encountered air traffic movement. Arrival and departure at an airport together are defined as air traffic movement as per given by international Civil Aviation Organization (ICAO). In case of air traffic movement, arrival and departure of single aircraft is considered as two movements.

Fig. 1. (a) Temporal representation of Mumbai Airport Traffic



(b) Graphical Representation of Air Traffic Movements Statistics at CSMIA

Domestic air traffic movement share of CSMIA as compared to all other airports in India in the financial year 2017-18 was 12.4% while in case of international movements, share percentage is 19.7. Thus overall 13.8% of total air traffic in India is handled by CSMIA alone as shown in Table 1.

Table 1 Chronological Statistics Air traffic movements at CSMIA (source: www.apaoindia.com)

(in '000s)	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18
Domestic	174	179.3	173.3	188.3	195.4	220.3	224.9	234.6
International	68.7	72.2	71.3	72.4	74.1	76.4	80.6	86.1
Total ATM's	242.7	251.5	244.5	260.7	269.5	296.6	305.5	320.7
Growth Y-o-Y (%)	5.6	3.6	-2.8	6.6	3.4	10.1	3	5
% of Air Traffic Movement handled by MIAL in comparison with all Airports								
Domestic	15.9	14.5	14.9	15.7	15.5	15.5	13.6	12.4
International	22.9	23.3	22.7	21.5	21.5	20.4	20.1	19.7
Total	17.4	16.3	16.5	17	16.8	16.5	14.9	13.8

3. Preliminary analysis

Real-time arrival information of all the flights arriving at CSMI airport, Mumbai on both the terminals T1 and T2 was collectively collected from the site https://www.mumbaiairport.com/arrivals.php for seven days picked from month of June, July and September. An average number of flights both domestic and international arriving at the airport was found to be around 750 out of which hardly 20 flights of the total sum of flights were found to be diverted to other airports for landing. Since the data count for such situation is statistically insignificant, this case is neglected from the analysis. Total 5163 flight data have been considered in the study. Raw data collected includes information about all the flights arrived at CSMI airport, Mumbai on selected days. Figure 2 gives overall outline of total flight numbers, number of flights arriving on time (early arrival or arrival delay equal to or more than 15 minutes) and also delayed arrivals (arrival delay more than 15 minutes) status for all selected seven days. For example, first coloured bar shows the total number of arrivals on that day; second colour bar shows a number of flights arrived early or having arrival delay time equal to or less than 15 minutes and the last bar shows the number of flights delayed more than 15 minutes on the same day.

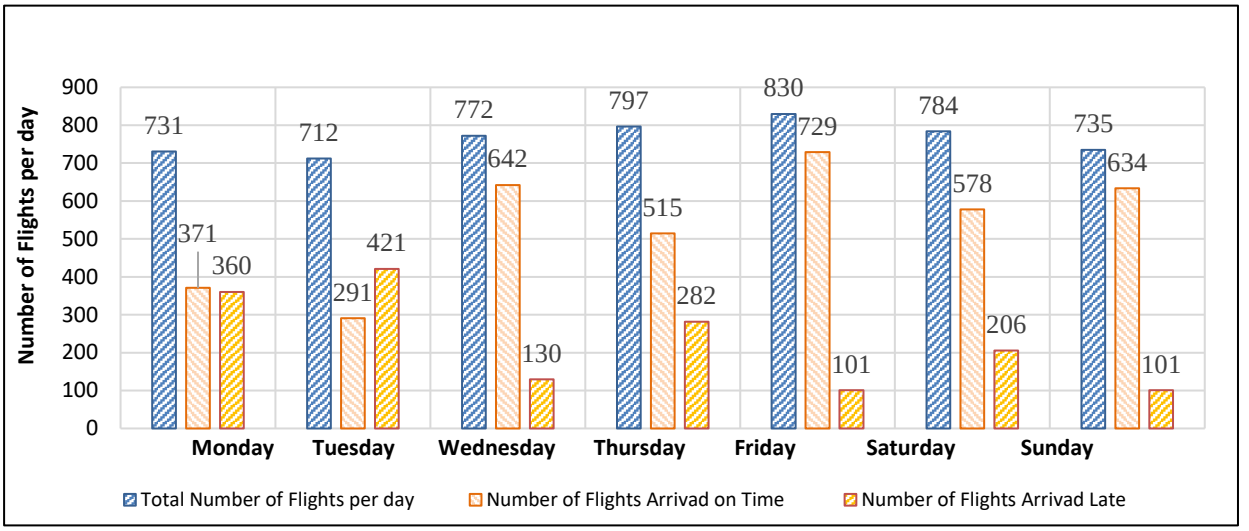


Fig. 2. Flight Data on The Dates of Collection

In our data, the delay is given in minutes wherein flights having a delay equal to or less than 15 min are considered as non-delayed while those getting delayed more than 15 minutes are considered as delayed. Here response variable is arrival delay and explanatory variables contains quantitative variables as scheduled hourly arrival flow rate, scheduled flight time, departure delay at origin airport and qualitative variables which include the terminal number, the day of the week on which flight arrived and interval of time in which it arrived. Arrival terminal shows whether flight landed at Terminal 1 or Terminal 2. Day hours have been divided into 8 intervals each comprises of 3 hours. Arrival delay is calculated deducting actual arrival time from scheduled arrival time. In this study, arrival delay is represented in minutes. In the figure 3 delays are grouped in the interval of 10 mins and is plotted against the percentage of flights experiencing corresponding delays.

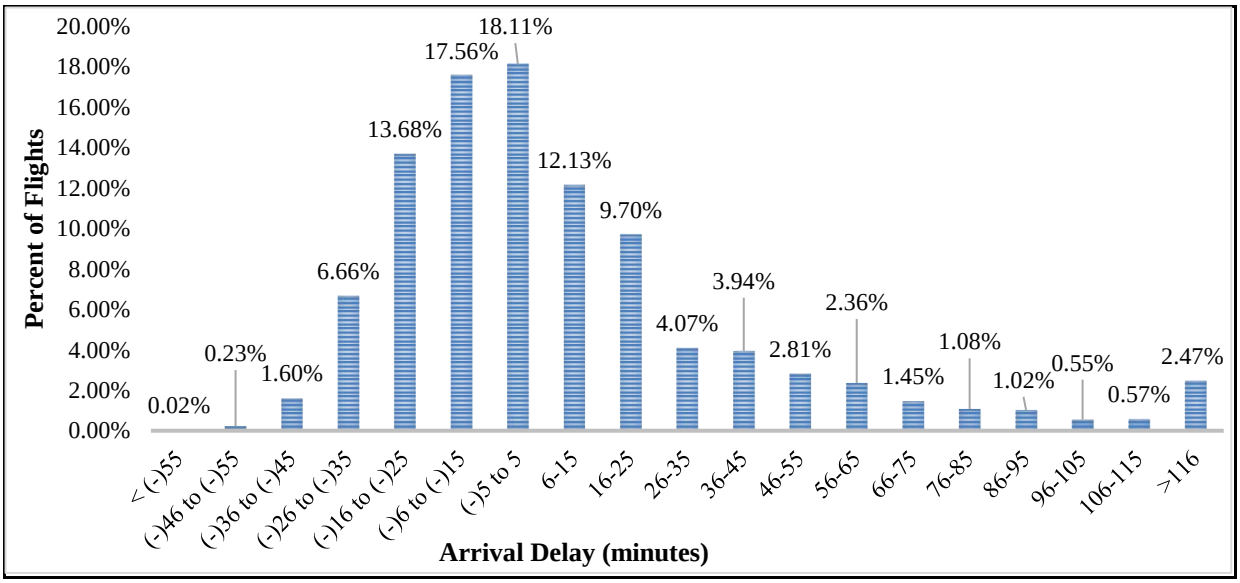


Fig. 3. Arrival Delay Histogram

Since we have defined delayed flight as flight having delay more than 15 minutes, it is imperative to understand the cumulative percentage of flights getting delayed by more than 15 minutes. To analyse this case, the arrival delays are plotted against the cumulative percentage of flights experiencing the corresponding delays. It is to be noted that

50 % of the total flights had an early arrival or on-time arrival at the airport. Almost 20% flights had delay ranging between [0 minutes – 15 minutes] of arrival delay. As per the definition, 30% of the flights experienced arrival delay more than 15 minutes as shown in figure 4.

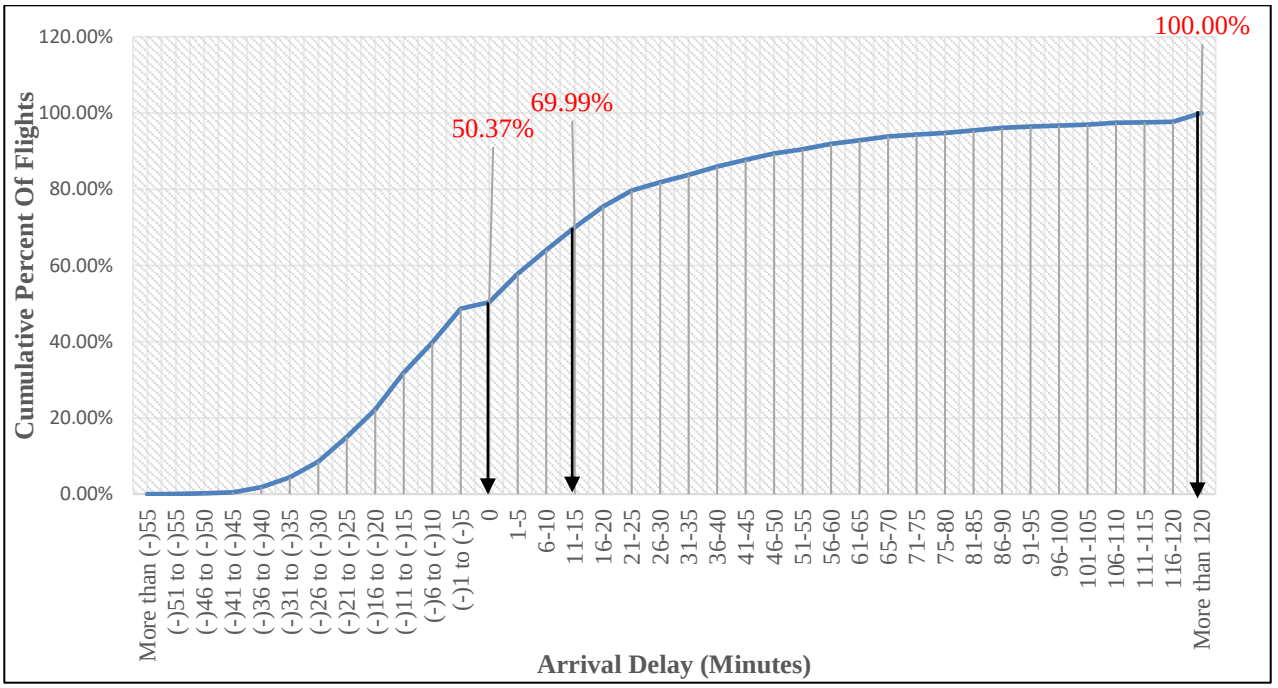


Fig. 4. Cumulative Percent of Flights Vs Arrival Delay

To characterize hourly variations in flights getting delayed more than 15 minutes, the clockwise frequency of delayed flights is shown in following figure 4. For example, between 23 hours to 00 hours, the highest number (243) of flights had got delayed more than 15 minutes followed by 119 between 00 hours to 01 hours. It is noticeable that the lowest frequency of delayed flights and hence the congestion was found least between 05 hours to 06 hours as shown in figure 5.

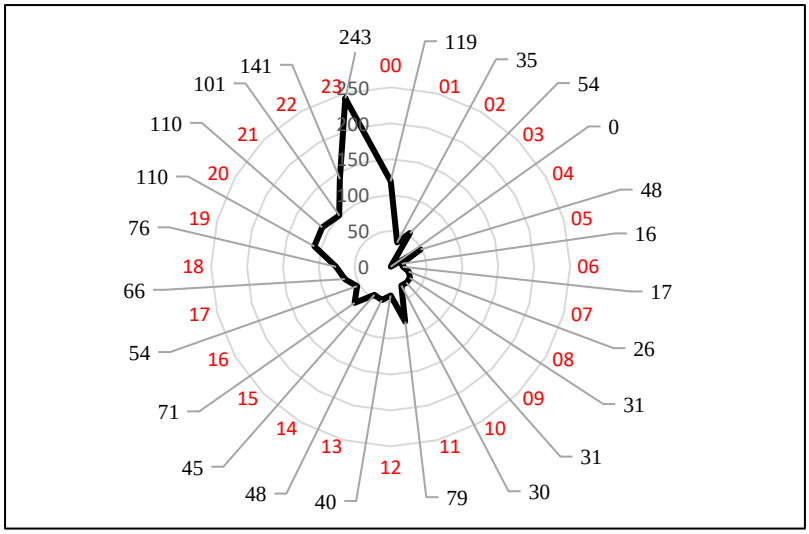


Fig. 5. Hourly Variation of Delayed Flights

4. Probability Distribution for Arrival and Airborne Delays

4.1. Probability Distribution for Arrival Delays

In a probability distribution delays are modelled against time probability density function. To model delays, probability density functions are developed by making use of arrival and departure data collected at the airport. Comparing critical statistics of Kolmogorov Smirnov, Anderson Darling, Chi-Squared, it has been found that, General Extreme Value is suitable and a better fit for the given data than others. In this study Gen. Extreme Value distribution is used to model arrival delays.
GEV (β, μ, k) density function is given as follows;

$$f(x) = \frac{1}{\beta}\left(1 + k\frac{x-\mu}{\beta}\right)^{-1/k-1} . \exp\left(-\left(1 + k\frac{x-\mu}{\beta}\right)^{-1/k}\right) \qquad \ldots\ldots\ldots\ldots\ldots\ldots \beta > 0$$

Total observed arrival delay considered in this study is 4801 with mean 10.937 and standard deviation of 42.195 for which General Logistic distribution has given values of estimated parameters k as 0.163, β as 22.788 and μ as -6.862 which is calculated using maximum likelihood method. The given model has a degree of freedom 30 and observed chi-square value has found to be much more than critical value as a result p-value is much less than 0.0001 for significance level 0.05. Figure 6 gives a graphical representation of observed and theoretical frequencies. It can be seen that Gen. Extreme Value distribution is fitted suitably for given raw data. Figure 7 shows cumulative relative frequency of both raw data and fitted distribution which moderately contradicts for delays more than 20 minutes with a cumulative probability of 0.76.

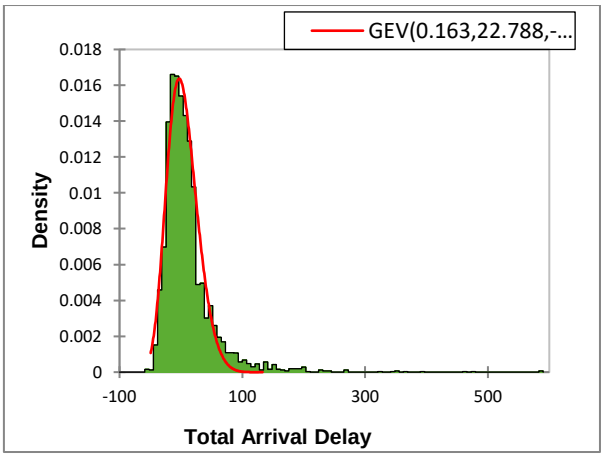


Fig. 6. Observed Frequency Distribution and Model Fit

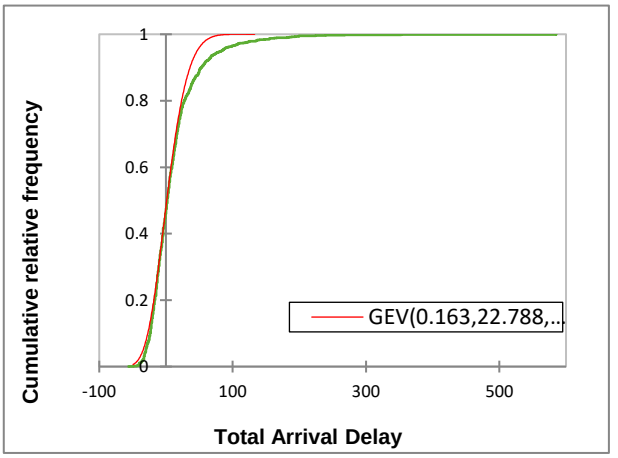


Fig. 7. Cumulative Histogram of observed and theoretical frequencies

4.2. Probability Distribution for Airborne Delays

Comparing critical statistics of Kolmogorov Smirnov, Anderson Darling, Chi-Squared, it is clear that, Logistic distribution is suitable and better fit than others and hence Logistic Distribution has been fitted with a significance level of 5% using maximum likelihood method. For given data, μ and s obtained are -2.920 and 264.946 respectively. The probability density function for logistic distribution is given below;

$$f(x, \mu, s) = \frac{e^{-\frac{x-\mu}{s}}}{s(1 + e^{-\frac{x-\mu}{s}})^2}$$

Figure 8 gives a graphical representation of observed and model fit of the delays. It can be seen that logistic distribution is fitted satisfactorily for given raw data. Figure 9 shows cumulative relative frequency of both raw data and fitted distribution which moderately contradicts for delays in a range (-)17 to (-)50 minutes and (-)18 to (-)58 minutes.

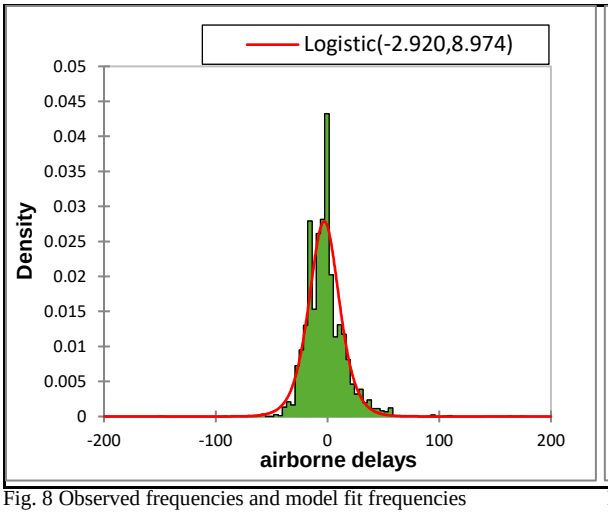


Fig. 8 Observed frequencies and model fit frequencies

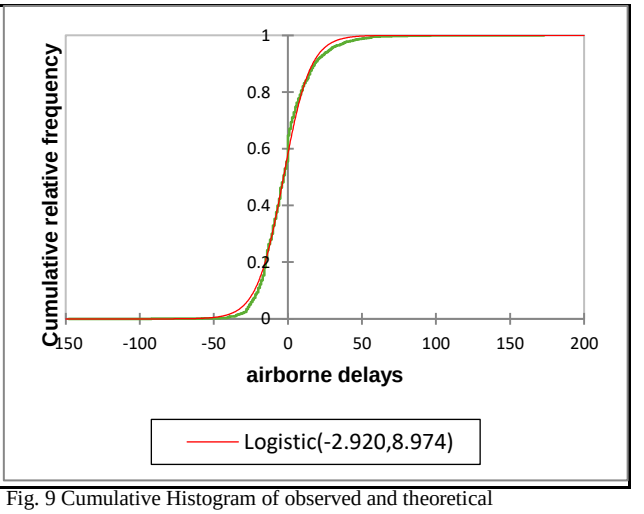


Fig. 9 Cumulative Histogram of observed and theoretical

5. Modelling arrival delays

5.1. ANCOVA Model for Arrival Delays

This model is used when a quantitative variable needs to be modelled using both qualitative and quantitative variables. In this study, quantitative variables considered in this study is departure delay and arrival delay. All these are treated as continuous vector features. In this study, qualitative variables considered are Day of a week, (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday. R^2 was obtained is 0.904, adjusted R^2 of 0.904 which means 90% of the variability in dependent variable 'Total Arrival Delay' is explained by the 2 explanatory variables, 'departure delay' and 'day of a week.

Model Equation

Total Arrival Delay = {-7.268+1.003(Departure Delay) +0.696(Friday) +11.68(Monday) +4.964(Saturday)-4.281(Sunday) +8.299(Thursday) +19.257(Tuesday)}

Monday, Tuesday, Wednesday, Thursday, Friday, Saturday Sunday are dummy variables which will take value either 0 or 1. The variable will take value 1 only if the flight arrives on that particular day else it will take zero for all other cases. Table 2 gives the t statistic and p statistic of the variables included in the model.

Table 2 Model Parameters

Source	Value	t	p-value
Intercept	-7.268	-12.994	< **0.0001**
Departure Delay	1.003	188.973	< **0.0001**
day of a week-Friday	0.697	0.893	0.372

day of a week-Monday	11.682	14.531	< **0.0001**
day of a week-Saturday	4.964	6.261	< **0.0001**
day of a week-Sunday	-4.281	-4.020	< **0.0001**
day of a week-Thursday	8.300	8.994	< **0.0001**
day of a week-Tuesday	19.257	23.084	< **0.0001**

All the days' coefficients are estimated considering Wednesday as base. It is quite important to study the results of variance analysis which allow us to determine if independent variables give significant information or not. In this model the fisher f test is used. Probability corresponding to F value is found to be less than 0.0001 which means we are taking 0.01% risk in concluding that null hypothesis is wrong which tell us that given explanatory variables do bring the important amount of information to the model.

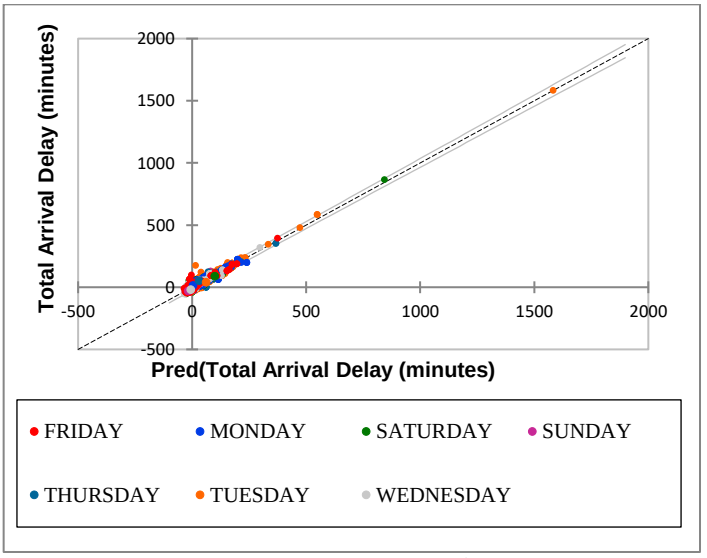


Fig. 10. Predicted values vs observed values of Total Arrival Delay

5.2. Logistic Regression Model for Arrival Delays

The fundamental principle behind the logit model is that it identifies the relation between occurrence or non-occurrence of independent variables. In this study, the model will identify whether the flight will have arrival delay more than 15 minutes or not with confidence level of 95% using Newton-Raphson algorithm for maximisation of the likelihood function. In this model, quantitative variables considered are scheduled hourly flow and departure delay. All of these are treated as continuous vector features while arrival delay has a binary response. Qualitative variables considered are a terminal number (T1, T2), Day of a week, (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday) and scheduled Interval of time in which flight was scheduled to arrive.

Model Equation

Logit model: $$p = \frac{\exp(\beta x)}{1 + \exp(\beta x)}$$

Where, βx comprises of variables as shown in table 3.

Table 3 Coefficient Value for Model Variables

Source	Value	p-value
Intercept	-2.952	< 0.0001
Scheduled Flight Time	-0.003	< 0.0001
Total Departure Delay	0.123	< 0.0001
Terminal-2	0.463	0.001
SIT -2	0.263	0.306
SIT -3	-0.665	0.004
SIT -4	-0.013	0.949
SIT -5	-0.314	0.167
SIT -6	-0.555	0.008
SIT -7	0.639	0.001
SIT -8	0.392	0.021
day of a week-MONDAY	1.824	< 0.0001
day of a week-SATURDAY	0.932	< 0.0001
day of a week-SUNDAY	-0.302	0.295
day of a week-THURSDAY	1.426	< 0.0001
day of a week-TUESDAY	2.454	< 0.0001
day of a week-WEDNESDAY	0.284	0.158

Where SIT is Scheduled Interval of Time in which flight arrived wherein SIT-1 and terminal-1 are taken as base. Monday, Tuesday…. Sunday are dummy variables which will take value either 0 or 1. The variable will take value 1 only if flight arrives on that day else it will take zero for all other cases. Similar approach has been applied for a terminal number and scheduled interval of time (SIT) coefficients of which are taken

The model considers 4280 observations with a degree of freedom of 4263 have R^2 value of 0. 522. The model has successfully passed chi-square on the log ratio test which is equivalent to Fisher's F test which tells if variables give significant information. In this model as the probability is less than 0.0001, it can be concluded that variables brought significant information to the model. Table 4 shows a number of miss-classified and well-classified observations. Wherein sensitivity is coming out as 93.85% while specificity is 71.89% and, on an average percentage of well-classified observations is 86.59%. Figure 7 Depicts Receiver Operating Characteristic which is a curve of points (1-specificity, sensitivity) and gives information about the performance of the model with the help of AUC and can be used to compare between models. Sensitivity is a share of well-classified positive occurrences while specificity is a share of well-classified negative occurrences. AUC is coming out to be 0.93 which means the model is excellent.

Table 4 Classification table for the training sample

from \ to	0	1	Total	% correct
0	2688	176	2864	93.85%
1	398	1018	1416	71.89%
Total	3086	1194	4280	86.59%

Category 0 represents arrival delay less than or equal to 15 minutes while category 1 represents arrival delay more 15 minutes

5.3. Support Vector Machine (SVM)

This method is used when the response variable is classified into binary variable explained by both quantitative and qualitative variables. It always targets to identify a separation between two class of an object assuming that more the separation, better and reliable is the classification. It makes use of sequential minimal optimisation algorithm which breaks the problem into smaller sub-problems and solves them analytically due to which computational burden reduces dramatically. In this model, explanatory variables considered are scheduled hourly flow, Scheduled flight time and departure delay. All these are treated as continuous vector features while arrival delay has a binary response. Qualitative variables considered are a terminal number, scheduled Interval of time, actual Interval of time and day of a week. All these are treated as categorical features while arrival delay has a binary response. Model has considered 4280 observations out of which 1943 lie close set to hyperplane or decision plane with the bias of -2.015 that means hyperplane does not go through origin which means it does not have maximum margin. Table 10 depicts how well the model gave the results. Wherein sensitivity is coming out as 97.00% while specificity is 67.86% and, on an average percentage of well-classified observations is 85.37%. AUC obtained in this model is 0.921 with an overall accuracy of 0.854 and precision of 0. 837.

Table 5 Performance Matrix (Total arrival delay = 0 / 1)

from \ to	0	1	Total	% correct
0	2778	86	2864	97.00%
1	540	876	1416	61.86%
Total	3318	962	4280	85.37%

Category 0 represents arrival delay less than or equal to 15 minutes while category 1 represents arrival delay more 15 minutes

6. Summary and Conclusion

Speedy increase in air passenger and freight transportation through air is causing airport congestion and flight delays since airport capacity cannot cope up with the rate of increase in demand and this problem is going to be severe in future. Thus, to increase airport efficiency, the delay patterns have been diagnosed at every stage of flight phase and are modelled using SVM and logistic regression. In this study, data has been collected about flight information and descriptive analysis has been carried out to analyses the daily as well as hourly variation in arrival and airborne delays at CSMIA. Based on the available literature, arrival delays are modelled using probabilistic density function, SVM, ANCOVA and logistic regression. Chi-square test has proved that generalised extreme value distribution is best fit if delays are modelled against the time probability density function and distribution fits very well even for a significance level of 0.01. ANCOVA model has shown R^2 value of 0.904 with a p-value for the whole model is much less than 0. 0001. Logistic regression model has shown an average percentage of well-classified observations as 86.59%. with AUC of 0.930 while SVM method has shown average percentage of well-classified observations as 85.37%, the accuracy of 0.854 with AUC 0.921. In the case of SVM model, high sensitivity has been obtained from the training sample which is 97% but specificity was compromised which is 61%. However, selection of a method for predicting delays will surely depend upon the validation results of these models.

However, these models can be refined further by including the effect of flight distance, time-space between consecutive arrival or departures at the airport, hourly variation in visual visibility, mean precipitation and wind speed. Similar to arrival models, departure delay models for the airport can be developed and its propagation on downstream airport delays can be analysed. The arrival and departure delay models together can be a great use for the analysing capacity-demand scenario at Mumbai airport. Delay predictions can used for dynamic scheduling of flight arrivals in real time. Since the type of airline in the models. The airline's companies can also use these models to by incorporating airline type in the flight information and can improve the on-time performance of their flights.

References

Abdel-Aty, Mohamed, Chris Lee, Yuqiong Bai, Xin Li, and Martin Michalak. 2007. "Detecting Periodic Patterns of Arrival Delay." *Journal of Air Transport Management* 13 (6): 355–61. https://doi.org/10.1016/j.jairtraman.2007.06.002.

"Association of Private Airport Operators." n.d. Accessed October 13, 2018. http://www.apaoindia.com/?page_id=923.

"CS229 Final Report : Modeling Flight Delays." 2008.

Jose, Juan, Hamsa Characterization, Citable Link, Juan Jose Rebollo, and Hamsa Balakrishnan. 2018. "Accessed Characterization and Prediction of Air Traffic Delays" 2007: 1–19.

Mueller, Eric, and Gano Chatterji. 2002. "Analysis of Aircraft Arrival and Departure Delay Characteristics." *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*, no. October: 1–14. https://doi.org/10.2514/6.2002-5866.

Sc, B I T, and M L Dahanukar College. 2016. "Flight Delay Analysis and Possible Enhancements with Big Data," 778–80.

Tu, Yufeng, Michael O. Ball, and Wolfgang S. Jank. 2008. "Estimating Flight Departure Delay Distributions - A Statistical Approach with Long-Term Trend and Short-Term Pattern." *Journal of the American Statistical Association* 103 (481): 112–25. https://doi.org/10.1198/016214507000000257.

Wu, W., C.-L. Wu, T. Feng, H. Zhang, and S. Qiu. 2018. "Comparative Analysis on Propagation Effects of Flight Delays: A Case Study of China Airlines." *Journal of Advanced Transportation* 2018. https://doi.org/10.1155/2018/5236798.

Yuan, Duojia. 2007. "Flight Delay-Cost Simulation Analysis and Airline Schedule Optimization A Thesis Submitted in Fulfillment of Requirement for the Degree of Doctor of Philosophy," no. February.