Analysis of Passenger Flow Prediction of Transit Buses Along a Route Based on Time Series

Reshma Gummadi and Sreenivasa Reddy Edara

Abstract India's transport sector has a prominent role in transportation of passengers. Andhra Pradesh State Road Transport Corporation (APSRTC) is a major transportation in the state. State has many routes, and there are so many towns on a particular route. Most of the population mainly depends on transportation system; hence, it is necessary to predict the occupancy percentage of the transit buses in a given particular period for the convenience of passengers. There is a need of advancement in transportation services for effective maintainability. Identifying passenger occupancies on a different number of buses is found to be a major problem. A promising approach is the technique of forecasting the data from previous history and the better predictive mining technology must be applied to analyze the passenger to predict the passenger flow. In this work, ARIMA-based method is analyzed for studying the APSRTC transit bus occupancy rate.

Keywords Transit buses · Passenger flow · ARIMA

1 Introduction

1.1 Overview

Andhra Pradesh State Road Transport Corporation (APSRTC) recently took an initiative for implementation of information technology in the state of Andhra Pradesh, and the effective use of IT helps in many ways for APSRTC. It helps in providing better services to passengers and also leads to effective maintenance

R. Gummadi (✉)
Acharya Nagarjuna University, Guntur 522510
Andhra Pradesh, India
e-mail: reshma.gorripati@gmail.com

S. R. Edara
Department of Computer Science & Engineering,
Acharya Nagarjuna University, Guntur 522510, Andhra Pradesh, India

S. C. Satapathy et al. (eds.), *Information and Decision Sciences*,
Advances in Intelligent Systems and Computing 701,
https://doi.org/10.1007/978-981-10-7563-6_4

management of vehicles. The role of Information Technology plays a key role in better inventory control and also better managerial controls. In view of that, Ticket Issuing Machines (TIMs) were introduced in APSRTC in May 2000. The main objective is to issue tickets even though ground booking is completed and also to pick up number of passengers in a route. These are introduced in a view that management can derive information like punctuality analysis and travel patterns of the public, generating MIS reports from the database. At present, there are 14,500 TIMs being utilized.

The data collected using TIMs from passengers includes starting stage, ending stage, ticket time, ticket percentage, number of adults, and number of children. The digitization of information helps in so many number of ways to increase the bus services. As the data is already collected, and this data is helpful for further investigations and forecasting the information can be done. By analyzing the data, we can predict the passenger occupancies which are helpful for estimating the flow of passengers at critical period of time also. The predictive analysis must be done by considering all the cases like festivals, weekdays, and normal days. As ticket time is also collected, the estimation of the peak times and proportion of passengers can be done accordingly.

1.2 Objectives

The objective of this paper is to analyze various ARIMA models to implement the estimation of the passenger flow at APSRTC. Based on the data which is collected from APSRTC Head Office, the passenger flow was analyzed with findings and suggestions. In this paper, the modeling methods for passenger occupancies were discussed and the analysis of modeling techniques on real time data is presented. There are so many model-based prediction algorithms like Kalman filtering, regression models, and artificial neural networks. As time plays a key role in this application, an ARIMA analysis model can be used to predict the passenger flow over a period of time. SARIMA model is analyzed as there is seasonality in monthly data for which the high values may tend to occur in few months, and low values may tend to occur in some other months. In this scenario, we consider $S = 12$ is the span of the periodic seasonal behavior. The model can be selected in such a way that total number of passengers to be predicted for a particular month in a year is based on previous year data.

2 Data Source

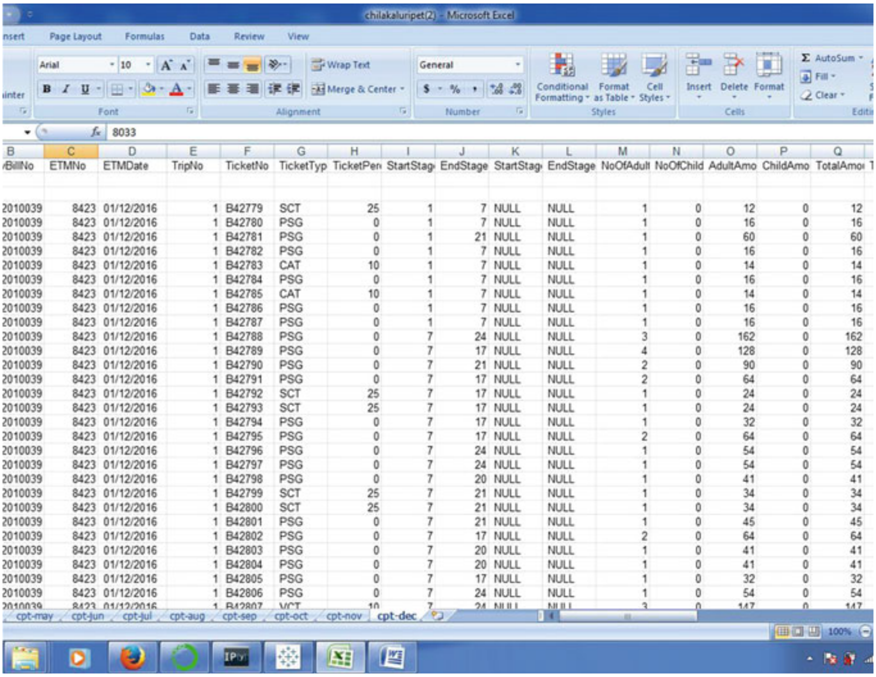
This paper examined novel methods to estimate the passenger occupancies over a period of time. The model dataset used in this paper based on the data which was the primary source is provided by APSRTC on transit buses information.

2.1 Transit Data

APSRTC provides transport services within AP. Among those services, the transit information of Macherla to Chilakaluripet route ticket-wise data of 24 services extracted from three depot databases for the period of April 1, 2016 to December 31, 2016 was given and the dataset snapshot is displayed below.

The dataset consists of various measures and dimensions but for analysis purpose, we choose ETMDate, FromStagecode, ToStageCode as dimensions, and NoOfAdults considered to be a measure.

From Macherla to Chilakaluripet, there are 20 intermediate stations and each station is identified by bus number which is a unique id as shown in Fig. 1. The data given in excel table consists of Date, Trip No, Ticket percentage, Start Stage, End Stage, No of adults, Number of Childs, and corresponding Ticket time given. Among those fields, Date, Start Stage, End stage, and total number of passengers are to be considered for analysis.



8033

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
/BillNo	ETMNo	ETMDate	TripNo	TicketNo	TicketTyp	TicketPer	StartStag	EndStage	StartStag	EndStage	NoOfAdult	NoOfChild	AdultAmo	ChildAmo	TotalAmo
2010039	8423	01/12/2016	1	B42779	SCT	25	1	7	NULL	NULL	1	0	12	0	12
2010039	8423	01/12/2016	1	B42780	PSG	0	1	7	NULL	NULL	1	0	16	0	16
2010039	8423	01/12/2016	1	B42781	PSG	0	1	21	NULL	NULL	1	0	60	0	60
2010039	8423	01/12/2016	1	B42782	PSG	0	1	7	NULL	NULL	1	0	16	0	16
2010039	8423	01/12/2016	1	B42783	CAT	10	1	7	NULL	NULL	1	0	14	0	14
2010039	8423	01/12/2016	1	B42784	PSG	0	1	7	NULL	NULL	1	0	16	0	16
2010039	8423	01/12/2016	1	B42785	CAT	10	1	7	NULL	NULL	1	0	14	0	14
2010039	8423	01/12/2016	1	B42786	PSG	0	1	7	NULL	NULL	1	0	16	0	16
2010039	8423	01/12/2016	1	B42787	PSG	0	1	7	NULL	NULL	1	0	16	0	16
2010039	8423	01/12/2016	1	B42788	PSG	0	7	24	NULL	NULL	3	0	162	0	162
2010039	8423	01/12/2016	1	B42789	PSG	0	7	17	NULL	NULL	4	0	128	0	128
2010039	8423	01/12/2016	1	B42790	PSG	0	7	21	NULL	NULL	2	0	90	0	90
2010039	8423	01/12/2016	1	B42791	PSG	0	7	17	NULL	NULL	2	0	64	0	64
2010039	8423	01/12/2016	1	B42792	SCT	25	7	17	NULL	NULL	1	0	24	0	24
2010039	8423	01/12/2016	1	B42793	SCT	25	7	17	NULL	NULL	1	0	24	0	24
2010039	8423	01/12/2016	1	B42794	PSG	0	7	17	NULL	NULL	1	0	32	0	32
2010039	8423	01/12/2016	1	B42795	PSG	0	7	17	NULL	NULL	2	0	64	0	64
2010039	8423	01/12/2016	1	B42796	PSG	0	7	24	NULL	NULL	1	0	54	0	54
2010039	8423	01/12/2016	1	B42797	PSG	0	7	24	NULL	NULL	1	0	54	0	54
2010039	8423	01/12/2016	1	B42798	PSG	0	7	20	NULL	NULL	1	0	41	0	41
2010039	8423	01/12/2016	1	B42799	SCT	25	7	21	NULL	NULL	1	0	34	0	34
2010039	8423	01/12/2016	1	B42800	SCT	25	7	21	NULL	NULL	1	0	34	0	34
2010039	8423	01/12/2016	1	B42801	PSG	0	7	21	NULL	NULL	1	0	45	0	45
2010039	8423	01/12/2016	1	B42802	PSG	0	7	17	NULL	NULL	2	0	64	0	64
2010039	8423	01/12/2016	1	B42803	PSG	0	7	20	NULL	NULL	1	0	41	0	41
2010039	8423	01/12/2016	1	B42804	PSG	0	7	20	NULL	NULL	1	0	41	0	41
2010039	8423	01/12/2016	1	B42805	PSG	0	7	17	NULL	NULL	1	0	32	0	32
2010039	8423	01/12/2016	1	B42806	PSG	0	7	24	NULL	NULL	1	0	54	0	54
2010039	8423	01/12/2016	1	B42807	VCT	10	7	24	NULL	NULL	3	0	147	0	147

Fig. 1 Snapshot of data given by APSRTC

2.2 Model Evaluation

To analyze the characteristics of historical data, time series plays a prominent role. In the dataset provided by APSRTC, they have given data with minute duration. Time series is chosen because data is available on daily basis. Here, data is in probabilistic in nature. It is practically impossible for 100% accuracy. Let us define $d^n(t)$ the demand of the passengers at time interval $(t - 1)$ for day n time series $T^n{}_m(t)$ consists of data n_m time intervals before $d^n(t)$ on the same day. As data is available in minutes, that should be grouped into day-wise and finally to monthly data. In this case, the previous time intervals to be considered with differences measured at each interval.

$$T^n{}_d(t) = \{d^n(t-1),\ d^n(t-2),\ \ldots d^n(t-n_m)\}$$

Seasonality can be defined in this case as number of passenger's data in months.

2.3 Analyzing Bus Schedule

The model is complex as there is difficulty in planning route definition. Let us consider a service operates with number of scheduled buses per minute duration, i.e., varies from 1 min duration to 15 min duration and these type can be treated as low-frequency services. High-frequency services are more number of buses per hour. For the performance measures to be accurate, reliability of bus location data is the essential starting point. Kalman filtering can be applied for the number of inputs from different sources to establish the connection of the bus.

The starting time of all buses that are in one route segment checks the previous time schedules and then all buses that have covered the same lag over the same period of time in the previous intervals.

For example, prediction is being calculated on 1 day in a month service. The figures for the previous minutes on that route segment would be compared by removing outliers from the dataset, and forecasting the passengers is derived and added to the current number of passengers to provide total number of passengers for that segment. This is potentially calculated with each bus location for each bus stop. Trip-wise pattern analysis checks whether the current trip has similar pattern as that of previous patterns; on the same day, Z-test can be applied in this case and by identifying positive skewness and negative skewness, it would give results for pattern analysis.

Graph is plotted for number of passengers that represents rows and from stage represents columns as shown in Fig. 2. The data over a particular route from each stage is plotted. Number of passengers observed is increasing over particular stages and decreasing over some stages. Similarly, after analyzing 42000 records,

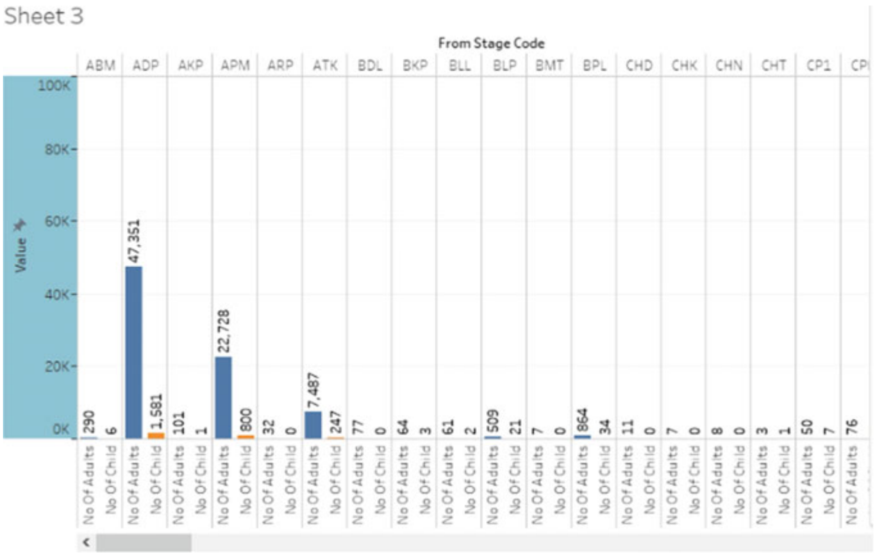


Fig. 2 Number of passengers on a particular stage

predicting the number of passengers along route from start stage to end stage for every month is to be done.

The data can be extracted month-wise, and for each month the amount can be calculated. In a similar way, the number of passengers can be calculated month-wise which is shown in Fig. 3. Once the data is plotted in monthly manner, ARIMA modeling can be applied to predict the passenger flow for each month in the next year. Hence, from next year number of buses to be increased or decreased can be known by forecasting the number of passengers.

2.4 ARIMA Modeling

ARIMA modeling can be applied to the available data. At the initial step, identify whether the variable which is to be predicted is stationary in time series or not; if the variable is not stationary, it should be made stationary using the following process. Let x_t denote the number of passengers at time t; the proposed general ARIMA formulation is as below.

Let $\{x_t|t$ Î $T\}$ denote a time series such that $\{w_t|t$ Î $T\}$ is an **ARIMA (*p, q*)** time series where $w_t = D^d x_t = (I - B)^d x_t =$ the dth order differences of the series x_t. Then, $\{x_t|t$ Î $T\}$ is called an **ARIMA (*p, d, q*)** time series (an **integrated autoregressive moving average** time series); here, B denotes the backshift operator which is used

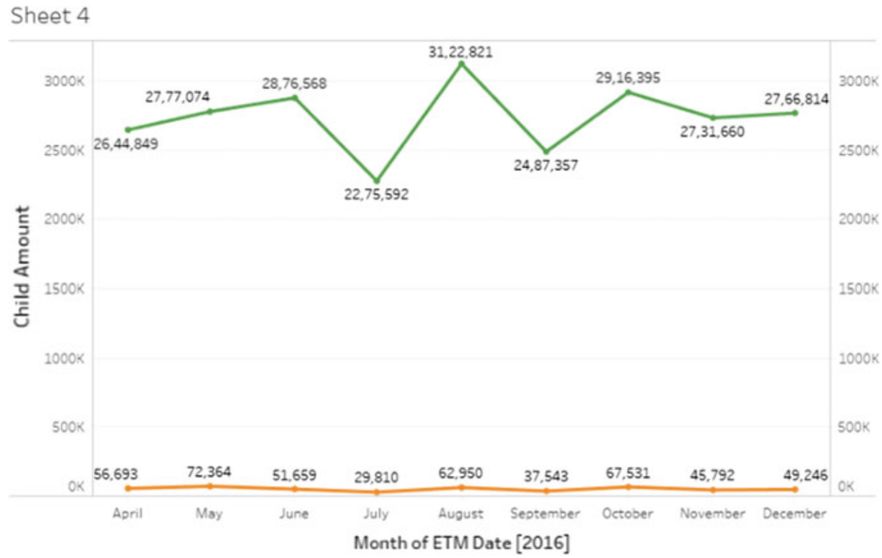


Fig. 3 Monthly report of passengers

to forecast the values based on previous values; p denotes the autoregressive parameters, d is the number of differencing passes, whereas q represents the moving average which was computed for the series after it was difference once. Input series of ARIMA must have a constant mean, variance, and autocorrelation through time.

The equation for the time series $\{w_t | t\ \hat{I}\ T\}$ is $b(B)w_t = d + a(B)u_t$. Suppose that d roots of the polynomial $f(x)$ are equal to unity, then f(x) can be written as $f(B) = \left(1 - b_1x - b_2x^2 - \cdots - b_px^p\right)(1-x)^d$, and f(B) could be written as $f(B) = \left(I - b_1B - b_2B^2 - \cdots - b_pB^p\right)(I-B)^d = b(B)D^d$. In this case, the equation for the time series becomes $f(B)x_t = d + a(B)u_t$. In **ARIMA (1,1,1)** $x_t = (1+b_1)x_{t-1} - b_1x_{t-2} + d + u_t + a_1u_{t-1}$. If a time series is $\{x_t$: $t\ \hat{I}\ T\}$, that is, seasonal we would expect observations in the same season in adjacent days to have a higher autocorrelation than observations that are close in time (but in different seasons in this case seasonality can be defined as month estimated for 1 year). This model satisfies the equation incorporating seasonality.

$$\left(I - B^k\right)^{ds}(I-B)^d\beta^{(s)}(B)\beta(B)x_t = \alpha^{(s)}(B)\alpha(B)u_t + \delta$$

Forecasting an ARIMA (p, d, q), time series can be done by the following process:

$$x_T(l) = E(x_{T+l}|P_T)$$

Let P_T denote $\{\ldots, x_{T-2}, x_{T-1}, x_T\}$ = the "past" until time T. Then, the optimal forecast of x_{T+l} given P_T is denoted by this forecast that minimizes the mean square error.

For the given data, there is a need of calculating the first difference, log first difference, seasonal, log seasonal, and seasonal first difference which apply SARIMA modeling with order SARIMA (0,1,0) * (0,1,1,12) and SARIMA (0,1,0) * (2,1,1,12) from which state-space model results can be obtained. The goodness of fitted value can be known by observing AIC value. The modeling is selected for lower AIC value. At the next step, the predicted values can be obtained by applying the above-selected model and calculates the new values of the series.

3 Conclusion

This paper has mainly focused the analysis of passenger flow prediction of transit buses based on time series. APSRTC dataset was used for analysis. ARIMA modeling is analyzed. Based on the investigated simulation and analysis report, comprehensive study was done. This information helps in improving the service efficiency and reduction of waiting time of passengers due to lack of buses when there is more passenger flow. The analysis carried out in this work will benefit the public transport authority in helping them to plan accordingly to place number of buses over a particular time span to reduce the overcrowding or under crowding. As a future research, other forecasting methods available in the literatures can be experimented.

References

1. Al-Deek, H., D'Angelo, M., Wang, M.: Travel time prediction with non-linear time series. In: Proceedings of the ASCE 1998 5th International Conference on Applications of Advanced Technologies in Transportation, Newport Beach, CA, pp. 317–324 (1998)
2. Chien, S.I.J., Kuchipudi, C.M.: Dynamic travel time prediction with real-time and historic data. J. Transport. Eng. **129**(6), 608–616 (2003)
3. Bin, Y., Zhongzhen, Y., Baozhen, Y.: Bus arrival time prediction using support vector machines. J. Intell. Transport. Syst. **10**(4), 151–158 (2006)
4. Chen, S. et al.: The Time Series Forecasting: From the Aspect of Network. arXiv:1403.1713 (2014)
5. Vagropoulos, S.I., Chouliaras, G.I., Kardakos, E.G., Simoglou, C.K., Bakirtzis, A.G.: Comparison of SARIMAX, SARIMA, Modified SARIMA and ANN-based Models for short-term PV generation forecasting. In: IEEE International Energy Conference, Leuven pp. 1–6 (2016)
6. Takaomi, H., Takashi, K., Masanao, O., Shingo, M.: Time series prediction using DBN and ARIMA. In: International Conference on Computer Application Technologies, IEEE (2015)