 doi:10.20944/preprints202210.0112.v1

Article

Intelligent Passenger Frequency Prediction for Transportation Sustainability using Kalman Filter Algorithm and Convolutional Neural Network

Onemayin David Jimoh[1], Lukman Adewale Ajao[*2], Oluwafemi Oyetunde Adeleke[3], Stephen Sunday Kolo[1] and Oyedeji Abdulwaheed Olarinoye[1]

[1] Department of Civil Engineering, Federal University of Technology, Minna, 92011, Nigeria
[2] Department of Computer Engineering, Federal University of Technology, Minna, 92011, Nigeria
[3] Department of Civil Engineering, University of Ilorin, Ilorin, 92011, Nigeria
*Corresponding author: ajao.wale@futminna.edu.ng

Abstract: The passenger prediction flow is very significant to transportation sustainability. This is due to some chaos of traffic jams encountered by the road users during their movement to the offices, schools, or markets at earlier of the days and during closing periods. This problem is peculiar to the transportation system of the Federal University of Technology Minna, Nigeria. However, the prevailing technique of passenger flow estimation is non-parametric which depends on the fixed planning and is easily affected by noise. In this research, we proposed the development of a hybrid intelligent passenger frequency prediction model using the Auto-Regressive Integrated Moving Average (ARIMA) linear model, Convolutional Neural Network (CNN), and Kalman Filter Algorithm (KFA). The passengers' frequency of arrival at the bus terminals is obtained and enumerated through the closed-circuit television (CCTV) and demonstrated using the Markovian Queueing Systems Model (MQSM). The ARIMA model was used for learning and prediction and compared the result with the combined techniques of using CNN-KFA. The autocorrelation coefficient functions (ACF) and partial autocorrelation coefficient functions (PACF) are used to examine the stationary data with different features. The performance of the models was analyzed and evaluated in describing the short-term passenger flow frequency at each terminal using the Mean Absolute Percentage Error (MAPE) and Mean Squared Error (MSE) values. The CNN-Kalman-filter model was fitted into the short-term series and the MAPE values are below 10%. The Mean Square Error (MSE) shows that the CNN-Kalman Filter model has the overall best performance with 83.33% of the time better than the ARIMA model and provides high accuracy in forecasting.

Keywords: ARIMA; convolutional neural network; Kalman filter; passenger flow; transportation; short-term prediction; stochastic model

1. Introduction

Transportation sustainability requires adequate planning for the traffic flow in urban centers, where bus shuttles are used as a public means to convey passengers from one location to another [1]. The passenger traffic flow is described as prevalent interactions between the travelers (passengers and vehicles) and the infrastructure (road networks, traffic control devices, or signage), for the optimal transport network in managing the traffic congestion problems [2, 3]. Traffic flow management can be achieved efficiently through timetable scheduling for passenger movement. This process will enhance decision-making by the transport manager in forecasting passenger flow and efficiently controlling the bus dispatch [4-6]. Understanding travel behavior characteristics of a university community will assist greatly in the smooth running of the inter-campus transportation sustainability, mobility factors, and control of the traffic flow [7-9].

This flow of passengers could be long-term or short-term, and prediction of this traffic flow is of great value in a diverse area of applications, including smart city, urban road network planning, city traffic monitoring, and the tracking of city-taxi movement [10-12]. A long-term passenger flow is not significant due to the close-fitting appearance of the patronage, and its prediction duration can take days, weeks, or longer [13]. While the short-term passenger prediction can be investigated within a short period and is more problematic due to the frequent patronage and congestion at the bus terminal [14].

The short-term passenger flow requires a real-time information system that will help in monitoring, predicting, and scheduling the public passenger flow in the transportation system [15, 16]. This transport scheduling system (TSS) could be intelligent by integrating automatic vehicle location, fare collection, passenger counters, route monitoring, and time and distance measurements for accurate timetable preparation [17]-[19]. The randomization variation of short-term passenger flow makes it tough to analyze all its features with a single time-series model [20]. This challenge leads to the use of hybrid models that involved linear and non-linear for better performances [21]. Although, there are attempts to combine linear and nonlinear models in forecasting short-term passenger flow the structures of such models become complex [22]. While a single model performs better in the prediction analysis flow of the short-term passenger with reasonable prediction outcomes and low cost using the linear regression technique [23]. The linear time series analysis approach can be statistical or descriptive which depends on the time domain or frequency domain analysis techniques [24].

This research justifies the adoption of the combined CNN-KFA model of big data technology and linear approach in the study. The Convolution Neural Network (CNN) is used as a learning model to reduce the error rate and improve the speed during the prediction stage using Kalman Filter Algorithm (KFA). This technique of CNN-KFA is evaluated and compared with a single linear ARIMA model for times series short-term passenger frequency flow prediction. This self-motivated learning model distinguishes between phenomena, and nominal, and deduces the state of knowledge for nominal from the phenomena [25].

The research focuses on the passenger flow prediction analysis for a sustainable transportation system between the inter-campuses arterial route of the Federal University of Technology, Minna, Nigeria using ARIMA and the CNN-Kalman Model. This research contributes to knowledge by:

1. Using a markovian model to analyze the passenger flow frequency obtained through the closed-circuit television (CCTV) at the bus terminals for complete 5 days.
2. Using convolution neural network (CNN) to extract, preprocess, and training of the data for simplicity and to reduce error.
3. Using ARIMA and CNN-KFM model for analyzing the passenger flow frequency
4. The ACF and PACF are used to investigate the stationary data with different features.
5. Using MAPE and MSE to evaluate the prediction model performance.

The research section is organized as follows. The related works are presented in section 2. Section 3 presented the methods and materials, site survey, and data collection process are presented in 3.1, which includes the model estimation of passenger arrivals, traffic flow monitoring processes, and the estimation of the passenger traffic flow frequency using Continuous-Time Markov Process. Section 3.2 presented a prediction model of passenger frequency flow using the ARIMA model, Kalman filter, and Convolution Neural Network technique. Section 4 discussed simulation results obtained from the study of the passenger frequency flow and the models for prediction using ARIMA and Kalman filter techniques. Section 4 is summarized with the performance analysis using MAE, MAPE, and others as evaluation metrics. Section 5 conclude the research with future research recommendation.

2. Related Works

The short-term passenger flow prediction called for technical and intelligent approaches in the public transportation planning due to the persistent challenges of passenger overflow at bus terminals. Several efforts have been put together to optimize the stochasticity of traffic flows condition, passenger operations demand, and timetable schedule. Gu et al. proposed ARIMA model for the prediction of passenger frequency in the Shanghai station [26]. The observed historical events were taken within five weeks with a sampling interval of 10 minutes. It was reported that the time series has a cyclical and slow attenuation trend. The cyclical and trend phenomena were removed to obtain stationary series. The selected ARIMA (2,1) is used to predict the frequency of passengers at the station with an accuracy of 80%.

The use of Support Vector Machine (SVM) with Deep learning (DL) theory for passenger flow prediction was proposed [27]. The system outperformed three other shallow prediction models from Qingdao metro experiments that were carried out. The system has a good depth structure which allowed for optimum performance with data sets that had a lot of randomnesses. However, the model cannot efficiently handle a dataset with a lot of noise. Guo et al. developed a passenger frequency prediction model using a merging of support vector regression (SVR) and long short-term memory (LSTM) neural network techniques [28]. The passenger frequency information training was conducted using the LSTM model to train and show the large unsteadiness of irregular flow and its approximation. A fusion method based on the real-time prediction errors of the SVR outputs and LSTM is combined into the final computation of the prediction model. The hybrid model performance was evaluated using MAPE, RMSE, and MAE with average values of (12.59%, 68.54, and 40.06) respectively for a one-step-ahead process. The two-step ahead processes evaluation was recorded as (14.84%, 80.12, and 47.58) respectively. The comparative performance analysis shows that the SVR-LSTM model precisely mirrors the irregular unsteadiness of passenger frequencies, which performs well and produces greater forecast precision than individual models.

Liu et al. proposed the hybridization of Kernel Extreme Learning Machine (KELM) with wavelet transform (WT) for the development of a passenger flow prediction system [29]. The system achieved an appropriate breakdown of data about passenger flow in both high and low-frequency sequences with an MSE of 2982.277. Zhang et al. use a K-Means clustering 2-step model for obtaining volume passenger flow characteristics and LSTM for the prediction of short-term passenger flow [30]. The system achieved efficient aggregate passenger flow. However, the model considered only local spatial correlations due to the LSTM network method adopted. Cats and Glück, integrate a dynamic model for the strategic phase frequency and vehicle capacity determination [31]. The frequencies and vehicle capacities consideration have significant consequences on the transport service and operational costs. The dynamic transit model estimate variations in service headways, crowding and the passenger distribution flow effect.

Zhang et al. presented the combination of Residual Network (ResNet) for spatial correlation capture and temporal correlations using LSTM and Graph Convolutional Network (GCN) [32]. This approach was used for the extraction of network topology information in the prediction of passenger flow based on the short-term. The system achieved a Mean-Absolute Error (MAE) of 16.6318 but has poor interpretability. The adoption of a 3-Dimension convolution neural network (CNN) with a Graph Convolutional Network (GCN) for extraction of spatiotemporal passenger flow features is developed. The performance of the system is evaluated using root means squared (RMS) error of 9.402%. The prediction of passenger frequency is optimized through the learning network with Adam's long and short-term memory network (Adam-LSTM) in urban rail transit [33]. The system ensured optimal prediction with a performance of 7.14% Mean Relative Error (MRE).

Li et al. developed a forecasting model for multi-source data set of rail transit in an urban city of Beijing using SARIMA and SVM for the training and prediction analysis

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2022 doi:10.20944/preprints202210.0112.v1

model to establish the traffic flow [34]. The model uses an intelligent data acquisition system integrated with sensors networks-based Internet of things (IoT) for the remote data collections on a large scale with precise passenger frequency data. The model is suitable for complex data, nonlinearity in nature, and periodic urban rail transit. The obtained prediction outcome fits well with the measured data and is suitable for short-term passenger frequency forecasting.

3. Materials and Methods

This research focuses on the passenger flow monitoring and prediction analysis for the dynamic time-table scheduling to manage the inter-campuses bus shuttle and the passenger queue system. The model uses a convolutional neural network as a learning stage and a Kalman filter algorithm for the prediction resolutions. This process is achieved through the analysis monitoring of passenger flow during the week (Monday-Friday) and between 6:30 am to 6:30 pm on the days. The information is collected through the mounted CCTV camera installation at the bus terminals, and the historical data clips of the passenger arrival frequency were collected for the extraction, processing, and training using the convolutional neural network (CNN) of big data technology. The CNN algorithm is used for the extraction of passenger headcount, training, and learning the hidden pattern for the simplified input and output parameters relationship in the model. The output (preceding information) of learning or called supervising model is fed directly into the prediction model (Kalman Filter) for better prediction accuracy as illustrated in Fig. 1. This prediction model is controlled by Kalman gain for tunning the parameters and updated immediately after each iteration using covariance matrix (Z) and error estimation. Although, the big data technology approach using CNN is programmed with R language for data extraction, processing, and training to prevent multi-dimensional objectives in this concept. The future variations were studied and simulated in MATLAB using the CNN-Kalman Filter and ARIMA model.

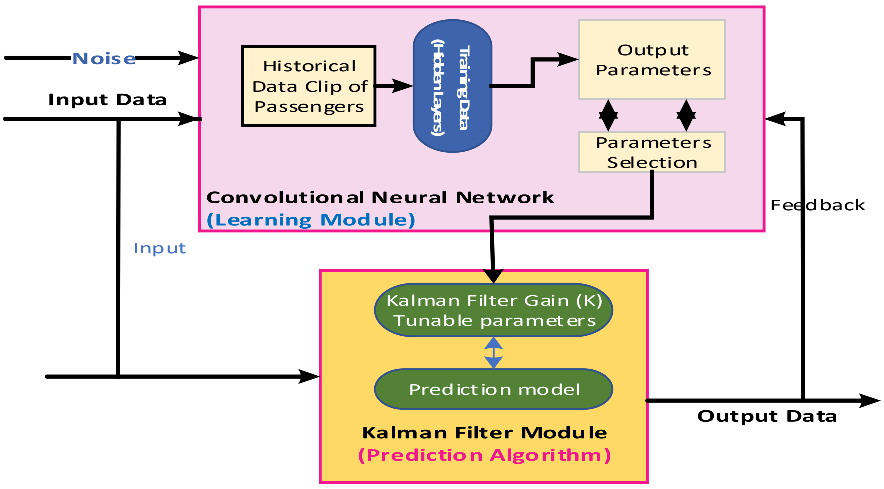


Figure 1. Intelligent hybrid architecture for the passenger frequency prediction.

3.1. Site survey and data collection process

A site survey was carried out to determine the traffic flow at the bus terminals of inter-campuses as shown in Fig. 2a. The transit route of FUTMinna (Bosso–Kpakungu-Gidan Kwano) has been surveyed and digitized to capture potential bus parks, markets, and public facilities within their corridors as illustrated in Fig. 2b. The facilities required at each terminal include an in-vehicle intelligent transportation system, CCTV camera, recording system, a mounted visual display unit powered with solar energy, router, network video recorder (NVR), ITB hard drive, and many others. However, the vehicular speed, density, and passenger flow characteristics were studied for the potential application in our research.

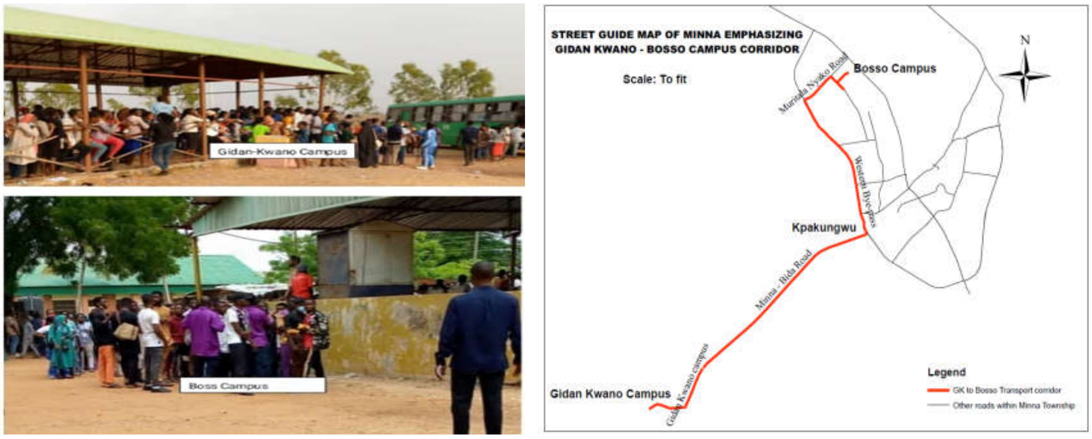


Figure 2. (a) Site survey and passenger frequency of FUTMinna inter-campuses. (b) Geographic map location of the FUTMinna campuses.

3.1.1. Traffic flow based on speed

This is the distance covered at time (t), with vehicle speed on the transit. This speed is difficult to measure and requires selection or sampling of vehicle travel within an area over time (t) for average speed measurement. The average speed of a vehicle's travel per time depends on two approaches such as mean time speed (MTS) and mean space speed (MSS). The average mean time speed (v_t) can be given as in "(1)", where, ψ is the number of vehicles traveling along with the fixed point, and v_j is the speed of vehicles number (when j = 1, 2, 3…nth).

$$v_i = \frac{1}{\psi}\sum_{j=1}^{\psi} v_j \quad (1)$$

The average speed of a vehicle measurement along the roadway over time *t* is compared to the meantime speed, and the average MSS (v_s) can be calculated as in (2), and the ω is the number of vehicles traveling along the route.

$$v_s = \left(\frac{1}{\omega}\sum_{j=1}^{\omega}\frac{1}{v_j}\right)^{-1} \quad (2)$$

3.1.2. Traffic flow-based density

Density *(σ)* is another factor or parameter considered in the prediction and measurement of traffic flow in the transportation system. Density (σ) is described as the population of vehicles per unit length or called spacing (∂) as expressed in "(3)". The spacing or called average density is equal to the inverse proportional of density (σ) within the roadway length (L) at a given period (t_1) as given in "(4)", and the average density time or spacing-time can be evaluated in a particular area or region (∀) as expressed in "(5)".

$$\sigma = \frac{1}{\partial} \quad (3)$$

$$\partial(L, t_1) = \frac{\omega}{L} = \frac{1}{\partial(t_1)} \quad (4)$$

$$\partial(\forall) = \frac{\omega}{L} = \frac{\omega dt}{L dt} \quad (5)$$

3.1.3. Traffic flow-based passenger

This refers to the number of passenger movements across reference points per unit of time and hour. The inverse flow of this passenger is the measure of elapsed time between the guest passing across the reference point in space (*ith*) and the subsequent time *(i + 1)* called headway (ℏ) as expressed in "(6)" and "(7)". The headway (ℏ) is constant when traffic jam tends to infinity. The inverse of the headway (ℏ) is expressed as in "(8)",

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2022 doi:10.20944/preprints202210.0112.v1

and the time-space over the region (β) with total distance (td) can be evaluated with expression in "(9)".

$$F = \sigma\delta \tag{6}$$

$$F = \frac{1}{\hbar} \tag{7}$$

$$F(T, x_1) = \frac{\psi}{T} = \frac{1}{\hbar(x_1)} \tag{8}$$

$$F(\beta) = \frac{\psi}{T} = \frac{\psi dx}{Tdx} = \frac{td}{\beta} \tag{9}$$

3.2. Markovian queue model for passenger flow monitoring

The Markovian queue model of continuous-time series using the Poisson arrival equation is implemented for the queue monitoring and transition of the passenger flow (arrival and departure). The passenger transitions that occur between the arrival rate λ and departure rate μ at the bus terminal of k are duly considered by assuming the average arrival as $(\lambda = \lambda_1, \lambda_2, \lambda_3, \lambda_4, ...\lambda_k)$, and departure $(\mu = \mu_1, \mu_2, \mu_3, \mu_4, ...\mu_k)$, as per unit time to the queue [35-36]. The markovian process of the passenger frequency monitoring is depicted in Fig. 4.

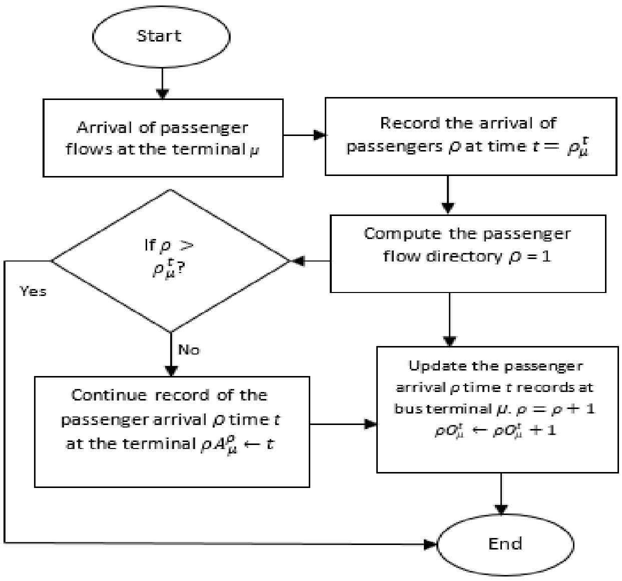


Figure 4. Flowchart for the passenger arrival and queue monitoring.

So, the state equation for the passenger arrival time and departure time can be monitored using probability in state n as in "(10)", and when the condition is tending to infinity can be expressed as in "(11)". The adjustment of the complete state probability of the passenger flow (arrival and departure) is given in "(12)" when $P_n(n \geq 1)$.

$$P_n = \frac{\lambda_{n-1}\lambda_{n-2}\lambda_{n-3}...\lambda_0}{\mu_n\mu_{n-1}\mu_{n-2}...\mu_1} P_0 = P_0 \prod_{i=n}^{n-1} \frac{\lambda_i}{\mu_{i+1}} \tag{10}$$

$$\sum_{n=0}^{\infty} P_n = P_0 + P_0 \sum_{n=0}^{\infty} \prod_{i=n}^{n-1} \frac{\lambda_i}{\mu_{i+1}} = 1 \tag{11}$$

$$P_0 = \frac{1}{1 + \sum_{n=0}^{\infty} \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}}} \tag{12}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2022 doi:10.20944/preprints202210.0112.v1

We assume that, if the arrival means the rate is λ, then $C_\theta^2 = \frac{\ell_\theta^2}{\left(1/\lambda\right)^2}$, and the service rate is μ, when $C_x^2 = \frac{\ell_x^2}{\left(1/\mu\right)^2}$. The waiting time of the passenger in the queue can be perfectly approximate as expressed in "(13)-(15)".

$$\delta_{nq} \approx = \frac{\ell^2 (C_\theta^2 + C_x^2)}{2(1-\ell)} g \tag{13}$$

$$g = \exp\left(\frac{2(1-\ell)(1+\ell^2 C_\theta^2)^2}{3\ell(C_\theta^2 + C_x^2)}\right), when C_\theta^2 \leq 1; \tag{14}$$

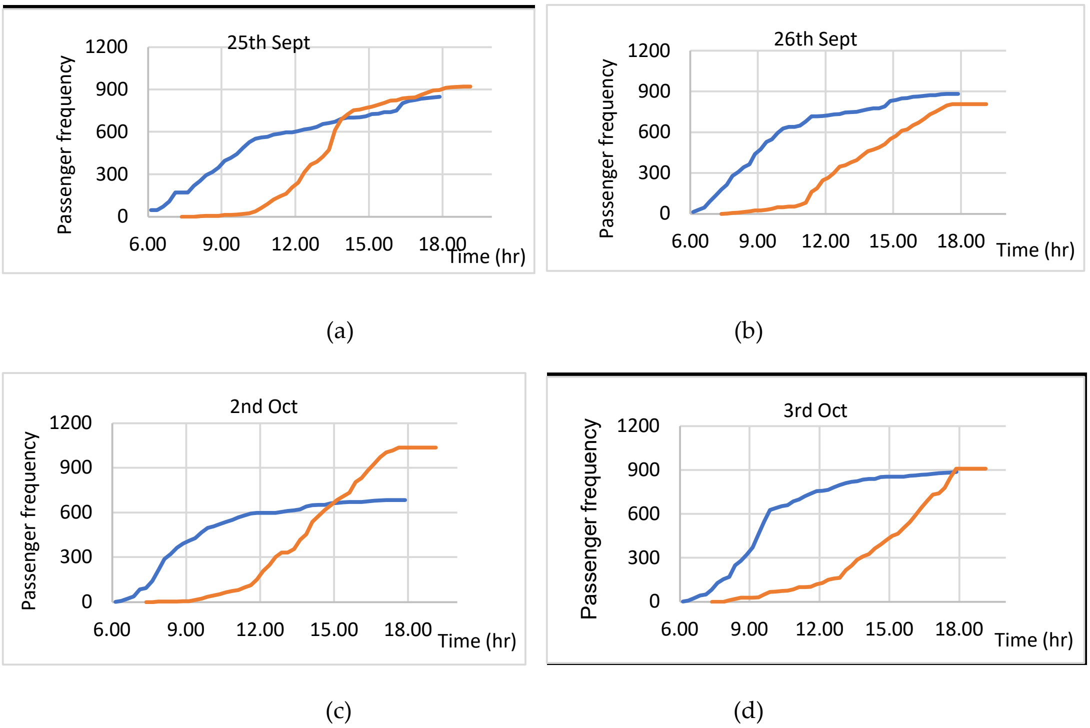
$$g = \exp\left(\frac{(1-\ell)(1-C_\theta^2)^2}{(C_\theta^2 + 4C_x^2)}\right), when C_\theta^2 \geq 1; \tag{15}$$

The passenger waiting time and departure time determinants can be derived as expressed in "(16)-(17)" [37,38].

$$P_0 = \frac{1}{\left[\sum_{M=0}^{c-1} \frac{(c\ell)^M}{M!} + \frac{(c\ell)^c}{c!(1-\ell)}\right]} \tag{16}$$

$$\varpi_q = \frac{\delta_{nq}}{\lambda} \tag{17}$$

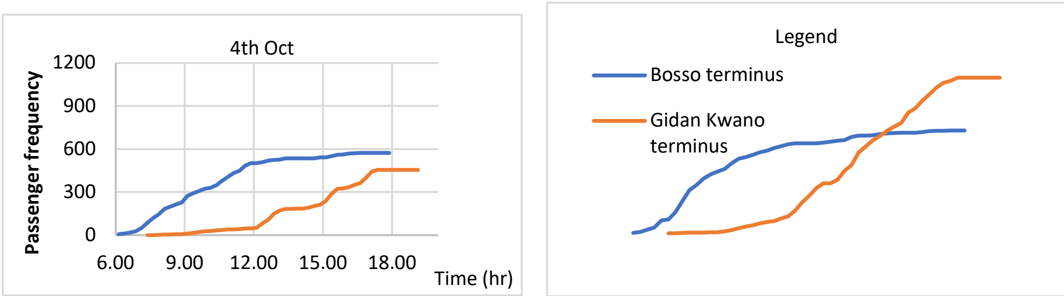
Furthermore, the data recording was preprocessed, trained, and analyzed to estimate the frequency of passengers arriving at the terminals from 6.00 am to 6.00 pm for Bosso station, and 7.00 am to 7.00 pm at Gidan-Kwano station. This recording data of passenger frequency covered two weeks between 25th September and 4th October 2021 were illustrated in Fig. 5.



(a) (b)

(c) (d)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2022 doi:10.20944/preprints202210.0112.v1



(e) (f)

Figure 5. Analysis of the passenger frequency arrival at the bus terminals between 25th September to 4th October 2021.

3.3. ARIMA model for the passenger frequency prediction

A stochastic time series model ARIMA was used in the forecasting of passenger frequency information observation at the bus terminals of the university inter-campuses. This stochastic model varies in time and predicts future values with a confidence interval. In the autoregressive part model p[AR], the degree of differencing order in the model is d[I], and the order of the moving average part of the model is q[MA]. The difference between the part model is D, the autoregressive seasonal arrangement of the model is P, Q is the order of the moving average seasonal part of the model, and *s* is the period of the model. If $y_t = \{X_t\}$ is in series with mean μ, the series of ARIMA model is *(p, d, q) (P, D, Q)*^{s}, and can be expressed as in "(18)-(19)".

$$\begin{cases} Y_t = (1-\beta)^d (1-\beta^s)^D X_t - \mu \\ \phi(\beta)\phi(\beta^s)Y_t = \theta(\beta)\theta(\beta^s)z_t, z_t \sim N(0,\sigma^2) \end{cases} \tag{18}$$

$$\begin{cases} \phi(z) = 1 - \sum_{i=1}^{p} \phi_i z^i, \quad \phi(z) = 1 - \sum_{i=1}^{p} \phi_i z^i \\ \theta(z) = 1 - \sum_{i=1}^{q} \theta_i z^i, \quad \theta(z) = 1 - \sum_{i=1}^{q} \theta_i z^i \end{cases} \tag{19}$$

These ARIMA models of *(p, d, q) (P, D, Q)*^{s}, are autoregressive, differencing, and moving-average respectively, *s* is the number of the observation periods [39]. The β represents backshift notation and can be expressed as $\beta^{\kappa} y_t = y_{t-\kappa}$, the estimation of ARIMA coefficient are $\phi(\beta) = (1-\phi_1\beta^s,\ldots,\phi_p\beta^{s*v})$ and z_t is a noise process error assumed, and σ^2 is a covariance of the forecasting model $y_t = \beta r_t + \eta_t$. The βr_t represent regression coefficient, η_t represent the regression error during the ARIMA method in forecasting. The passenger frequency prediction flowchart using the ARIMA model is illustrated in Fig. 6.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2022 doi:10.20944/preprints202210.0112.v1

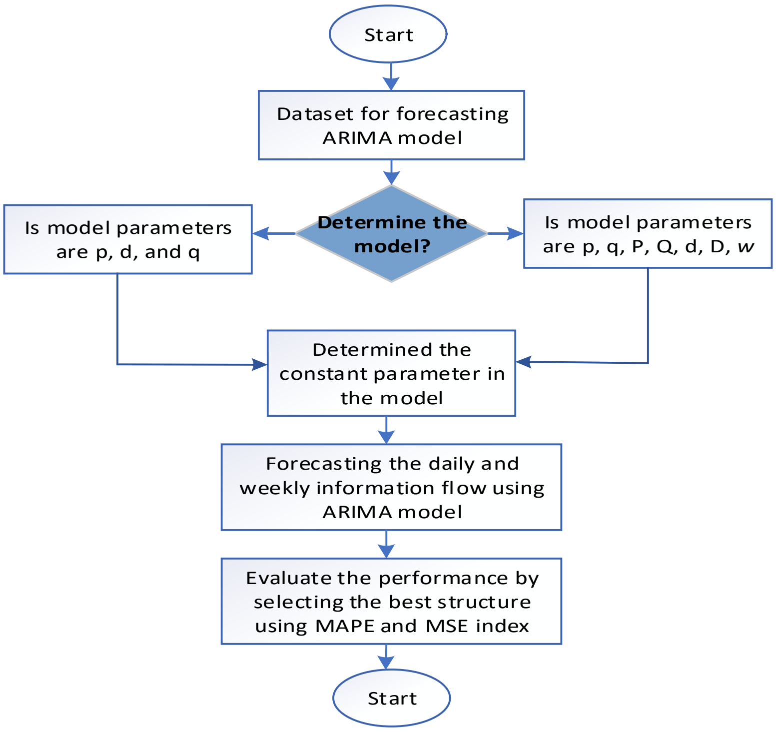


Figure 6. Passenger frequency prediction flowchart using ARIMA model technique.

3.4. LEARNING AND PREDICTION MODEL USING CNN-KFA

The hybrid model which involved a convolution neural network and Kalman algorithm was used as an intelligent learning module for the passenger features extraction, analysis, and prediction. The extraction of the passenger's video clip is collected from the storage device which feeds into the R programming environment for the data cleaning (removing any outlier and filling the missing parameters) using CNN. After the preprocessing (data cleaning), the dataset is subjected to training and validation which was evaluated using MSE and MAPE. The modeling of the output dataset is fed into the KF algorithm for the prediction resolutions as illustrated in Fig. 7.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2022 doi:10.20944/preprints202210.0112.v1

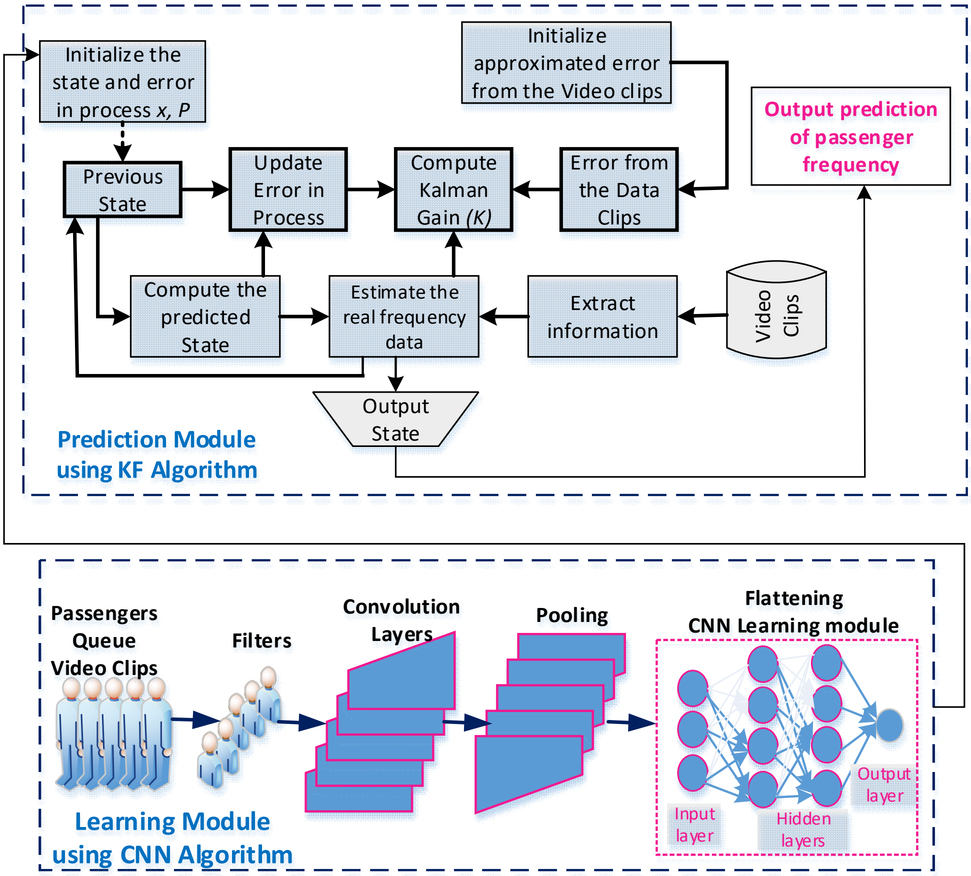


Figure 7. Hybrid model architecture for the passenger frequency prediction using CNN-KFA.

The KF vector and its associated covariance are predicted, while the state vector and its covariance are corrected using the measured data. The Kalman observation model can be expressed as in "(18)-(21), where, (φ_1, φ_1) are the state vector of AR(2) parameters.

$$y_t = \varphi_1, y_{t-1} + \varphi_2, y_{t-2} + v) \quad (20)$$

$$H = [y_{t-1}, y_{t-2}, \ldots y_{t-p}]^T \quad (21)$$

$$H = [y_{t-1}, y_{t-2}]^T \quad (22)$$

$$x_t = [\varphi_1, \varphi_1] \quad (23)$$

Since we are modeling time-series data, the observation model was formulated as an autoregressive model of order 2 AR(2) and embedded in the KF to resolve the non-stationary time series data as discussed in Table 1.

Table 1. Kalman filter algorithm.

Stages	Methods
1.	The Kalman system model is formulated with the dynamic and observation model thus: $x_t = F_x * x_{t-1} + F_n * n$ (Dynamic model) $y_t = H * x_t + v$ (Observation model) where, x_t is the state vector of the current time step x_{t-1} is the state vector of the previous time step n is the perturbation vector (noise for the dynamic model with zero) y_t is the measurement vector v is the measurement noise with zero mean F_t is the transition matrix F_n is the perturbation matrix H is the measurement matrix H_t is the transformation matrix into the measurement domain for the state vector parameters mapping n and v are white noise that independent of each other, and they follow normal probability distributions of $p(n) \sim N(0,Q)$, and $p(v) \sim N(0,R)$ where Q and R, are covariance matrix for processing noise
2.	Initialize F_x, F_n, H, and x_t, then initialize covariance matrices P, Q, and R.
3.	Begin the time loop by a predict the state vector x_t with its covariance P, and correct the x_t with its covariance P by computing the following expressions; Expectation, $e = H*x_t$ Covariance of expectation, $E = H*P*H'$ Innovation, $z = y - e$ Covariance of innovation, $Z = R + E$ Kalman gain, $K= P + H'' * Z^{-1}$ then, calculate the new x_t and P as expressed in "(24)-(25)" $x_t = x_t + K * P$ (24) $P = P - K * H * P$ (25) The standard Kalman filter algorithm model with the transformation matrix H_t that allow the prediction and measurement are summarized from Kalman gain in "(26)-(27)". $K = \frac{H\sigma_1^2}{H^2\sigma_1^2 + \sigma_2^2} \to K_t$ (26) $K = P_{t\|t-1} H_1^2 (H_t P_{t\|t-1} H_t^T + R_t)^{-1}$ (27) The conversion of the transformation matrix into the domain for the prediction and measurement is expressed as in "(28)-(29)". $\hat{x}_{t\|t} = \hat{x}_{t\|t-1} K_t (z_t = H_t \hat{x}_{t\|t-1})$ (28) $P_{t\|t} = P_{t\|t-1} R_t H_t P_{t\|t-1}$ (29)
4.	End the time loop
5.	Plot results.

 doi:10.20944/preprints202210.0112.v1

4. Results and Discussion

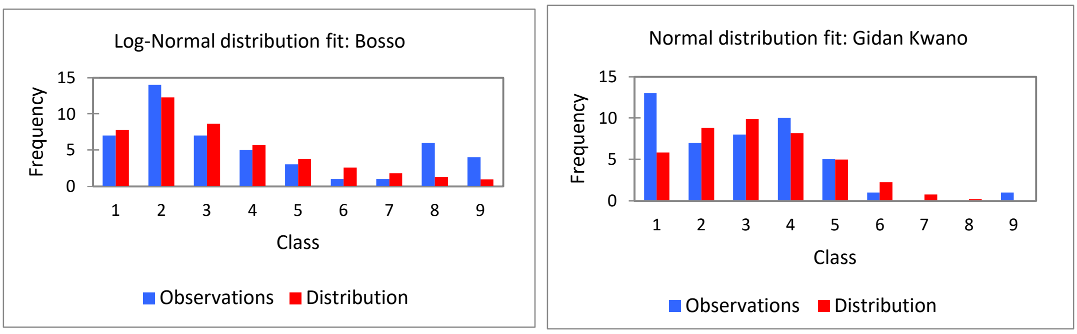
4.1. Frequency distribution fit of passenger flow prediction model

The frequency of the passengers waiting time at the terminal is a significant problem in the public transport bus park. The average passenger flow waiting times at the bus terminal is the half headway service to the randomness of passenger arrivals. When considering the randomness of the variables $\xi_1, \xi_2, \xi_3, \ldots, \xi_n$ as the frequency of passengers arriving at the terminals, the mean distribution process is observed μ_ξ. The performance of the distributions was assessed using a statistics p-value for the hypothesis testing. The *p*-value conveys information about the weight of evidence against the null hypothesis (H_0). The null hypothesis states that the frequency of passengers arriving at a bus terminal follows a normal distribution will reject the hypothesis if the weight of evidence (p-value) against H_0 is not significant. That is, if smaller than the significant level of 0.05 (α = 5%), and accept if the p-value is greater than the significant level. The smaller the P-value, the greater the evidence against H_0. The frequency of passenger flow at both terminals is recorded, which varies with the days and time. The passenger arrival frequency framework was evaluated under the conditions of characteristics distribution of Normal, Log-normal, Logistic, and Gamma functions respectively as presented in Table 2. However, the graphical result analysis that shows that Log-Normal distribution fits Bosso passenger flow, and Normal distribution fits GK passenger flow is presented in Fig. 8a and 8b. So, more commuters are arriving at Bosso terminal than at GK terminal between the period 5:00 GMT and 16:00 GMT. However, there are more commuters at the GK terminal after 16 GMT hoping to return to their destinations. The passenger frequency at Bosso terminal is high as 250 by 6:00 GMT and rises to 650 by 13.00 GMT of the day. The situation at GK terminal differs from that of Bosso terminal where passenger flow of about 110 at 12.30 GMT with a peak of 900 at 17:30 GMT.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2022 doi:10.20944/preprints202210.0112.v1

Table 2. Passenger flow characteristics distribution.

Distribution	P-value	
	Frequency for Bosso terminal	Frequency for GK terminal
Normal	0.080	0.659
Log-normal	0.539	0.036
Logistic	0.221	0.588
Gamma	0.274	0.192
Best fit	Log-normal values μ = 2.512 (error = 0.108) $\sigma = 0.889$ (error=1.795)	Normal distribution values μ = 18.956 (error=0.116) $\sigma = 14.351$ (error=1.977)
	Kolmogorov-Smirnov test	
	D = 0.113 p-value = 0.539 Alpha = 0.05 H_0: Lognormal distribution rejects the null hypothesis H_0 while it is true is 53.93%	D = 0.106 p-value=0.659 Alpha = 0.05 H_0: Normal distribution reject the null hypothesis H_0 while it is true is 65.87%



(a) (b)

Figure 8. Distribution fit of the passenger frequency at the terminals.

4.2. ARIMA and CNN-KALMAN simulation model

The analysis of the time-domain for passenger frequency prediction (short-time) was carried out and evaluated using ARIMA and CNN-KFA. A stationary time series model with autocorrelation is adopted in the data analysis of short-term passenger frequency. This includes autoregressive process with order p AR(p), moving average process with order q MA(q), and autoregressive moving average process with orders p and q (ARIMA (p, q). The parameter of the selected model is observed using the differential operation method to transform the non-stationary series to a stationary one, and measured with the ACF and PACF with different features. The prediction analysis of the Bosso terminal shows the ARIMA models that performed best are ARIMA (2,1,1), ARIMA (3,1,1), ARIMA (3,1,3), ARIMA (1,1,4), and ARIMA (1,1,1) model for the respective days between 25th Sept., to 4th Oct as presented in Table 3. Similar analysis was carried out for Gidan-Kwano terminal and the best model is ARIMA (1,1,1), ARIMA (1,1,1), ARIMA (1,1,1), ARIMA(1,1,1), ARIMA (1,1,1), and ARIMA (1,2,1) for the respective days as shown in Table 4. The graphical illustrations of the prediction models at both terminals fo series I and II are in Fig 9-12.

Table 3. ARIMA parameters' statistics for bosso and GK series I.

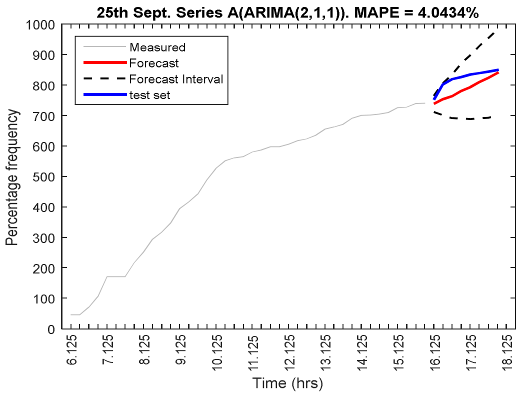
Model	Conditional Probability Distribution: Gaussian			
ARIMA(2,1,1)	Parameter	Value	Standard Error	t- Statistic
	Constant	14.5826	3.21618	4.53415
	AR{1}	-0.277553	0.0505743	5.48803
	AR{2}	0.323114	0.0863819	3.74053
	MA{1}	1	0.0788999	12.6743
	Variance	182.449	37.0802	4.9204
ARIMA (3,1,1)	Constant	1.1153	3.7413	0.298104
	AR{1}	0.87982	0.6438	1.36661
	AR{2}	0.114374	0.326041	0.350797
	AR{3}	-0.0709957	0.321207	-0.221028
	MA{1}	-0.549939	0.569557	-0.965556
ARIMA(3,1,3)	Constant	18.3373	3.54223	5.17677
	AR{1}	0.0400275	0.135642	0.295097
	AR{2}	-0.548672	0.0590592	-9.2902
	AR{3}	0.559627	0.114938	4.86895
	MA{1}	0.549308	0.283821	1.9354
	MA{2}	1.00185	0.193999	5.16421
	MA{3}	0.00339307	0.252061	0.0134613
	Variance	181.594	53.8445	3.37256
ARIMA(1,1,4)	Constant	3.19206	5.08127	0.628201
	AR{1}	0.764812	0.206109	3.71072
	MA{1}	-0.07163	0.191964	-0.373143
	MA{2}	-0.0822544	0.144216	-0.570356
	MA{3}	-0.0382082	0.167825	-0.227667
	MA{4}	0.307267	0.184992	1.66098
	Variance	133.893	24.808	5.39718
ARIMA(1,4,1)	Constant	4.04007	6.63474	0.608927
	AR{1}	0.789722	0.246886	3.19873
	MA{1}	-0.107717	0.289063	-0.372642
	MA{2}	-0.142334	0.222155	-0.640696
	MA{3}	-0.108311	0.188883	-0.573429
	MA{4}	0.228137	0.131839	1.73043
	Variance	250.249	41.7286	5.99707
ARIMA(1,1,1)	AR{1}	0.830206	0.150675	5.5099
	MA{1}	-0.31076	0.178112	-1.74474
	Variance	80.616	15.2273	5.29418

 doi:10.20944/preprints202210.0112.v1

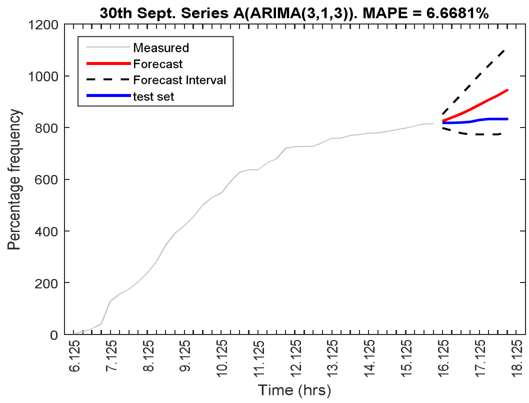
Table 4. ARIMA parameters' statistics for bosso and GK series II.

Model	Conditional Probability Distribution: Gaussian			
ARIMA(1,1,1)	Parameter	Value	Standard Error	t- Statistic
	Constant	6.00631	8.38266	0.716515
	AR{1}	0.696391	0.193262	3.60336
	MA{1}	-0.055414	0.232832	-0.238
	Variance	351.409	43.4869	8.08079
ARIMA (1,1,1)	Constant	2.68909	5.76032	0.46683
	AR{1}	0.849509	0.259418	3.27467
	MA{1}	-0.540168	0.350765	-1.53997
	Variance	201.592	49.0742	4.10791
ARIMA(1,1,1)	Constant	3.28838	5.60331	0.586864
	AR{1}	0.856178	0.170058	5.03462
	MA{1}	-0.380161	0.229661	-1.65531
	Variance	233.506	42.2953	5.52086
ARIMA(1,1,1)	Constant	7.48495	7.72266	0.96922
	AR{1}	0.611154	0.255924	2.38802
	MA{1}	0.0144436	0.254074	0.056848
	Variance	224.163	64.4252	3.47943
ARIMA(1,1,1)	Constant	1.77054	6.73689	0.262813
	AR{1}	0.901718	0.284548	3.16895
	MA{1}	-0.786899	0.366814	-2.14523
	Variance	666.683	239.948	2.77844
ARIMA(1,2,1)	Constant	4.04687	3.98424	1.01572
	AR{1}	0.820808	0.642674	1.27718
	AR{2}	-0.246971	0.497713	-0.496212
	MA{1}	-0.00623382	0.671024	-0.00929002
	Variance	94.5432	23.0806	4.09622

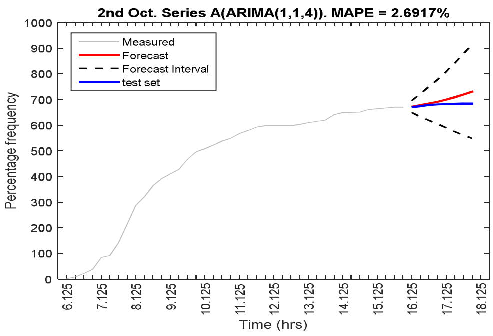
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2022 doi:10.20944/preprints202210.0112.v1



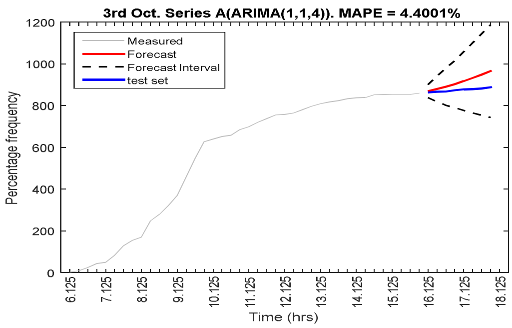
(a)



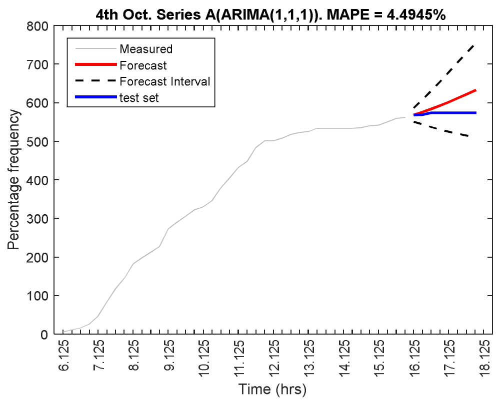
(b)



(c)



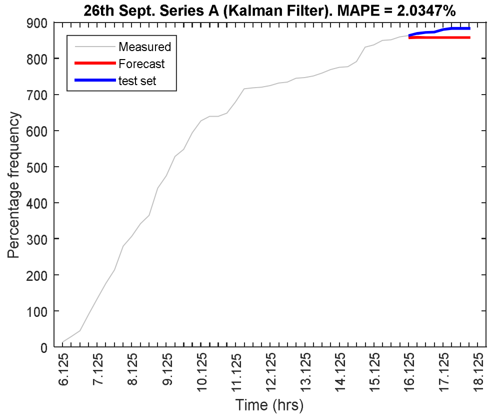
(d)



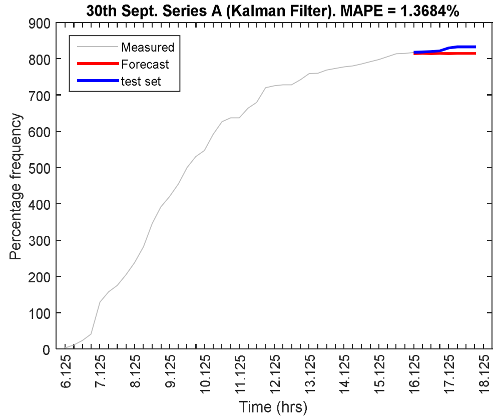
(e)

Figure 9. MAPE estimate for ARIMA model in bosso series I between 26th and 4th October 2021.

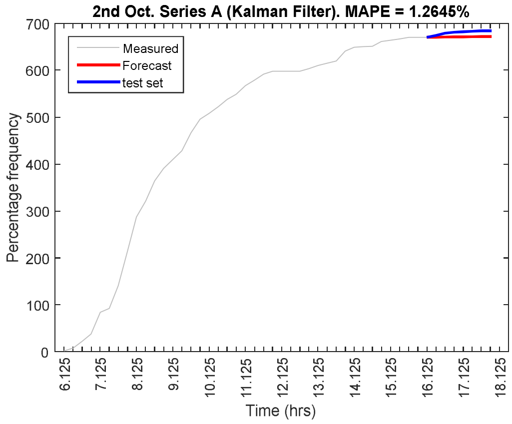
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2022 doi:10.20944/preprints202210.0112.v1



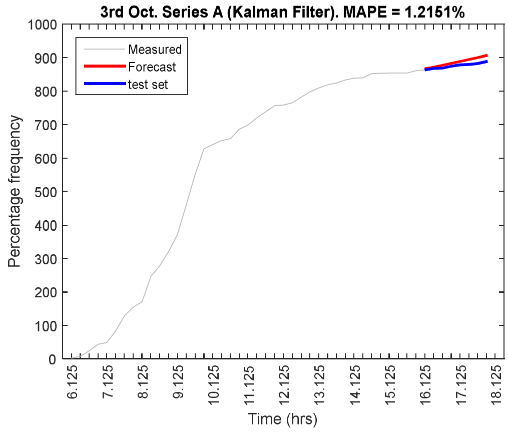
(a)



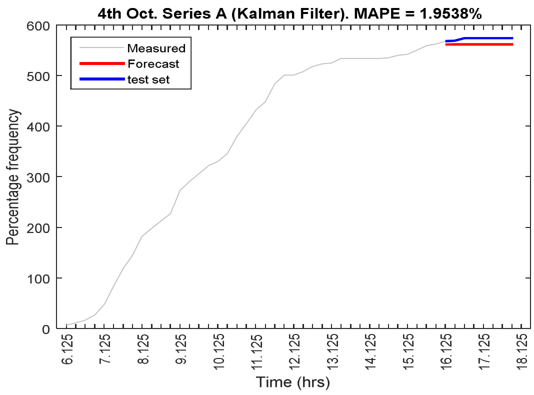
(b)



(c)



(d)



(f)

Figure 10. MAPE estimate for kalman model in bosso series I between 26th and 4th October 2021.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2022 doi:10.20944/preprints202210.0112.v1

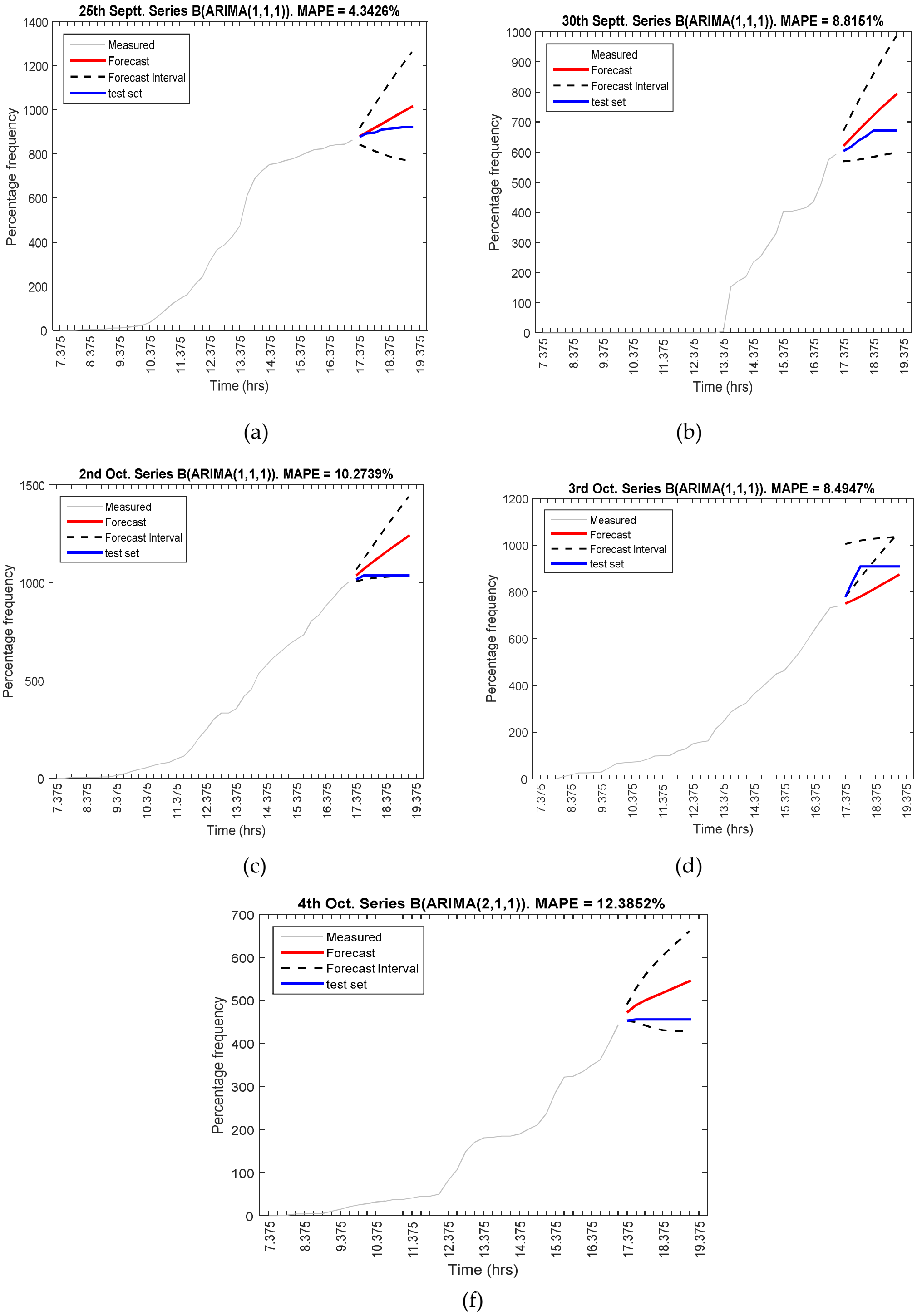


Figure 11. MAPE estimate for ARIMA model in bosso series II between 26th and 4th October 2021.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2022 doi:10.20944/preprints202210.0112.v1

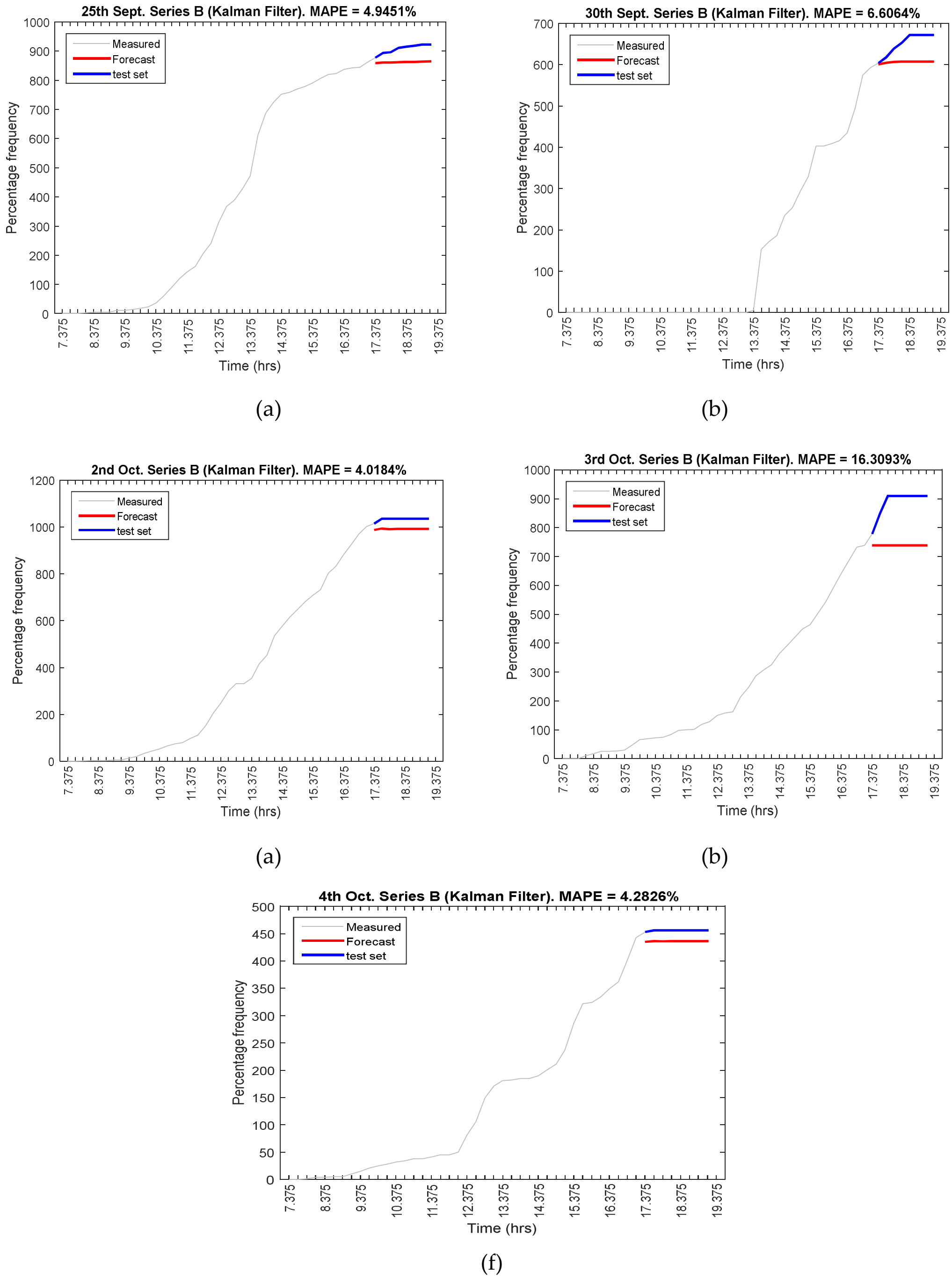


Figure 12. MAPE estimate for kalman model in GK series II between 26th and 4th October 2021.

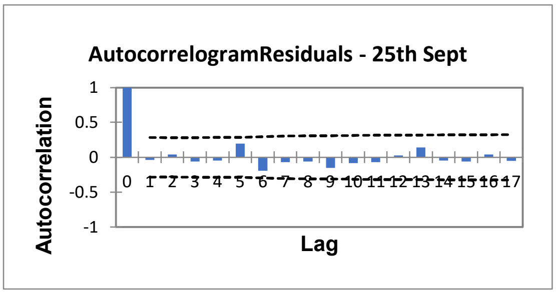
However, the autocorrelation functions (ACF) and partial autocorrelation functions (PACF) were used to test the ordered of the model structure, and to select the best fit in the model. This process is achieved by capturing the data structure in the model, when ACF(Lag K=0, or 1). We compute the mean values of the original time series data *(x)*, the difference between real time series data and observed mean values x' are calculated. The squared of the output result is taken and calculate the sum of the squared differences for mean observations *(n-k)*. The product of the difference between the original data and mean observation is computed and find the output summation to determine the ACF lag 1 or 0. The AR1 model included in the residual will help to reduces error in determining value for the autocorrelation's coefficient. The autocorrelation coefficient and partial autocorrelation coefficient is expressed as given in "(30)-(31)". The result of autocorrelations

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 October 2022 doi:10.20944/preprints202210.0112.v1

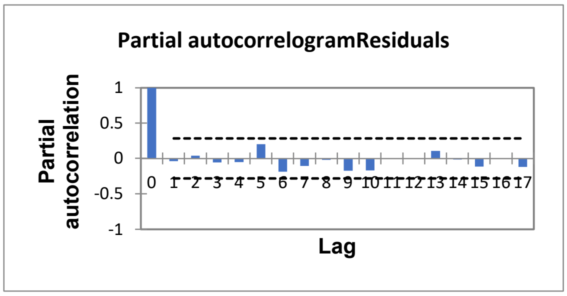
residual and partial autocorrelation residuals of the series in Bosso and GK are graphical illustrated in Fig. 13.

$$ACF(x,y)=\frac{\frac{1}{n-1}\sum_{i=1}^{n}((x_i-\bar{x})(y_i-\bar{y}))}{\sigma x\,\sigma y} \quad (30)$$

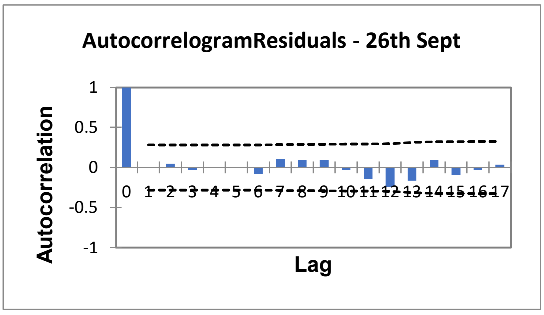
$$PACF(y_i,y_{i-k})=\frac{\sigma(y_i,y_{i-k}\,|\,y_{i-1},y_{i-2},\ldots y_{t-k+1})}{\sigma y_i\,|\,y_{i-1},y_{i-2},\ldots y_{i+k}\,\sigma y_{i-k}\,|\,y_{i-1},\ldots} \quad (31)$$



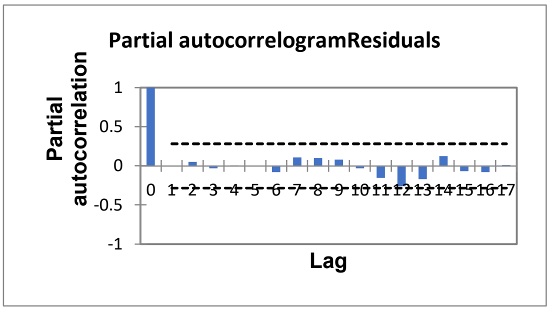
(a)



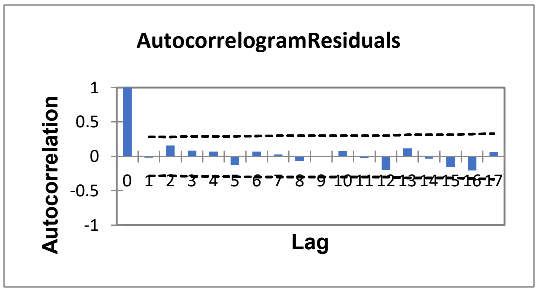
(b)



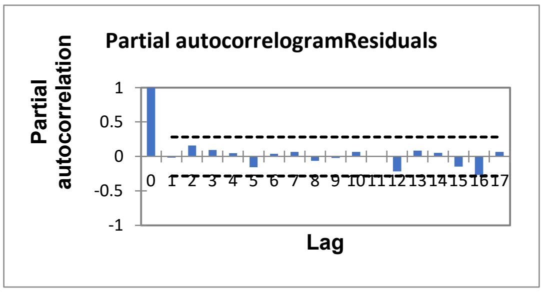
(c)



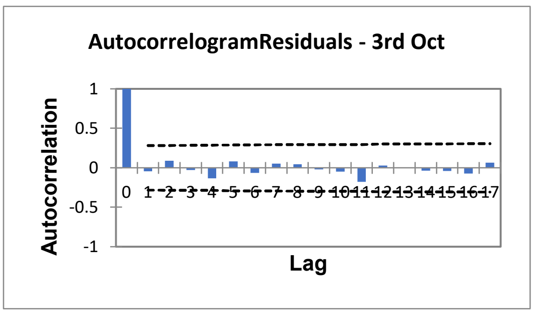
(d)



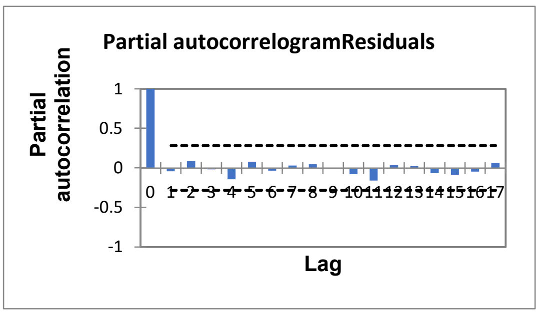
(e)



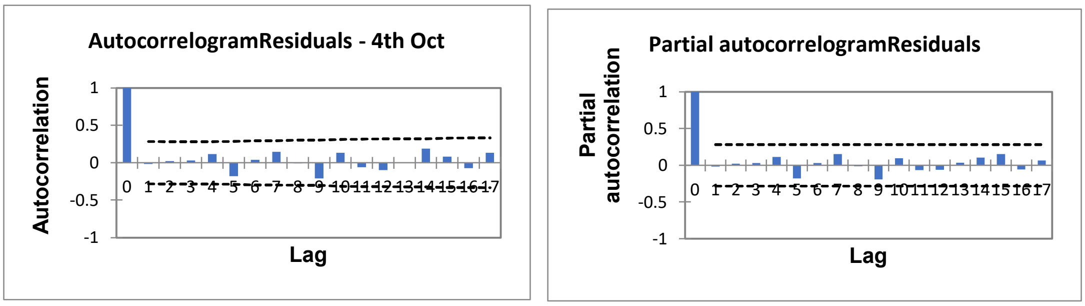
(f)



(g)



(h)



(i) (j)

Figure 13. The ACF and PACF analysis of passenger frequency prediction.

This observed time-series data were used to estimate the parameters and for predictions beyond the instance of the observed time-series information. The prediction values of both ARIMA and CNN-KFA were compared using Mean Absolute Percent Error (MAPE) and the Mean Squared Error (MSE) as expressed in "(32)-(33)", and the graphical results are presented in Fig. 14.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i^{observed} - y_i^{forecast}}{y_i^{forecast}}\right| * 100 \quad (32)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i^{observed} - y_i^{forecast}\right|^2 \quad (33)$$

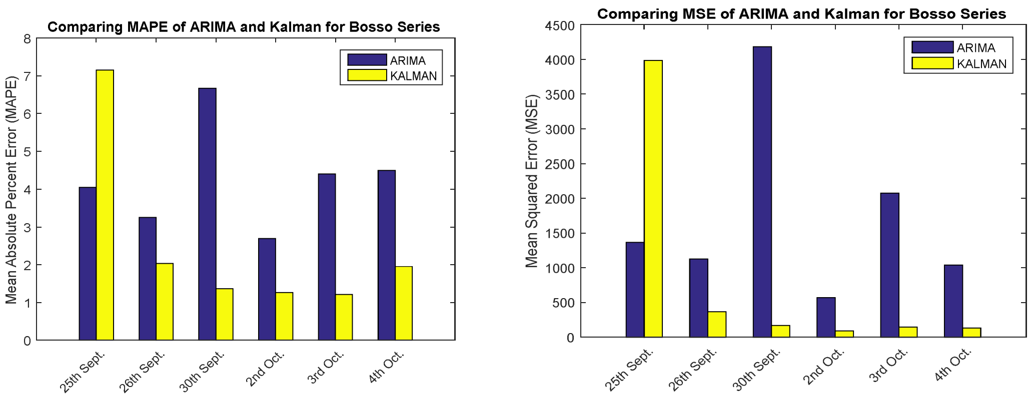


Figure 14a.Comparison analysis between the models using MAPE and MSE for bosso series I.

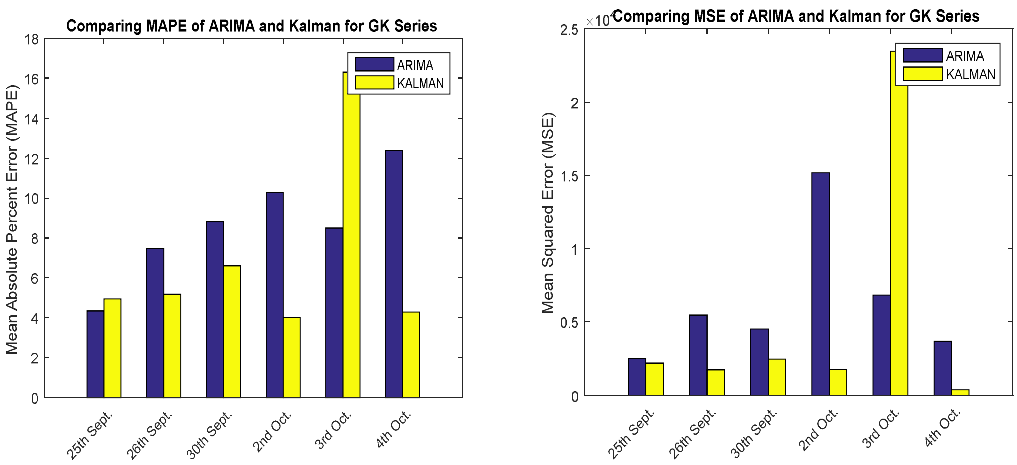


Figure 14b.Comparison analysis between the models using MAPE and MSE for GK series II.

5. Conclusion

The passenger flow prediction for the transportation sustainability and planning of inter-campuses shuttle on the city arterial route was investigated and analyzed for the future preparation of a timetable schedule. A markovian model was used in analyzing the arrival of the passenger flow frequency and queue system through the stationary CCTV position at the bus stations. This helps to manage the travel pattern and allows recognition of the population group, and patronage periods. The Auto-Regressive Integrated Moving Average (ARIMA) and CNN-KFM were used for learning and prediction purposes. The performances are evaluated using MAPE and MSE statistical analysis. The MAPE results for the Bosso terminal using ARIMA and CNN-KFM model shows that CNN-KFM has better performance with 83.33% times than the ARIMA model. Similar results were obtained for the GK terminal which shows that CNN-KFM has the overall best performance in terms of the Mean Absolute Percent Error (MAPE) with 85.5%:66.66% which shows high accuracy. This investigation will help to control environmental pollution and other related factors. The future works will focus on the development of time-table scheduling for public transportation sustainability and, transport fare collection systems using a genetic meta-heuristic programming concept.

Acknowledgments: This research was funded by Tertiary Education Trust Fund (TETFund), Nigeria under the National Research Fund grant Tetfund/DR&D/CE/NRF/STI/34/VOL1.

References

1. A. H. Alkhereibi, S. Tahmasseby, S. Mohammed, D. Muley, "Blue-collar laborers' travel pattern recognition: Machine learning classifier approach", Transportation research interdisciplinary perspectives, vol. 12, pp. 100506, 2021.
2. L. A. Ajao, J. G. Kolo, B. O. Abisoye, I. Z. Jibil, U. M. Jonah, "In-vehicle traffic accident and roadsides barrier detection and alerting system using distance-time based parameters and radar range algorithm", IEEE PES/IAS Power Africa Conference, Nairobi, Kenya, pp. 1-5, 2020.
3. O. D. Jimoh, L.A. Ajao, O.O. Adeleke, , S. S. Kolo "A vehicle tracking system using greedy forwarding algorithms for public transportation in urban arterial", IEEE Access, vol. 8, pp. 191706-191725, 2020.
4. H. K. Suman, N.B. Bolia, G. Tiwari, "Comparing public bus transport service attributes in Delhi and mumbai: policy implications for improving bus services in Delhi", Transportation Policy, vol. 56, pp. 63–74, 2017.
5. O. C. Inalegwu, J. Agajo, L. A. Ajao, M. O. Abdulsalam, O. D. Sadiq, S.C. Peter, "design of an unmanned aerial vehicle for campus area surveillance", Journal of Engineering Science and Applications (JESA), vol. 11, no. 1, pp. 78-86, 2019.
6. Alexandrakis, "Cycling towards sustainability: The transformative potential of urban design thinking in a sustainable living lab", Transportation research interdisciplinary perspectives, vol. 9, pp. 100269, 2021.
7. N. A. El-Taweel, H. Khani, H. E. Farag, "Optimal sizing and scheduling of LOHC-based generation and storage plants for concurrent services to the transportation sector and ancillary services market", IEEE Transactions on Sustainable Energy, vol. 11, no. 3, pp. 1381-1393, 2019.
8. K. Gkiotsalitis, O. Cats, "Reliable frequency determination: incorporating information on service uncertainty when setting dispatching headways", Transportation Research Part C, vol. 88, pp. 187–207, 2018.
9. K. Hamad, P. T. T. Htun, L. Obaid, "Characterization of travel behavior at a university campus: a case study of Sharjah University City, UAE", Transportation research interdisciplinary perspectives, vol. 12, pp. 10048, 2021.
10. S. Z. Zhao, T. H. Ni, Y. Wang, X. T. Gao, "A new approach to the prediction of passenger flow in a transit system", Computers and Mathematics with Applications, vol. 61, no. 8, pp. 1968-1974, 2011.
11. X. Kong, F. Xia, Z. Fu, X. Yan, A. Tolba, Z. Almakhadmeh, "TBI2Flow: travel behavioral inertia based long-term taxi passenger flow prediction", World Wide Web, vol. 23, no. 2, pp. 1381-1405, 2020.
12. S. Liang, M. Ma, S. He, H. Zhang, "Short-term passenger flow prediction in urban public transport: kalman filtering combined k-nearest neighbor approach", IEEE Access, vol. 7, no. 120937-120949, 2019.
13. F. Su, H. Dong, L. Jia, Y. Qin, Z. Tian, "Long-term forecasting oriented to the urban expressway traffic situation", Advances in mechanical engineering, vol. 8, no. 1, pp. 1687814016628397, 2016.
14. H. Li, Y. Wang, X. Xu, L. Qin, H. Zhang, "Short-term passenger flow prediction under passenger flow control using a dynamic radial basis function network", Applied Soft Computing, vol. 83, pp. 105620, 2019.
15. J. B. Ingvardson, O. A. Nielsen, S. Raveau, B. F. Nielsen, "Passenger arrival and waiting for time distributions dependent on train service frequency and station characteristics: A smart card data analysis", Transportation Research Part C: Emerging Technologies, vol. 90, pp. 292-306, 2018.
16. Y. Sun, B. Leng, W. Guan, "A novel wavelet-SVM short-time passenger flow prediction in the Beijing subway system", Neurocomputing, vol. 166, pp. 109-121, 2015.

17. J. Ge, H. Pei, D. Yao, Y. Zhang, "A robust path tracking algorithm for connected and automated vehicles under i-VICS", Transportation Research Interdisciplinary Perspectives, vol. 9, pp. 100314, 2021.
18. C. Nie, H. Wei, J. Shi, M. Zhang, "Optimizing actuated traffic signal control using license plate recognition data: methods for modeling and algorithm development", Transportation Research Interdisciplinary Perspectives, vol. 9, pp. 100319, 2021.
19. F. Hoeft, "Internal combustion engine to electric vehicle retrofitting: Potential customer's needs, public perception, and business model implications", Transportation Research Interdisciplinary Perspectives, no. 9, pp. 100330, 2021.
20. S. Wang, E. Yasmeen, "Forecasting functions for time series and linear models", 2020. URL:http://pkg.robjhyndman.com/forecast. r package version 8.12.
21. K. Gkiotsalitis, O. Cats, "Reliable frequency determination: incorporating information on service uncertainty when setting dispatching headways", Transportation Research Part C, vol. 88, pp. 187–207, 2018.
22. Z. Chunhui, S. SUN-Yang, "Kalman filter-based short-term passenger flow forecasting on bus stop", Journal of Transportation Systems Engineering and Information Technology, vol. 11, no. 4, pp. 154-159, 2011.
23. J. Liu, Y. Heng, H. Zhao, X. Gao, P. A. Wang, "A Prediction model of short-term passenger flow for urban transit hubs", Journal of Transportation Information Safety, vol. 32, pp. 41-44, 2014.
24. J. Zhao, M. Ye, Z. Yang, Z. Xing, Z. Zhang, "Operation optimizing for minimizing passenger travel time cost and operating cost with time-dependent demand and skip-stop patterns: nonlinear integer programming model with linear constraints", Transportation Research Interdisciplinary Perspectives, vol. 9, pp. 100309, 2021.
25. M. S. Grewal, "Kalman filtering", International Encyclopedia of Statistical Science; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany; pp. 705–708, 2011.
26. Y. Gu, Y. Han, X. Fang, "Method of hub station passenger flow forecasting based on ARMA model", J. Transportation Information Safety, vol. 29, pp. 5-9, 2011.
27. K. Zhu, P. Xun, W. Li, Z. Li, R. Zhou, "Prediction of passenger flow in urban rail transit based on big data analysis and deep learning", *IEEE Access*, vol. 7, pp. 142272-142279, 2019.
28. J. Guo, Z. Xie, Y. Qin, L. Jia, Y. Wang, "Short-term abnormal passenger flow prediction based on the fusion of SVR and LSTM", IEEE Access, vol. 7, pp. 42946-42955, 2019.
29. R. Liu, Y. Wang, H. Zhou, Z. Qian, "Short-term passenger flow prediction based on wavelet transform and kernel extreme learning machine", IEEE Access, vol. 7, pp. 158025-158034, 2019.
30. J. Zhang, F. Chen, Q. Shen, "Cluster-based LSTM network for short-term passenger flow forecasting in urban rail transit", IEEE Access, vol. 7, pp. 147653-147671, 2019.
31. O. Cats, S. Glück, "Frequency and Vehicle Capacity Determination using a Dynamic Transit Assignment Model", Transportation Research Record, vol. 2673, no. 3, pp. 574 -585, 2019.
32. J. Zhang, F. Chen, Y. Guo, X. Li, "Multi-graph convolutional network for short-term passenger flow forecasting in urban rail transit", IET Intelligent Transport Systems, vol. 14, no. 10, pp. 1210-1217, 2020.
33. B. Wang, M. Ye, Z. Zhu, Y. Li, Q. Liang, J. Zhang, "Short-term passenger flow prediction for urban rail stations using learning network based on optimal passenger flow information input algorithm", IEEE Access, vol. 8, pp. 170742–170753, 2020.
34. W. Li, L. Sui, M.Zhou, H. Dong, "Short-term passenger flow forecast for urban rail transit based on multi-source data", EURASIP Journal on Wireless Communications and Networking, vol. 1, pp. 1-13, 2021.
35. F. R. Cruz, M. A. Almeida, M. F. D'Angelo, T. van Woensel, "Traffic intensity estimation in finite markovian queueing systems", Mathematical Problems in Engineering, 2018.
36. M. Anokye, A. R. Abdul-Aziz, K. Annin, F. T. Oduro, "Application of queuing theory to vehicular traffic at a signalized intersection in Kumasi-Ashanti region, Ghana", American International Journal of Contemporary Research, vol. 3, no. 7, pp.23-29, 2013.
37. D. Fiems, H. Bruneel, "Discrete-time queueing systems with Markovian preemptive vacations", Mathematical and Computer Modelling, vol. 57, no.3, pp. 782-792, 2013.
38. Y-T., Zhu, B-H., Mao, L. Liu, M-G. Li, "Timetable design for urban rail line with capacity constraints", Discrete Dynamics in Nature and Society, vol. 1, pp. 1-11, 2015.
39. Z. Chenga, M. Trepanier, L. Sun, "Incorporating travel behavior regularity into passenger flow forecasting", Transportation Research Part C: Emerging Technologies, vol. 128, pp. 103200, 2021.