IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 20, 2019, accepted May 7, 2019, date of publication May 22, 2019, date of current version July 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918217

Multi-Objective Transport System Based on Regression Analysis and Genetic Algorithm Using Transport Data

MOUBEEN FAROOQ KHAN, SOHAIL ASGHAR, MANZOOR ILLAHI TAMIMI, AND MUHAMMAD ASIM NOOR
Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan

Corresponding author: Sohail Asghar (sohail.asg@gmail.com)

ABSTRACT Intelligent transportation system (ITS) is a cutting-edge traffic solution employing state-of-the-art information and communication technologies. Optimized bus-scheduling, being an integral part of ITS, ensures safety, efficiency, traffic congestion-reduction, passengers' forecast, resource allocation, and drivers' experience enhancement. Nevertheless, of its significance, recent years have witnessed limited research carried out in this context. In this paper, we apply a uni-variate multi-linear regression over the past three years of data from a renowned bus company and forecasted potential passengers for different days in a week. Moreover, a minimum number of different types of buses have been calculated, and bus optimization has been performed in a genetic algorithm. The results of accurateness has been validated by using mean absolute deviation (MAD) and mean absolute percentage error (MAPE). The values of MAD (99.14) and MAPE (8.7%) advocate that the results are quite rational.

INDEX TERMS Intelligent transportation systems, mean absolute deviation (MAD), mean absolute percentage error (MAPE), genetic algorithm, regression analysis.

I. INTRODUCTION

With the drastic increase in the world's population, along with all other resources, the need for efficient and economical transportation has become a major concern of the modern age. Bus-system; being an integral part of a transportation system, provides greater capacity at the economy of scales. An efficiently managed bus-system can be very profitable in developing countries. This management includes various factors; allowing a specific number of buses for a time-slot and route, when to increase the number of buses and how to schedule bus-departures. Other indispensable factors for bus optimization and management are; the total number of buses required, passengers' count, bus-capacity, route and trip management [1]. Depending on all these factors, optimized bus scheduling is considered a complex task. Data-mining algorithms play a vital role in solving such complex issues [2].

For optimized bus scheduling; bus-timetables are eagerly required as the passengers need those regarding arrivals and departures. A good bus-system must be able to find the optimal number of departures (frequencies) from the point of departure to arrival. The bus-timetables are made in accordance with passengers' demand and existing constraints; in off-peak time, the number of departures will decrease, and during rush-hours departures will increase. If fleet-size is not managed according to the number of passengers, then passengers will be un-served. During off-peak times, several buses will be more than the demand and result in enormous operating costs. To make a timetable with optimized departures without computer programming needs high processing-time and high-accuracy. Bus companies, most often improve the timetables when faces the changes in passengers' demand and operations' constraints. So, high processing time is needed for making a timetable, which affects the operational cost and service quality of a bus company [3].

From the aforementioned discussion, optimization of bus scheduling depends on factors like passengers' count and minimum buses required. To maximize the profit, the fundamental part is to have the equilibrium between passengers and the buses that company has. Any fault in the equilibrium will result in:

The associate editor coordinating the review of this manuscript and approving it for publication was Congduan Li.

TABLE 1. Data set features.

Date	Day	Route	Time	Total_ Bus_Seats	Booked _Seats	Is_Bus _Dropped	Single_Seat _Fare	Bus _Type	Fare_Basic	Fare_Total
20160604	Saturday	LHR-RWP	0100	39	38	N	1270	1	49530 PKR	48260 PKR
20160715	Friday	LHR-RWP	0145	39	33	N	1300	1	50700 PKR	42900 PKR

1) Excess of buses and fewer passengers.
2) Excess of passengers and fewer buses available

This will result in larger waiting time. Passengers' count can be obtained by using forecasting methods and it will provide valuable results for making decisions. Forecasting is used for predicting undefined thing; there exist more than 300 different methods and based on their respective characteristics, it can be divided into qualitative and quantitative forecasting. The quantitative forecasting can also be sub-divided into time-series and regression analysis [4], [5].

Time-series analysis; being a prominent artificial intelligence-based forecasting technique, is a sequence of vectors depending upon time. It is used when the prediction is required based on the historical data. Time-series analysis is a type of uni-variate method and in the context of bus-system; it can be used to predict passengers' demand. This technique uses passengers' demand over time as a variable and does not use any other variable [6]. The demand varies because of change in other various factors and methods of time-series analysis have the disadvantage that they cannot use them.

Regression-analysis urges to find a functional relationship among predictor and response or dependent variable, explanatory and independent variable. Regression-analysis is further divided into univariate-regression when dealing with one response variable and multivariate regression or with two or more response variables. The number of software's is available for regression analysis such as SPSS, NCSS, SAS, and SIMCA. Moreover, with the advances in intelligent algorithms, prediction methods have also been improved from traditional methods to genetic-algorithm, fizzy-sets, rough-sets and neural network algorithm etc [7].

After predicting the accurate, real-time and reliable passengers' count, a bus company needs an optimized-timetable for bus-departures. For this purpose, Genetic algorithm (GA) is performed on the data for finding the optimized departure schedules. The Genetic Algorithm is an artificial intelligence-based algorithm that can be used for optimizing the departure schedules. It helps the data evolving gradually to a status closed to the optimal solution. In our work, we have conducted a study using the past three years' dataset, obtained from a renowned bus service company.

Concluding all, our proposed study has the following contributions:

1) Regression-analysis has been used to forecast the passengers for a day.
2) In addition to the forecasted users, we have calculated the minimum number of different types of buses required and, in this way, optimized the timetable for departure schedule through a Genetic Algorithm.
3) Huge data helps in forecasting the passengers accurately, that's why we have used data set of a bus company that is generated by 500 buses over a period of 3 years, data set is represented in the Table 1. Empirical result shows, models that are used in this paper can efficiently and effectively predict the potential passenger.

However, we have calculated the optimized bus departures for a single day on two routes. In the same way this model could be applicable on all other routes as well. The rest of the paper has been organized as follows. Section 2 summarizes the literature-work related to regression analysis and Genetic Algorithm. In section 3, the forecasting of potential passengers' count has been calculated. In section 4 and 5 based on the forecasted passenger's minimum number of buses required are calculated and based on the number of buses optimized departure schedule is generated using a Genetic Algorithm. In section 6, tests have been applied to the results obtained from the regression analysis. In section 7 results have been concluded and discussed. Moreover, the direction for future work has also been stated.

II. BACKGROUND

In this section, we review some of the concepts related to Multi-objective transport system using big data. These concepts are used in later Sections of this paper. It will be quite helpful for the reader to understand the proposed scheme.

A. MULTIVARIATE LINEAR REGRESSION

Linear regression is considered the key aspect as an autonomous variable to describe the variations in the dependent variable. Different studies reveal that some significant factors are involved which are accountable for changes in the dependent variable. So there is a need to use more than one factor as independent variables to illustrate the changes in the dependent one, which is also recognized as multiple linear regression [8]. Multiple linear regression analysis is the natural generalization of simple linear regression, used to evaluate the connotation among more than one independent variable to a single dependent variable. The mathematical representation of MLRA is given below:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

setting up $\hat{y}$ as the projected value of dependent variable, whereas β_0 is the constant and $x_1 \ldots x_k$ represents the distinct predictor independent variables. However, β_0 becomes the value of $\hat{y}$ when all values from x_0 through x_k are zero while regression coefficient are $\beta_1, \beta_2, \ldots \beta_k$. Estimation of multiple linear regression is the same as linear regression.

The change in $\hat{y}$ signifies by each regression coefficient comparative to a one unit change in the respective independent variable.

B. OBJECTIVE FUNCTION

Objective function refers to the function that needs to be either maximized or minimized in the particular optimization problem. Thus, this expression can be used in the mathematical optimization context. For instance, in the context of machine learning a model $\mathcal{M}$ is defined. To train $\mathcal{M}$ need to define a loss function, which needs to be minimized. This $\mathcal{L}$ could represent the objective function of the respective problem. Likewise, in search algorithms, objective function will be the cost of a solution such as in travelling salesman problem (TSP), the objective function is represented by $\mathcal{C}$, which sums up all the edges weight on tour. Hence, one could find the inexpensive tour which is associated with minimization of $\mathcal{C}$.

C. GENETIC ALGORITHM

Genetic algorithm represents an intellectual utilization of random search that provides a solution to the optimization problems by determining such input values which results in best outcomes. GAs is adaptive heuristic search algorithm, they also provide a solution for both constrained and unconstrained problems that are based on natural selection. GAs is designed to exploits the historical shreds of evidence in order to direct the exploration into the better performance region within the search space. Moreover, they stimulate the process in natural systems that are prerequisite for progression. GAs is considered better than conventional Artificial Intelligence, as it is comparatively more robust [9]. Unlike AI schemes, on changing input values or in the presence of noise they do not break easily. In addition to this, it provides substantial benefits compared to other existing optimization techniques such as linear programming, depth-first, breadth-first, and praxis etc. However, it is based on the analogy of the structure and behaviour of the chromosome within individuals' population. GA is consisting of three basic operations: selection, genetic operations, and replacement. Being an optimization algorithm, its advantage relative to traditional gradient-based algorithms lies in its ability to locate the global minimum and also operate with discrete or integer design variables.

D. ENCODING

Encoding is the initial part of a GA process, as problem-related information is encoded into a structure; chromosome or string. A chromosome can generally be defined as a variable's sequence of a problem placed in an organized manner.

E. FITNESS EVALUATION

Basically, GAs mimics nature's survival-of-the-fittest principle for performing a searching process. That is how GAs is naturally suitable for solving maximization problems where a fitness function is maximized. The following fitness function is widely used:

$$f(y) = 1/1 + f(y)$$

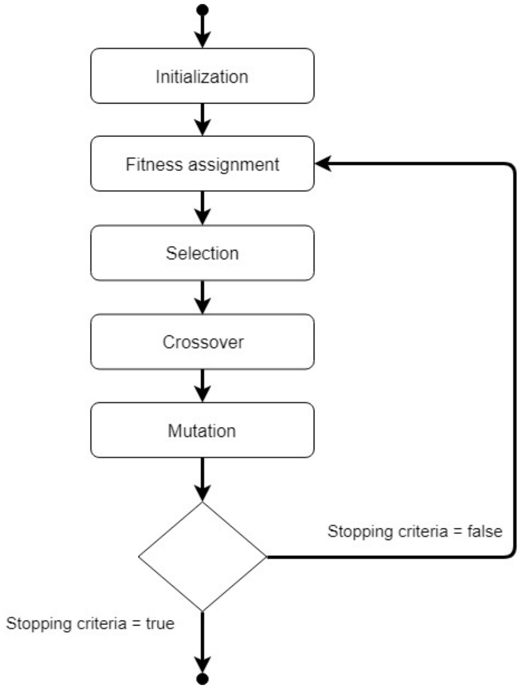


FIGURE 1. Workflow of Genetic Algorithm.

F. GENETIC OPERATIONS

The Genetic Algorithms' operations initiate with a population of random strings showing the design variables. Each string is then evaluated to find the fitness function. There are three main operators; reproduction, operators, crossover, and mutation, to create a new population of points. These operators can be elaborated as follow:

1) REPRODUCTION

Being the initial operator applied to a population. It chooses the good strings in a population and generates a mating-pool.

2) CROSSOVER

As the reproduction process terminates, the mating-pairs get selected. The crossover process mimics the idea of genetic information exchange between male and female.

3) MUTATION

It refers to a random change in chromosome or string's information. Basically, it defines a chromosome's variation. The variation can either be local or global. The GA offers better global searching capability. The working flow of GA is as follows:

III. LITERATURE REVIEW

In recent years, with the progress of operational research and system-engineering, new concepts have been put forward;

in the optimization of forecasting passengers and generation of optimized schedules to improve the transit network. This section encompasses some representative methods:

Osman *et al.* [1] have proposed a deterministic optimization model (D-FSSB) for selecting and scheduling a fleet of buses in the private transit network (PTN). To improve the PTN, integer-programming formulation tool has been used for the decision-making process which further optimizes the criterion function by determining the number and size of buses allocated in particular interval to a specific trip. Furthermore, the formulation of the proposed model has been expressed in general algebraic modeling system (GAMS) 22.6.

In [10], the authors have presented a model that combines the multiple regression principles (MRP) and multivariate time series analysis(MTSA) that overcomes the random factors of time-series. This approach has improved the forecasting accuracy and escalates the prediction reliability. They reported that their model outperforms when used with known factors.

Chebbi *et al.* [11] have addressed the problem of the squandered capacity of transport in the PRT system by presenting a two-tier transit model. Further, a hybrid multi-objective GA is designed that provides a solution to the proposed problem. It is assimilated with multiple crossover operators and linear programming methods. Additionally, a study is conducted on relevant multi-objective unfilled automobile redistribution problem that endeavours to reduce the vacant movement and number of automobiles used. The evaluation and simulation results of the proposed system are promising.

In [12], the authors have developed a novel technique for rural transit management that takes into account two factors: cost and equity in multiple objectives. However, a heuristic procedure is built on data enclosure analysis that is used for characterization and evaluation of route design. For extensive data processing that is vital for analysis; a GIS-based decision support system is generated. The proposed approach is implemented by Quinte Acess, a renowned organization supports to remodel automobile routes. This will give new insight regarding transit service operation in rural regions.

In airport operations, for appropriate operating plans forecasting passengers is a critical aspect. To address this issue the authors in [6], have presented a model for forecasting short-term air passenger demand from search engine queries using big data. To ensure predictive ability, time shifts ranging from 0-11 months at 1-month interval are used to develop this forecast model. The results by applying simple regression analysis on the optimal model demonstrate that slope and intercept are significant and displaying a coefficient of 0.886. The Error rate between actual values and forecasted values ranged from 2.96 to 11.01% with a mean error of 5.30%. The results found that the model is promising and could be used to forecast short-term passenger demand.

Smith *et al.* [13] have achieved the constructive results by applying t-Test on two-step clustering framework to time-series. In every possible combination, both K-Means and Fuzzy Art have been applied collectively and individually. Their results have outperformed the previous state-of-the-art techniques used in isolation. In [14], the authors have presented a prediction model for Bus passenger capacity in which time-series data has been assorted into regression analysis predict. Integration of these two forecasting procedures; in this study collectively contains the merits between them. The respective model reflects the pragmatic condition of bus passenger and better fits practical data.

In [17], the authors have improved the forecasting methods by using the Mean Absolute Deviation (MAD) and Mean Absolute Percentage Error (MAPE). Both are used to compute the percentage of mistakes in the least square (LS) methods in 9.77%. It decided that the least square method is functioned for time-series and trend data. In [4], they have presented a method for time-table optimization that is based on hybrid automobile size to handle the bus demand instability. Moreover, it has solved the issues faced by the heuristic algorithm. The outcome has suggested that the hybrid model outperforms against small and large automobile size models. However, the downside of the proposed approach has no capacity and quantity restrictions on automobile size.

Hairong *et al.* [19] have presented a multi-object GA to achieve a better public transit plan. The optimization of the model has performed in such a manner that minimizes the transfer time of passengers on different stops by satisfying the constraints such as traffic demand, maximum departure time and minimum headway. However, it is a mixed integer non-linear programming problem which has hard to solve by classical techniques. For this reason, modifications have been made and most proficient outcomes achieved from an improved genetic algorithm (IGA). IGA imports the simulated annealing algorithm for the enhancement of premature and slow evolution speed phenomenon. This study has also revealed that after applying tests on the data ensures positive and effective results that leads towards reasonable decisions.

Sun *et al.* [18] have conducted a study on head-way optimization and scheduling combination of bus rapid transit automobiles. The proposed model has following features that include minimum passenger's travel cost, automobiles operation cost and constraints encompass a volume of passengers, time and frequency. Their scheduling combination has comprised of the normal, zone and expresses scheduling. In normal-scheduling, the automobile run along routes and stops at every station from initial stop to end, in zone-scheduling automobiles runs only on a high-traffic-volume section while the express vehicle stops at certain stations. The respective model was solved by using Genetic Algorithm and proven scientifically. They reported that their optimized results can save 69.92% cost. However, the study has following limitations e.g. vehicle has run at constant speed, uniform passenger arrival rate and have a sufficient number of vehicles.

In [3], the authors have formulated an integer programming model for an optimized time-table. This model was developed by feeding route data, number of passengers, number of fleets, vehicle capacity, departure time, travel time, headway, employees active hours, vehicle revenues and operation cost into it. The model results in 99.10% accuracy with 99.99% algorithm efficiency. In [25] they have conducted a study on simple and multiple linear regressions along with quadratic regression analysis on hourly and daily statistics from research house. Likewise, they have obtained an improved coefficient of determination by using out-door temperature and solar radiation in the multiple regression model. The results were promising after applying the selected model.

Zhu [20] has constructed an ARIMA model for forecasting daily passenger-flow. The iterative and recursive algorithm used by the proposed model to computes 2-new sequences of daily capacity individually. The experimental assessment depicts that the relative error (RE) of recursive forecasting is smaller than 0/iterative. Sengupta and Gupta [21] have presented a nonlinear programming model in order to determine fleet selection and an optimal automobile-mix whereas criterion optimization function encompasses various components such as fleet operational cost, implicit passengers' waiting cost, per trip and automobile revenue. Bookbinder and Edwards [22] have developed an optimization model to resolve the scheduling problem that targets pupil transit with common source and destination.

In [15], the authors have modeled the automobile routing problem with time windows by considering a tour as a virtual stop. Besides, they proposed 2-assignment problems; exact approaches for special-cases and a heuristic algorithm for general cases. In [2] the authors have evaluated the urban-level spatiotemporal features of resident tours from real GPS statistics of a cab in the Beijing city. Based on these features they have proposed the Multi-objective Bus Route Plan Model relied on the ant-colony algorithm. To overcome the problem of computational complexity caused by traditional density-based clustering algorithm has been replaced by a two-stage clustering algorithm. The empirical result shows that the proposed approach has provided a better solution to bus routes with robust transit and minimized mobile cost.

However, Fink *et al.* [16], have addressed the problem of vehicle routing by employing multiple synchronization constraints. The proposed AVRPWVS deals with routing workers in order to handle the ground job while meeting each job time window such as offloading luggage or refilling the fuel in a jet. Moreover, two mathematical multi-commodity flow formulations have presented which rely on time-space networks. The results found that the proposed model provides an optimal solution comparative to the two-integer model. In [23], they have developed a multi-object genetic optimized model which has involved passengers and bus corporations. They have obtained the optimal results of bus departure interval by using the genetic algorithm.

By thorough review, we have found the following limitations that need to be addressed such as no vehicle constraint, identical buses; passenger arrival is uniform at the stops, same fare for all the routes and no vehicle capacity constraint. If the fare is same for all the routes then cost-benefit analysis will not provide accurate results as the bus has to be deported only if the revenue is more than the cost of the bus that company is enduring. Passenger arrival rate can be different on different days which mean bus company needs a different number of buses on different days, a solution cannot be optimized until company know the potential passenger that can on a particular day. Optimizing departure schedules for a bus company having unlimited number will be an ideal solution, as companies have limited number of resources and these resources need to be managed in a way that buses should depart having a minimum number of empty seats. Due to the above limitations, more exertion is required in the area of generating optimized bus departure schedule.

IV. FORECASTING PASSENGERS FOR BUS DEPARTURES

For creating an optimized timetable for departure schedule, a bus company first need to know; how many potential users can come for a particular day. Secondly, the management of resources according to passenger counts. Therefore, statistical approaches such as time series and regression analysis have been applied to the historical data of a renowned company. After computing the forecasted users and number of buses; optimized time-table generated by applying GA.

However, the relation among proposed forecasting model and optimized bus scheduling algorithm is that forecasting model forecasts the passengers and these predicted passengers are used as an input in optimized bus scheduling.

Hence, in this study, we have predicted passengers for a particular day using Uni-variate multiple linear regression analysis on historical data. Accordingly, we have taken the past three years of data from a renowned company that provides the basis of an accurate forecast. Dataset features are given in Table 4.

In Table 4, Date, Day and Time represents the time of departure. Total_Bus_Seats shows the total number of seats bus has. Booked_Seats represents the total number of seats that are sold for that particular day on a specific time. Is_Bus_Dropped shows if the bus is dropped or not. Bus_Type represents the type of bus; its value is 1 for 32 seat bus, 2 for 39 and 3 for 44 seat bus. Fare_Basic value is obtained by the multiplication of Single_Seat_Fare to total number of seats sold (Booked_Seats).Whereas, Fare_Total is the amount that we have got after multiplication of Total_Bus_Seats with Single_Seat_Fare. Figure 2 demonstrates the passengers' trend of the last three years on different weekdays.

A. UNI-VARIATE MULTIPLE LINEAR REGRESSION

Uni-variate multiple linear regression is generalized form of simple linear regression technique, see eq 2. This technique allows more than one independent variables as shown in Eq (1) while Table 2 represents symbols description:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e \qquad (1)$$

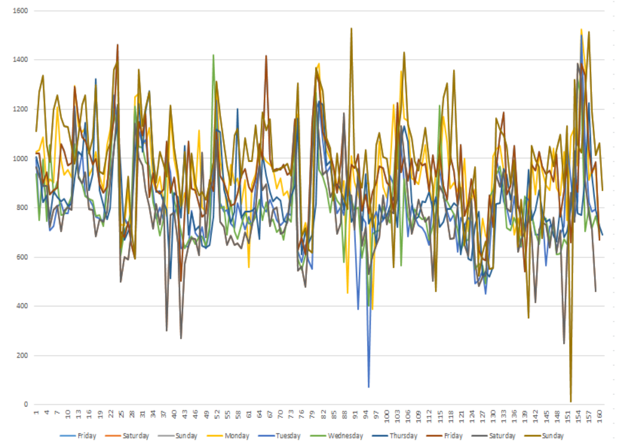


FIGURE 2. **Passengers' Trend on Weekdays.**

TABLE 2. **Symbols Description.**

Symbol	Description
Y	Regression Variable
X	Predictor Variable
β	Regression Coefficient
P	Number of Variables
e	Error

In the proposed solution we have modified the above mentioned Eq (1) as follows:

$$Y = Intercept + Xcode \times \mathrm{X}_1 + ND \times \mathrm{X}_2 + HD \times \mathrm{X}_3 + LW \times \mathrm{X}_4 \qquad (2)$$

whereas, Xcode represents the coding given to the total number of observations, ND signifies the segregation of day on which passenger count is moderate. However, HD represents the days when passenger count is higher than ND while LW represents the days on which passenger count is higher than HD.

1) *NORMAL DAYS (ND)*

The term ND symbolizes week days on which the passenger count is moderate, which includes Tuesday, Wednesday and Thursday. Since passenger commute less in these days therefore, rate of passengers' is moderate comparative to other days as shown in Figure 3 and Figure 4.

2) *WEEKENDS (HD)*

The respective term explains the days of week on which passenger count is relatively higher than ND including Friday, Sunday and Monday see Figure 2 and 3 for passenger count on these days.

3) *LONG WEEKENDS (LW)*

LW represents long weekend, it means when a normal weekend is prolonged from 2 to 3 days or more. It include, day before the start of LW, day on which the LW has started, day on which LW is ended and a day after LW ends. On long weekends passenger count is higher than HD and ND. In Figure 3, week 1 has a long weekend ranges from Friday to Monday.

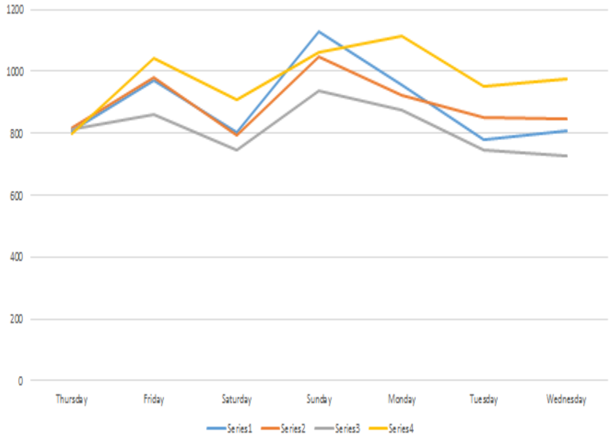


FIGURE 3. **Trend on Normal Weekdays.**

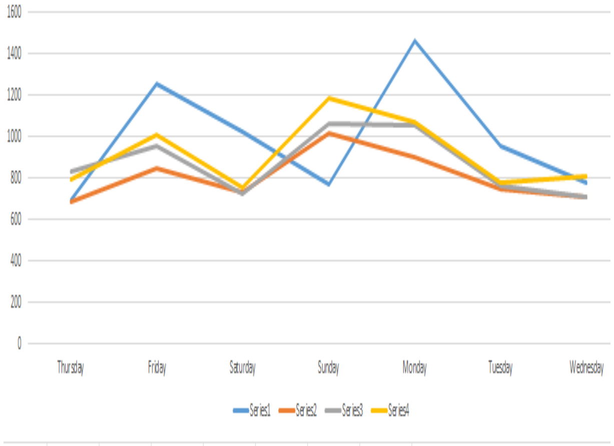


FIGURE 4. **Trend on one-long Weekend & three Normal Weekends.**

After applying uni-variate multiple linear regression on the dataset of past three years for Route-1, obtained results are shown in Table IV-A3. For Other It will be applied in the same way it is applied for Route-1.

tableMultivariate regression analysis results

Regression statistics			
R^2		0.22	
F		78.2	
	Coefficients	**t Stat**	**P-value**
Intercept	828.8934045	51.0123	6.3488E-294
Xcode	-0.091347811	-5.76434	1.06002E-08
ND	30.16311895	1.918363	0.055319946
HD	167.1777423	10.6069	4.14368E-25
LW	231.6697649	7.826143	1.16219E-14

By putting values in Eq (2) from Table 2, we get (3)

$$Y = 828.9 + (-0.09) \times \mathrm{X}_1 + 30.16 \times \mathrm{X}_2 + 167.17 \times \mathrm{X}_3 + 231.66 \times \mathrm{X}_4 \qquad (3)$$

TABLE 3. Symbol definition.

Symbols	Names
Cob	Cost of bus
Wt	Waiting time
Tax	Road tax
Dc	Driver commission
Bhc	Bus hostess commission
Lp	Cost of lunch packets
Dr	Cos of drinks
Cof	Cos of fuel per bus trip
Ods	Optimized bus departures
Mbr	Minimum buses requirement

V. OPTIMIZED DEPARTURE TIMETABLE

In this section, we have calculated the minimum number of buses required and the creation of an optimized timetable for departure schedule through a Genetic Algorithm.

A. MODEL FORMULATION

Different variables description that are used in model formulation are summarized in Table 3.

1) OBJECTIVE FUNCTION

Objective function is composed of following parameters such as waiting time; minimum buses required for a particular day and cost of the bus for a particular trip that company has to endure. Total time span is of 24 hours which is converted into minutes. Where minimum_buses_required() is the method in which forecasted bus passengers are used as an input.

$$
\begin{aligned}
Ods &= Cob + Mbr + Wt \\
Cob &= Dc + Bhc + Lp + Dr + Cof \qquad (4) \\
Mbr &= Minimum\ Buses\ Requirement \\
&\quad \times (Forecasted\ Passengers) \qquad (5) \\
Wt &= 24 \times 60 \div Mbr \qquad (6)
\end{aligned}
$$

2) COST OF BUS

The cost of the bus is calculated by summation of all the cost that the company has to pay for a particular trip that includes the commission of the driver, lunch-packets cost, drinks and bus fuel. This will give us the minimal amount that company has to endure for a particular route.

3) MINIMUM BUSES REQUIREMENT

Three types of buses are available by the company having 32, 39, and 44 seating capacity. Such combinations of buses should be made hence there should be no or minimal empty seats left. Whereas Cost-Benefit Analysis (CBA) is performed on the bus which has empty seats and based on the outcome decision is made, whether the bus will depart or not.

To generate the combination of buses backtrack subset sum algorithm [24] has been used and this will give us a combination of buses in ascending order.

In the Algorithm 1, if the subset sum algorithm is not able to generate the buses for forecasted users than the value of users are saved in a temporary variable and its value is increased until the algorithm is able to find a combination. From this combination, the smallest bus is selected because empty seats are added and the algorithm performs a cost-benefit analysis to check if the bus will depart or not.

4) WAITING TIME

Firstly, calculate the total time in minutes in order to compute the waiting time between departures. Each day has 24 hours and these hours are converted into minutes then divided by the total number of buses required on a particular day. The obtained result provides the minimum waiting time among bus departures.

B. CONSTRAINTS

1) COST BENEFIT ANALYSIS

A bus cannot depart if the revenue from a specific bus on a particular route is less than the cost of the bus (Cob) that the company is enduring.

$$Cob > Passengeronbus \times Fare.$$

2) NUMBER OF BUSES

The total number of buses required for the passengers all day long is also a constraint because if the number of passengers exceeds than the total seating capacity of the buses then these passengers will be un-served.

C. BOUNDARY CONDITION

When the total number of optimized schedules is equal to the minimum number of buses which are required for a particular day, then the algorithm stops for that particular day.

VI. GENETIC ALGORITHM BASED SOLUTION

A Genetic Algorithm is a heuristic search that has inspired by Charles Darwin's theory of natural evolution. It is a method of solving both unconstrained and constrained optimization problems which are based on natural selection. In addition to this, it is commonly used to solve optimization problems that are not suited for standard optimization problems and in which the objective function is discontinuous, non-differential, stochastic, or highly nonlinear. However, Genetic Algorithm is comprised of following steps such as initialization, parameter selection, fitness value and genetic operator etc.

Whereas, total bus cost is represented by λ, κ is used to find bus cost that company is bearing, σ typifies the $backTrackSubSetSum()$ while ω represents the $potentialUsers$ = Forecasted Bus Passengers.

VII. MULTI-OBJECTIVE TRANSPORT SYSTEM REGRESSION-ANALYSIS & GENETIC-ALGORITHM MODEL

The solution of the proposed model is relying on its implementation. By using Genetic Algorithm we have converted all variables that are listed below from phenotype to genotype. Based on concrete data, a long number string is obtained

Algorithm 1 Find Combination of Buses

```
procedure FCB($potentialUsers[]$)
    $int \rightarrow tempUsers \rightarrow 0$
    $array \rightarrow buses[]$
    $int \rightarrow bus \rightarrow 0$
    $int \rightarrow emptySeats \rightarrow 0$
    $int \rightarrow Fare = 1600$
    $int \rightarrow \lambda \rightarrow 0$
    $int \rightarrow revenue \rightarrow 0$
    $isEmptySeatsAdded \rightarrow false$
    for i =1 to $\omega$ do
        buses=$\sigma$($\omega$[i])
        if(buses.length==0)

        {
        tempUsers = $\omega$ + 1
        while $buses.length == 0$
         buses=$\sigma$(tempUsers)
         tempUsers++
        isEmptySeatsAdded = true
        endwhile

        }
        if(isEmptySeatsAdded)
        {
         $\lambda$ = $\kappa$()
         revenue = Fare $\times$ $\omega$[i]
        if(revenue<=$\lambda$)
        {
        buses.remove(0)
        }
        }
        }
    end for
end procedure
```

that encompasses date, departure time, passenger count and bus seating capacity. Therefore, multi-parameter encoding technique is used for such data.

The largest bus that company owned has 44 seats so if 44 are taken as 100% then the length of the string of binary grid spacing in each departure is at least 7. However, if the day is divided into a K number of departures then the chromosome length will become 7K. Once encoding is reversed, the process of decoding is initiated and binary code is converted into a decimal number and this volume reveals the actual bus occupancy [4], [23].

A. OBJECTIVE FUNCTION

The fitness function has used in the proposed model to minimize the expenses and for optimized departure schedules. The fitness function is used here $F(x) = f(x) > Cob$ whereas $f(x)$ is objective function and Cob represents the cost of bus. The model will be considered fit if the value of the objective function is more than the cost of the bus.

B. SELECTION PROCESS

Chromosomes need to be selected from the population in order to give input to a crossover. Good chromosomes are survived and inherited to create new offspring. The method used in the proposed study is the roulette wheel selection. When the population size is n, the fitness function of an individual I is Fi, then the selection probability is Pi.

$$Pi = \frac{Fi}{\sum_{i=1,l=1}^{N} \mathrm{Fi}}$$

where Pi illustrates the fitness level of individual I in the sum fitness of the entire group, the greater the fitness value more become the chances of selection.

C. CROSSOVER

In a GA, the crossover is one of the core operators, by crossover, it is meant that a group of two parent individuals are reshaped in such a way that they are exchanged and recombined to the corresponding genome on the chromosome which results in new individuals. The crossover is applied according to the generated exchange cross rate. The rate at which the crossover happens explains the frequency of crossover, and the results we get for optimization are faster when crossover frequency is high. The value of the crossover rate generally lies between 0.4 to 0.9 [23] whereas we have selected 0.6. Moreover, their exit different crossover types such as one point crossover, two crossovers, multi-point crossover and uniform crossover etc. One-point crossover for the recombination of genes is used in this study.

D. MUTATION

Mutation brings the genetic change at the chromosome level. A simple example of the binarily coded population can be taken as a change of $0 \rightarrow 1$ or $1 \rightarrow 0$. Just like crossover, the selection is also a problem relating to the variation rate in mutation. The value of the mutation is affected by the size of the population, length of the chromosome and other factors. The general range of mutation which starts at 0.001 and ends at 0.1 [23]. The selected mutation rate in the proposed study is 0.01.

VIII. RESULTS & DISCUSSION

Empirical results in Table 3 illustrates that all variables have a significant effect on a number of passengers (dependent variable). The regression equation includes explanatory variables such as coding mechanism (Xcode), moderate passenger count (ND), high passenger count (HD) and long weekend (LW). The result demonstrates the effect of ND, HD and LW is positive and significant on the dependent variable, whereas the effect of coding is negative but it remains significant throughout the analysis. If all the factors remain constant then intercept will remain 828.89 which are also significant. On the whole, explanatory factors are significantly effecting

TABLE 4. Actual and forecasted passengers for route-1.

Route-1			
No	Day	Actual passengers	Forecasted passengers
1	Day 1	692	757
2	Day 2	1257	1125
3	Day 3	1022	958
4	Day 4	770	726
5	Day 5	1461	1124
6	Day 6	955	988
7	Day 7	775	756

TABLE 5. Actual and forecasted passengers for route-2.

Route-2			
No	Day	Actual passengers	Forecasted passengers
1	Day 1	850	792
2	Day 2	655	723
3	Day 3	1225	1095
4	Day 4	701	740
5	Day 5	1165	1110
6	Day 6	935	892
7	Day 7	1092	904

at conventional standards. Value of R^2 ranging from 0 to 1 and 0.22 means 22% of the dependent variable is explained by independent variables. Finally, a value of $F(78.2)$ shows that the overall model is fit at the conventional standard.

However, a low or high value of R^2 does not mean it fits the model well. Thus, a detailed analysis is needed to ensure that the model is satisfactory to use [25].

Mean absolute deviation (MAD) and Mean absolute percentage error (MAPE) are other parameters to compute the quality of the model. Errors in forecasting are common and almost all the methods used for forecasting have errors in the results. The errors in forecasting which use regression analysis will occur frequently are calculated using MAD and MAPE. These procedures give choice to an organization to consider the utilization of the method used for prediction [17]. Actual number of passengers who came for seven days are shown in Table 3.

Eq (1) is used for predicting the passenger count for seven days. Table 8 displays the predicted passengers' count for Route-2.

Based on original data contained in Table 8 and forecasted results using regression analysis in Table 8, error calculation is performed by using MAD and MAPE. After applying formulas available in [14] on Table 8 and Table 8, value for MAD is 99.14 for Route-1 and 83 for Route-2 and for MAPE is 8.7% for Route-1 and 8.5% for Route-2. This value of MAPE is rational as it is not exceeding 10% margin of error.

After testing the predicted results, minimum number of different types of buses needed for particular day is calculated, these buses are calculated based on the forecasted passengers in Genetic Algorithm by using Algorithm 1. see Table 8.

Table8, shows the optimized bus departure time table for Route-1 and Route-2, it is generated by using Genetic Algorithm. It has specified the number of buses needed in one day and waiting time among departures.

TABLE 6. Actual vs forecasted passengers for route-1.

Route-1			
No	Actual Values	Forecasted Values	lAt-Ftl
1	692	757	65
2	1257	1125	132
3	1022	958	64
4	770	726	44
5	1461	1124	337
6	955	988	33
7	775	756	19

TABLE 7. Actual vs forecasted passengers for route-2.

Route-2			
No	Actual Values	Forecasted Values	lAt-Ftl
1	850	792	58
2	655	723	68
3	1225	1095	130
4	701	740	39
5	1165	1110	55
6	935	892	43
7	1092	904	188

TABLE 8. Different type of buses for route-1.

Route-1			
Days	No. of 32 Seat Buses Required	No. of 39 Seat Buses Required	No. of 44 Seat Buses Required
Day 1	11	7	3
Day 2	15	3	12
Day 3	15	10	2
Day 4	16	6	1
Day 5	15	12	4
Day 6	15	4	8
Day 7	16	4	2

TABLE 9. Different type of buses for route-2.

Route-2			
Days	No. of 32 Seat Buses Required	No. of 39 Seat Buses Required	No. of 44 Seat Buses Required
Day 1	13	4	5
Day 2	11	5	4
Day 3	15	9	6
Day 4	12	8	1
Day 5	15	6	9
Day 6	14	8	3
Day 7	13	8	4

Figure 5, shows total number of buses needed for the whole week for Route-1, while Figure 6, shows total number of buses required for 7 days for Route-2.

A. SENSITIVITY ANALYSIS

1) IMPACT OF CUSTOMER VOLUME ON BUSES CATEGORY

The bus company can run a maximum of 45 buses in one day. However, 15 buses with 32 seating capacity, the other

TABLE 10. Combination of buses for route-1 & route-2.

S.No	Route-1 Departure Time	Route-2 Departure Time
1	0:06	0:00
2	1:08	1:08
3	2:38	2:16
4	3:59	3:24
5	4:59	4:32
6	05:55	5:40
7	06:51	6:48
8	07:52	7:56
9	09:11	9:04
10	10:12	10:12
11	11:42	11:20
12	13:00	12:28
13	13:54	13:36
14	14:57	14:44
15	15:52	15:52
16	17:22	17:00
17	19:03	18:10
18	20:29	19:20
19	21:28	20:28
20	22:37	21:36
21	23:39	22:50
22		24:00

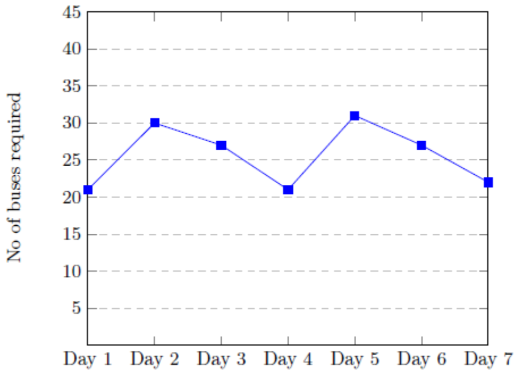


FIGURE 5. Buses Required for 7 Days for Route-1.

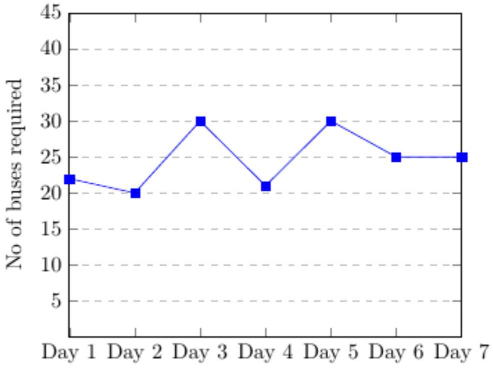


FIGURE 6. Buses Required for 7 days for Route-2.

pair of 15 have 39 and 44 seats respectively. If the passenger volume exceeds after filling all the buses than the excessive passenger will be dropped alternatively as in Table 11 for Route-1, we can see that for day 7, 16 (32 seater bus) buses are required which exceed the total limit of the company for that particular type. Therefore, one bus from this type needs to be removed and add the next smallest bus, in this case, it is 39 seater bus. It is obvious that this bus will have 7 empty seats which will result in expense.

TABLE 11. Combination of buses.

	No. of 32 Seat Buses	No. of 39 Seat Buses	No. of 44 Seat Buses	Empty Seats
1st Solution for Day7	16	4	2	0
2nd Solution for sssDay7	15	5	2	7

IX. CONCLUSION

Optimized bus scheduling is the main focus area for every transit company. This study is done in the context of predicting potential passengers; likewise, we have also calculated minimum buses required for a particular day. Based on the required buses optimized timetable for departures has been produced. However, for forecasting potential passenger's univariate multi linear regression has applied over past 3 years data that is obtained from the renowned company whereas Genetic Algorithm is utilized for finding the minimum buses required and for creation of optimized table. Furthermore, we have applied absolute deviation (MAD) and Mean absolute percentage error (MAPE) tests on the results of univariate multi linear regression. The value of MAD (99.14) and MAPE (8.7%) shows that the results are promising and this model can be applicable in different transit companies.

For future work the day can be divided into different groups to develop timetable by finding peak and off-time

ACKNOWLEDGEMENT

The authors would like to thank Ms. Iqra Mustafa, who provided her suggestions related to the English proof reading of the paper and contributed to improve the quality and readability of the paper.

REFERENCES

[1] M. F. S. Osman and M. M. Al-Sanousi, "A deterministic IP model for optimizing bus scheduling in a private transportation system," in *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manage.*, Dec. 2015, pp. 244–248.

[2] S. Lei, Z. Li, B. Wu, and H. Wang, "Research on multi-objective bus route planning model based on taxi GPS data," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery (CyberC)*, Oct. 2016, pp. 249–255.

[3] F. D. Wihartiko, A. Buono, and B. P. Silalahi, "Integer programming model for optimizing bus timetable using genetic algorithm," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 166, no. 1, 2017.

[4] W.-X. Sun, T. Song, and H. Zhong, "Study on bus passenger capacity forecast based on regression analysis including time series," in *Proc. Int. Conf. Measuring Technol. Mechatronics Automat. (ICMTMA)*, vol. 2, Apr. 2009, pp. 381–384.

[5] M. Chen, X. Liu, J. Xia, and S. I. Chien, "A dynamic bus-arrival time prediction model based on APC data," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 19, no. 5, pp. 364–376, Sep. 2004.

[6] S. Kim and D. H. Shin, "Forecasting short-term air passenger demand using big data from search engine queries," *Autom. Construct.*, vol. 70, pp. 98–108, Oct. 2016.

[7] C. Z. Chang, X. M. Chen, and M. Wang, "Study on combinational scheduling optimization of bus transit rapid based on tabu search & genetic algorithm," *Appl. Mech. Mater.*, vols. 744–746, pp. 1827–1831, Mar. 2015.

[8] Y. S. Kong, S. Abdullah, D. Schramm, M. Z. Omar, and S. M. Haris, "Development of multiple linear regression-based models for fatigue life evaluation of automotive coil springs," *Mech. Syst. Signal Process.*, vol. 118, pp. 675–695, Mar. 2019.

[9] A. S. Tasan and M. Gen, "A genetic algorithm based approach to vehicle routing problem with simultaneous pick-up and deliveries," *Comput. Ind. Eng.*, vol. 62, no. 3, pp. 755–761, 2012.

[10] Y. Li and F. Ying, "Multivariate time series analysis in corporate decision-making application," in *Proc. Int. Conf. Inf. Technol., Comput. Eng. Manage. Sci. (ICM)*, vol. 2, Sep. 2011, pp. 374–376.

[11] O. Chebbi and J. Chaouachi, "Reducing the wasted transportation capacity of Personal Rapid Transit systems: An integrated model and multi-objective optimization approach," *Transp. Res. E, Logistics Transp. Rev.*, vol. 89, pp. 236–258, May 2016.

[12] C. Chen, G. Achtari, K. Majkut, and J.-B. Sheu, "Balancing equity and cost in rural transportation management with multi-objective utility analysis and data envelopment analysis: A case of Quinte West," *Transp. Res. A, Policy Pract.*, vol. 95, pp. 148–165, Jan. 2017.

[13] C. Smith and D. Wunsch, "Time series prediction via two-step clustering," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–4.

[14] B. Yu, W. H. K. Lam, and M. L. Tam, "Bus arrival time prediction at bus stop with multiple routes," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 6, pp. 1157–1170, Dec. 2011.

[15] B.-I. Kim, S. Kim, and J. Park, "A school bus scheduling problem," *Eur. J. Oper. Res.*, vol. 218, no. 2, pp. 577–585, 2012.

[16] M. Fink, G. Desaulniers, M. Frey, F. Kiermaier, R. Kolisch, and F. Soumis, "Column generation for vehicle routing problems with multiple synchronization constraints," *Eur. J. Oper. Res.*, vol. 272, no. 2, pp. 699–711, 2019.

[17] U. Khair, H. Fahmi, S. Al Hakim, and R. Rahim, "Forecasting error calculation with mean absolute deviation and mean absolute percentage error," *J. Phys., Conf. Ser.*, vol. 930, no. 1, 2017, Art. no. 012002.

[18] C. Sun, W. Zhou, and Y. Wang, "Scheduling combination and headway optimization of bus rapid transit," *J. Transp. Syst. Eng. Inf. Technol.*, vol. 8, no. 5, pp. 61–67, 2008.

[19] Y. Hairong and L. Dayong, "Optimal regional bus timetables using improved genetic algorithm," in *Proc. 2nd Int. Conf. Intell. Comput. Technol. Automat. (ICICTA)*, vol. 3, Oct. 2009, pp. 213–216.

[20] H.-Y. Zhu, "N days average volume based ARIMA forecasting model for Shanghai metro passenger flow," in *Proc. Int. Conf. Artif. Intell. Educ. (ICAIE)*, Oct. 2010, pp. 631–635.

[21] J. K. Sengupta and S. K. Gupta, "Optimal bus scheduling and fleet selection: A programming approach," *Comput. Oper. Res.*, vol. 7, no. 4, pp. 225–237, 1980.

[22] J. H. Bookbinder and S. H. Edwards, *School-Bus Routing for Program Scheduling*. 1990.

[23] D.-R. Tan, J. Wang, H.-B. Liu, and X.-W. Wang, "The optimization of bus scheduling based on genetic algorithm," in *Proc. Int. Conf. Transp., Mech., Elect. Eng. (TMEE)*, Dec. 2011, pp. 1530–1533.

[24] J. J. McGregor, "Backtrack search algorithms and the maximal common subgraph problem," *Softw., Pract. Exper.*, vol. 12, no. 1, pp. 23–34, 1982.

[25] N. Fumo and M. A. R. Biswas, "Regression analysis for prediction of residential energy consumption," *Renew. Sustain. Energy Rev.*, vol. 47, pp. 332–343, Jul. 2015.

MOUBEEN FAROOQ KHAN was born in Islamabad, Pakistan. He received the B.S. degree in computer science from International Islamic University, Islamabad, in 2013, and the M.S. degree in computer science from COMSATS University Islamabad, Pakistan, in 2019.

From 2013 to 2018, he was a Developer with Stafona, Andpercent, and Daewoo. He has command on different languages and tools, such as Java, ASP.NET, Android, iOS, and Oracle database administration on Linux servers. He is a Service/Support Engineer with Trovicor. Besides having academic qualification and professional experience, he has completed trainings in DataBase Administrator, Oracle Data Integrator, SQL, Oracle Business Enterprise Edition, and Linux Administration.

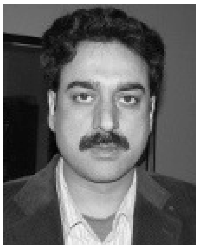
SOHAIL ASGHAR graduated with honors in computer science from the University of Wales, U.K., in 1994. He received the Ph.D. degree from the Faculty of Information Technology, Monash University, Melbourne, Australia, in 2006.

In 2011, he joined the University Institute of Information Technology, PMAS-Arid Agriculture University, Rawalpindi, as a Director. He is currently a Professor and the Chairman of Computer Science with COMSATS University Islamabad. He has taught and researched in data mining (including structural learning, classification, and privacy preservation in data mining and text and web mining), big data analytics, data science, and information technology areas. He has published extensively (more than 150 publications) in international journals as well as conference proceedings. He has also consulted widely on information technology matters, especially in the framework of data mining and data science.

Dr. Asghar is also a member of the Australian Computer Society (ACS) and the Higher Education Commission Approved Supervisor. In 2004, he received the Australian Postgraduate Award for Industry. He is on the Editorial Team of well-reputed scientific journals. He has served as a Program Committee Member of numerous international conferences and regularly speaks at international conferences, seminars, and workshops.

MANZOOR ILLAHI TAMIMI received the master's degree in computer science from Gomal University, Dera Ismail Khan, in 1998, and the Ph.D. degree from the Intelligence Engineering Laboratory, Institute of Software, Chinese Academy of Sciences, supervised by Prof. H. Wang. His Ph.D. thesis was titled: Detecting Outliers for Data Stream Under Limited Resources.

In 2009, he rejoined the Department of Computer Science, CUI, as an Assistant Professor, where he is currently an Associate Professor. His major research interests include mining data streams, ubiquitous data mining, intrusion detection, distributed data mining, machine learning and data mining, real-time systems, publish/subscribe systems, wireless sensor networks, active databases, and ECA rules. In 2005, he received the CUI Scholarship for Ph.D. studies at GSCAS, Beijing, China.



MUHAMMAD ASIM NOOR received the Ph.D. degree in computer science from Johannes Kepler University Linz, Austria, in 2008.

He is currently an Assistant Professor with the Department of Computer Science, COMSATS University, Islamabad, Pakistan, where he teaches software engineering, requirement engineering, and software project management at undergraduate and graduate levels.