IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 26, 2021, accepted April 12, 2021, date of publication May 14, 2021, date of current version May 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3080508

A Novel Approach to the Optimization of a Public Bus Schedule Using K-Means and a Genetic Algorithm

YASUKI SHIMA[1], RABIAH ABDUL KADIR[1], AND FATHELALEM ALI[2]
[1]Institute of IR 4.0, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
[2]Department of Management and Information Sciences, Meio University, Nago 905-8585, Japan

Corresponding author: Yasuki Shima (y.shima@meio-u.ac.jp)

This work was supported in part by the Universiti Kebangsaan Malaysia under Project ZG-2018-001, and in part by the Kojin-2019 Research Fund of Meio University.

ABSTRACT With the continuing economic growth of developing countries, the populations of their urban areas are increasing dramatically. In view of this trend, the optimization of bus service scheduling has become an important task. The efficiency of a transport system depends on several different planning processes, and the balance between these elements is rather complex. In this paper, we consider timetabling and vehicle allocation as the bases for our work. With the aim of providing a reliable service to passengers at a reasonable cost, we focus on the optimization of a bus schedule using a method based on K-means and a genetic algorithm. Our approach starts with parameter setting and data preparation, using a dataset of real bus operating schedules. Three elements are identified from this dataset: the time zones in which the bus service operates, the number of stops made by each bus in each trip, and the dwell time at bus stops. K-means clustering is used to identify moderate operation conditions. The outcome of the K-means algorithm is used as the objective fitness value for optimization of the bus schedule using a genetic algorithm. The results of experiments show that the proposed optimization model can improve the dwell time while maintaining the operating cost at its current level or less, and a remarkable increase in the operation rate is achieved in the case study. The proposed model is able to both effectively optimize the financial outlay and enable bus operators to meet passenger demand in a mutually satisfactory way.

INDEX TERMS Bus service reliability, clustering methods, genetic algorithm, K-means, optimal scheduling, optimization, public transportation, transportation.

I. INTRODUCTION

The provision of a reliable and efficient public transportation service for passengers is a challenging problem. In this domain, the planning of an optimal schedule for the transport service is one of the most important challenges. Optimal scheduling should meet the demand from passengers for a reliable service while maintaining the operational cost within affordable levels for the bus operator. For passengers on public transport, the overall satisfaction with the service depends largely on the accuracy of the timetable, on-schedule operation and the dwell time [1]. From the operator's point of view, an effective bus schedule also minimizes the operational cost with an optimal frequency for bus trips, and can respond well to the different levels of demand under peak and non-peak conditions. In this way, it can achieve a balance between supply and demand [2].

The efficiency of a transport system depends on several planning processes. Four different aspects are involved when planning a public transportation system: network design (ND), timetabling (TT), vehicle scheduling (VS) and crew scheduling (CS) [3]. The coordination between these processes is rather complex; planning is extremely difficult, especially for medium to large fleet sizes, and usually requires a separate treatment for each activity [4]. The first two processes in planning a transport system are related to the type and quality of the service to be offered, while the latter two are related to the allocation of resources. In this paper, we consider TT and VS as the bases for our proposed optimization framework. Handling these two processes is

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

complex in terms of solving an integrated model of the two planning operations [4], [5]. Ceder has elaborated on this optimization and the various approaches by which it can be achieved, as well as the complexity associated with TT and VS due to the relationship between these processes [6]. In regard to the theoretical foundation and processes of operational planning decomposition and service standards, Ceder found that passengers are ultimately seeking a reliable service, and that inadequate and/or inaccurate timetables tend to confuse them [6]. He also identified other attributes of passenger demand, including the travel time, service frequency, routing, transfers and other elements that can affect passenger comfort [6]. From the bus operator's perspective, the service provider must respond satisfactorily to the conditions of passenger demand. At peak hours, the number of passengers is high, and based on this number, service providers must be able to determine the optimal number of bus departures and the frequency of departure of buses from the place of origin to the destination [7].

A suitable schedule for bus operation in each time zone of a given day can contribute to creating an appropriate timetable and lower dwell times, and hence can increase passenger comfort. As mentioned by Bai *et al*., one of the reasons for dissatisfaction with a public bus service arises when passengers must wait for a long time at a bus stop and cannot predict the time of arrival at the destination [8].

Oort *et al*. found that the quality of public transport depends on the reliability of the service, and that the transit time of passengers must be guaranteed [9]. They concluded that the reliability of the service is a deciding factor in passenger choice behavior, based on a review of research papers from the previous 10 years.

Kornfeld *et al*. optimized a timetable using nonlinear and Poisson algorithms, with the aim of shortening the passenger waiting time [10]. Their model measured the traveling time of passengers and the arrival time at each bus stop, and defined the measured data as a variable. Based on the results of their study, Bais *et al*. identified time zones in which bus services were less frequent to match passenger demand, and also suggested a controlled number of bus trips within a given time zone [11]. With the aim of reducing the operational costs, they converted the number of buses (and hence the number of seats) into an objective function [11]. They set a threshold and ensured that the number of passengers did not drop below 75% of the total capacity. For example, if the bus service in a given time zone satisfied passenger demand, it was not deemed necessary to modify the timetable; otherwise, it was modified [11]. In their approach, they used a genetic algorithm (GA) and formulated an objective function to maximize the frequency of the service. They set constraints on the fleet size, a loading factor, the number of buses, and the minimum stopping time.

Regarding passenger demand, Wagale *et al*. found that ''the reliability of service is important to ensure the quality of public bus service'' [2]. In the same vein, other researchers have focused on either cost reduction from the viewpoint of the operator or reductions in passenger dwell time [10]. The parameters used in these studies were diverse, and even if an optimal solution could be found in practice, the bus operator would have difficulty adopting this solution. An optimal public transportation service is one that maintains a balance between the project cost and the number of trips based on passenger demand. We can therefore see that most prior research on timetable optimization falls short of providing easy-to-adopt and efficient improvements for both operators and passengers.

In the optimization method presented in this paper, we focus on the reliability of the bus transport service. Reliability forms a measure of service quality, in addition to the traffic and route characteristics, and is related to the passenger and operational characteristics of the public transport system. From the passenger's point of view, the reliability of the service is strongly related to the headway, the travel time, and the wait time for passengers. From the bus operator's perspective, reliability is closely related to operational aspects such as the size of the fleet and the cost of allocating vehicles throughout the day. Hence, the reliability of a public transport service is much more sensitive to the schedule reliability than to the service frequency [12].

In this paper, we use the operational cost (in terms of the operating size of the fleet per day) and the dwell time at bus stops as key parameters in order to optimize both the cost and the dwell time, and thus to provide better reliability.

In an earlier literature review [13], we studied different solution models for the optimization of bus schedules. Of the many different methods that have been used for the optimization of public transportation, we focus here on the use of K-means and a GA. The K-means clustering algorithm is an unsupervised algorithm that is one of the best-known, simplest and most commonly used clustering methods [14]. It enables the grouping of data points that are close to centroids, and is therefore widely used in the analysis of traffic clusters [15], [16]. The GA is a global optimization method that has proven useful in many optimization problems with multiple parameters.

The GA is a heuristic search method inspired by Charles Darwin's theory of natural evolution and represents the process of natural selection, where the fittest individuals are selected for reproduction and produce offspring for the next generation [17]. GAs are also widely used in transportation and areas related to scheduling optimization [7].

II. RELIABLE BUS SERVICES

The objective of this work is to find the optimal solution to the operator's demand for lower operational costs while simultaneously meeting passenger demand for reliable bus services with lower waiting times. Fig. 1 illustrates the approach used to address this problem, in which we formulate the objective, determine parameters on which to focus and link these to the wider objective and the demands of the stakeholders, namely the service provider and the passengers [6], [9], [18]–[20].

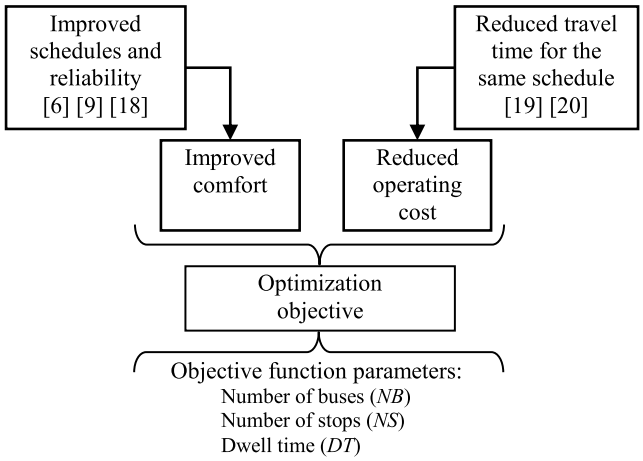


FIGURE 1. The objectives of improving passenger comfort and reducing operational costs.

In this research, we use actual data as the basis for bus timetable planning. The main drivers motivating this study are as follows:

1) Achieving a high level of reliability for passengers at a reasonable cost to bus operators continues to pose a major challenge for bus operation systems, despite the availability of several optimization algorithms.
2) There is a high demand and a real need for practical and adaptable solutions that can be developed easily and made available for operators to adopt, and which can be effective in bringing benefits to many stakeholders on a daily basis.
3) Vital data related to real bus operations can be easily obtained and collected, and can be used to develop solutions for improved services.

We therefore seek to improve a daily bus schedule, and we use the volume of passengers and the waiting time as core guiding parameters for optimization of this schedule. Our optimization is intended to provide both lower wait times (and hence greater comfort to passengers) and a cost-controlled public bus service. The current routes and operational data of the bus service are investigated and used to assess and guide the optimization process.

In the most general case, the objective of this optimization problem is to minimize the total travel time. At the same time, there is a constraint on the maximum number of buses; this is the main component that limits any increase in the frequency of buses, which is desirable from the passengers' viewpoint since it reduces the waiting time and is therefore essential for passenger comfort and satisfaction.

One challenging aspect is that the two issues of passenger demand for improved services and a reduction in the operating costs of public buses are not directly proportional issues. Finding a trade-off between a reliable service and improved waiting times on the one hand and improved revenue from the same number of passengers on the other requires a theoretical framework.

III. PROPOSED OPTIMIZATION MODEL

To achieve the objectives of this study, we use a three-phase optimization model. We establish a framework for our model that includes data preparation, clustering and optimization using a GA, as shown in Fig. 2. The dataset used in our experiments includes GPS tracking data, the current timetable, and the locations of authorized bus stops along two bus routes.

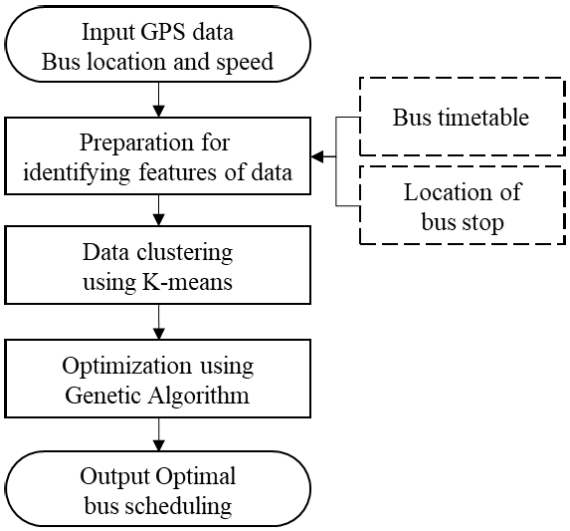


FIGURE 2. Proposed framework for the optimization of a public bus schedule.

Current operational data for this public bus service form an essential component of our optimization framework. We aim to enhance the current operation plan by exploring favorable patterns and identifying other points of interest from current operational data. We therefore need to identify the optimum patterns and elements of the existing bus service, and then apply these to create a heuristic optimization process for scheduling. We employ a clustering algorithm for this process. In clustering analysis, an unrelated set of data points are first divided into meaningful groups (clusters) of data [21]. We use K-means, a practical and efficient data mining clustering technique, to discover these operational patterns.

We also apply a GA, a stochastic optimization technique, to make use of the patterns obtained from the clustering process as input to guide in finding a better allocation of the current resources (number of vehicles) from the viewpoint of the bus service provider and to improve the service for customers in terms of the dwell time [7].

Our approach to optimization in this study is data-driven, with a stochastic search using a GA. The GA uses an objective fitness function with two parameters, and favorable values for these parameters are found for the given operation dataset using a clustering technique.

A. DATA PREPARATION AND PARAMETERS

The primary dataset contained the global positioning system (GPS) location coordinates of the vehicles and their time stamps. These data elements were extended by calculating the

number of stops (*NS*) made by the bus during each trip. The duration of each stop was derived and added to the dataset, and was labeled as the *dwell time* (*DT*). Table 1 shows a sample record from the GPS dataset and the dwell time. The GPS data were set to an interval of a few seconds to allow for automatic vehicle location (AVL), meaning that the distance traveled could be measured using the law of cosines [22]. In Table 1, records 3 and 4 show a bus that has stopped near an authorized bus stop (i.e., in the "nearby zone"). The bus spent 38 s and 13 s at each location.

TABLE 1. Samples of bus GPS data with dwell times.

No.	Time	DT	Latitude	Longitude	Nearby zone
1	17:39:56	0:00:09	26.161486	127.666322	
2	17:40:06	0:00:10	26.161703	127.667338	NO
3	**17:40:44**	**0:00:38**	**26.161853**	**127.667955**	**YES**
4	**17:40:57**	**0:00:13**	**26.161853**	**127.668005**	**YES**
5	17:43:17	0:01:07	26.161886	127.668638	NO
6	17:43:27	0:00:10	26.161870	127.668655	NO

The data preparation step illustrated in Fig. 2 uses GPS log data for the buses, in addition to the time zones and authorized bus stop locations provided by the bus operator. When these GPS time and location data are combined, we can determine whether a bus stopped at an authorized bus stop. The implementation involves a two-step process: calculating the dwell time when the bus stops, and then identifying stops at authorized bus stops. Any stops due to traffic jams or adverse traffic conditions are excluded by referring to the information on the bus stop locations on the specific route provided by the bus operator. The number of times the bus stops and the dwell times at these bus stops can be identified for each bus trip, and are accumulated for each operation time zone throughout the day. Fig. 3 shows examples of the total number of stops (*NS*) and the dwell time (*DT*) displayed per time zone (t).

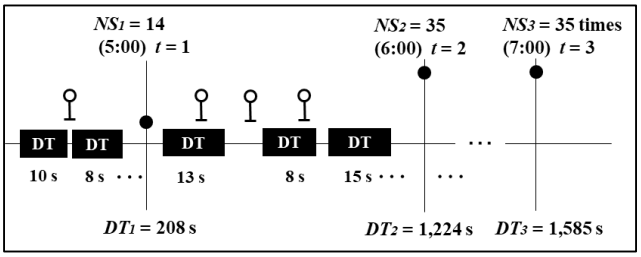


FIGURE 3. Dwell time at bus stops per time zone.

The total number of stops and the total dwell time per time zone are used as input features in the clustering process. These parameters are calculated in the data preparation step that is carried out prior to the clustering step below. We use the dwell time (*DT*) and the number of stops (*NS*) as two indicators of passenger demand for the service and its quality.

B. K-MEANS CLUSTERING FOR BUS OPERATION DATA

NS and *DT* are used as input features. The values of these parameters were calculated in the data preparation step, which was carried out prior to the clustering step. The groups resulting from this clustering process can be used to identify different densities of bus operations in terms of *NS* and *DT*, making it possible to identify and estimate peak and off-peak periods.

A high value of *NS* with a high *DT* represents a situation with many passengers and a high demand for bus services. Under these conditions, passenger comfort is compromised, although the revenue will be higher. Conversely, for low *NS* with low *DT*, there are fewer passengers and low demand for bus services. Under these conditions, the operational cost to the bus operator is compromised. Theoretically, a combination of moderate values of the dwell time (*DT*) at bus stops and frequent stops (*NS*) should result in an acceptable trade-off between passenger comfort and operational cost. Fig. 4 shows bus service revenue levels and a comparison between *DT*, *NS*, passenger demand and service provider demand.

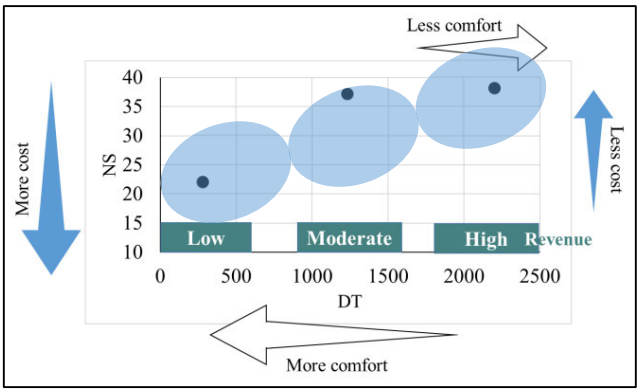


FIGURE 4. Comparison of bus operator demand for higher revenue (lower cost as result of higher NS) versus passenger demand for higher comfort (lower DT). Revenue levels are shown per dwell time.

Three zones can be identified, and these are shown as *low*, *moderate*, and *high* on the graph. A moderate *DT* provides a trade-off between good quality and comfort for passengers while still giving acceptable operational conditions for the bus provider in terms of cost. With an empirically reasonable pick-up of three clusters for the clustering algorithm, the clusters correspond to the low, moderate, and high operation patterns, for the input features of NS_t and DT_t that were previously mentioned in Section A.

In Fig. 4, the *high* group of data (i.e. high *DT* and *NS*) are associated with higher revenue to the bus operator, since higher values of *DT* and *NS* will result in larger volumes of passengers. *Low* data refer to low values of both *DT* and *NS*. The other points in between are classified as *moderate*; this group represents a compromise between passenger demand for lower *DT* and the provision of a bus service with reasonable revenues and operational costs. Fig. 5 shows the parameters used for K-means clustering.

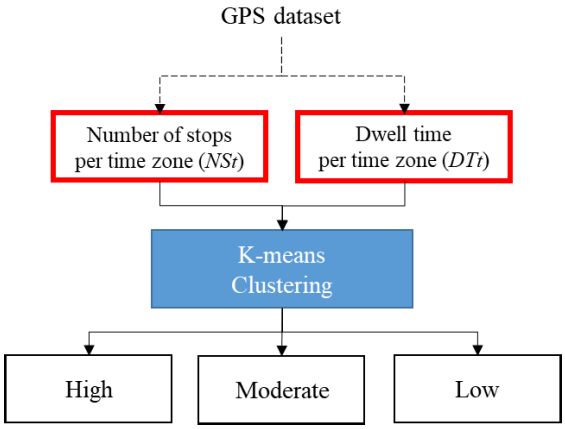


FIGURE 5. Levels of operation with NS_t and DT_t as inputs, clustered into three groups; high, moderate and low.

C. GENETIC ALGORITHM FOR OPTIMIZATION OF BUS SCHEDULING

In our optimization model, the process of applying the GA to generate an optimal timetable involves three steps:

1) Individual encoding: The population of GA individuals is established as follows:
Individual $= \{b_1, b_2, b_3 \ldots, b_n\}$, where $n =$ the number of time zones per day, $b_t =$ the number of bus vehicles allocated per time zone t, and $t \in \{1, 2, \ldots, n\}$.
2) Objective function: A function $f(DT_t)$ is used to assess the individual fitness value, where DT_t is the total dwell time in time zone t. The fitness value is the absolute value of the difference between the actual DT per time zone (DT_t) and the moderate DT per time zone (mDT_t), calculated over the daily time zones.
3) GA operators: Standard GA operations are applied to the populations of individuals, each of which encodes the parameters of a solution for the problem at hand.

Basic pseudocode for the genetic algorithm is as follows:

1) SELECTION: Stochastic, uniform.
2) REPRODUCTION:
   - Elite count: default (5% of population size).
   - Crossover fraction: default (0.8).
3) CROSSOVER: Intermediate.
4) MUTATION: Adaptive, feasible.

IV. EXPERIMENT AND ANALYSIS

A. DESCRIPTION OF THE DATASET AND EXPERIMENTAL SETUP

The datasets used in our experiments consisted of operational data obtained from a bus service provider in Okinawa Prefecture, Japan. The details of the dataset are given below. Three types of data were used:

- GPS location data for each bus along each route
- Authorized bus stops and their locations
- Daily time zones
- Number of vehicles operating per time zone.

The GPS information contained data on two bus lines, A and B, over 92 days, both inbound and outbound, and on weekdays and weekends. The numbers of data logs were as follows:

- Line A: 1,435,000 logs (15,600 logs per day)
- Line B: 1,985,000 logs (21,600 logs per day).

MS Excel, Anaconda (Python) and MATLAB were used to carry out the experiments. Table 2 shows the experimental setup, environment and tools.

TABLE 2. Summary of experimental setup.

Process	Algorithm	Environment/tools
Data preparation	Manual	MS Excel
Clustering	K-means	Python
Optimization	GA	MATLAB

B. DATA PREPARATION AND PARAMETER SETUP

The dataset used in the experiment was composed of bus GPS log data, bus timetables for each hourly time zone of each day, and the locations (latitude/longitude) of authorized bus stops. The GPS dataset was collected over 92 days for two bus lines, Line A and Line B. A breakdown of these datasets is shown in Table 3. The table shows the bus GPS logs, the total number of bus stops (*NS*), dwell time (*DT*), and total number of bus vehicles (*NB*) for the 92-day period.

TABLE 3. Bus GPS log dataset, number of buses (*NB*), *NS*, and *DT*.

Bus	GPS logs	Total			Days
		NS^{*1}	DT^{*2}	NB^{*3}	
Line A	1,435,000	1,955	53,626	5,040	92
Line B	1,985,000	1,711	59,013	9,768	92

*1 *NS* was calculated from GPS logs.
*2 *DT* was calculated using *NS*.
*3 *NB* was obtained from the timetable provided by the bus operator.

Based on this dataset, three parameters were identified: the departure *time zone* (*TZ*), the dwell time (DT_t) in each time zone, and the number of stops (NS_t) in each time zone. Table 4 shows the values of DT_t and NS_t for Bus Line B.

It can be seen from Table 4 that the bus that departed in time zone 7:00 stopped 38 times at authorized bus stops, with a total *DT* of 2,203 seconds, the longest of all the time zones. In time zone 16:00, the bus stopped 39 times, with a total *DT* of 2,043 seconds. These two time zones, 7:00 and 16:00, therefore represent a reasonable suggestions for the peak morning and afternoon hours of passenger commuting.

We were then able to make an estimate of the peak and off-peak hours based on the values of *NS* and *DT*. Of the total logs in the GPS dataset, only those related to the *NS* at authorized bus stops were used to calculate *DT*, which in turn was used as an indicator of the volume of passengers.

TABLE 4. NS_t and DT_t for different departure time zones (*DTZ*).

Departure time zone (*DTZ*)	Bus Line B	
	NS_t	ΣDT_t (s)
5:00	22	278
6:00	37	1,233
7:00	**38**	**2,203**
8:00	31	833
9:00	28	629
10:00	28	762
11:00	26	749
12:00	22	463
13:00	29	888
14:00	30	1,398
15:00	24	910
16:00	**39**	**2,043**
17:00	29	1,309
18:00	33	1,923
19:00	30	786
20:00	21	818
21:00	9	133

TABLE 5. Average weekday values for *NB*, *NS* and *DT* (bus line B).

Departure time zone	NB_t	NS_t	DT_t (seconds)
5:00	1	14	208
6:00	4	35	1,224
7:00	4	35	1,585
8:00	3	31	1,015
9:00	3	28	1,040
10:00	4	28	1,054
11:00	4	28	973
12:00	3	25	723
13:00	4	29	1,135
14:00	4	29	1,286
15:00	4	31	909
16:00	5	34	1,685
17:00	3	30	1,100
18:00	3	28	919
19:00	3	25	658
20:00	4	20	461
21:00	1	8	119

C. CLUSTERING OF BUS OPERATION DATA

The input to the clustering step was the classified dataset obtained as a result of the previous preparation phase. Clustering was applied to a minimized dataset containing the time zone *TZ*, the number of stops (NS_t) and the total dwell time (DT_t) per time zone (TZ_t). The average values of the *DT* in each time zone (DT_t) per day during the months of July (31 days), August (31 days) and September (30 days) and the number of stops per time zone (NS_t) were used in the K-means clustering process. Table 5 shows values for *NB*, the average values for *NS* and *DT* per daily time zone, as well as the daily average over the whole period (92 days).

We used silhouette analysis to select the appropriate number of clusters for our K-means clustering operation [24]. To determine the optimal value, we applied the K-means algorithm with two, three, four and five clusters to bus operation data containing the numbers of stops and dwell time per time zone (NS_t and DT_t) for inbound weekdays and weekends, and outbound weekdays and weekends, for Bus Line B. The outcome of this is shown in Table 6, and it can be see that a cluster value of three was on average a good choice for the test data.

The graph in Fig 7 shows the clustering results for the data on Line B (outbound) (Table 6).

Fig. 7 shows a cluster of data with moderate dwell time, as generated by K-means using the data in Table 5. The cluster with moderate values (mDT_t) is indicated with a circle. Smaller values of *DT* are seen for time zones 5:00 and 21:00,

TABLE 6. Silhouette analysis scores for 2-5 cluster * (line Am NS_t & DT_t as features, 17 sets).

Number of clusters	Outbound		Inbound	
	Week day	Week end	Week day	Week end
2	**0.638***	0.56	0.416	0.504
3	0.536	**0.633***	**0.565***	0.479
4	0.555	0.563	0.491	0.525
5	0.505	0.522	0.451	**0.535**

*Initialized with K-means++, 10 reruns limited to 300 steps

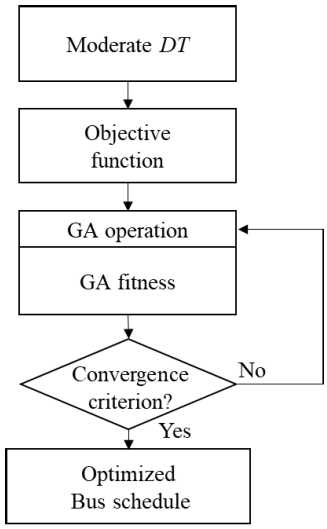


FIGURE 6. Workflow of the proposed GA schedule optimization scheme.

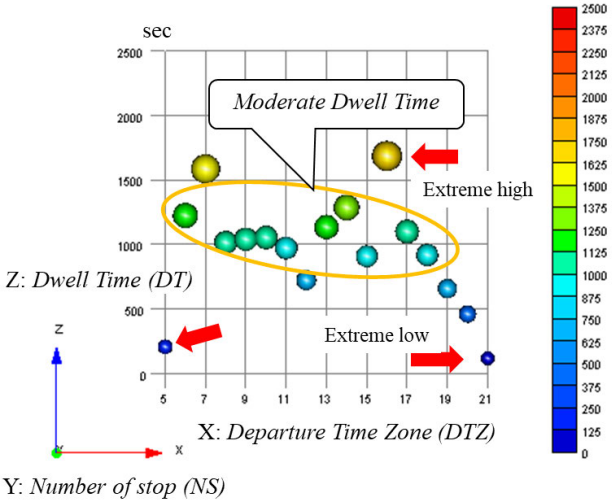


FIGURE 7. K-means clusters, with moderate values of *DT* circled (line B outbound, weekdays, based on the data in Table 5).

TABLE 7. Moderate and average moderate dwell time (*Ave mDT*) (bus line B).

Departure time zone (*mTZ=10*)	NS_t	DT_t (s)
6:00	35	1,224
8:00	31	1,015
9:00	28	1,040
10:00	28	1,054
11:00	28	973
13:00	29	1,135
14:00	29	1,286
15:00	31	909
17:00	30	1,100
18:00	28	919
Ave *mDT*	**-**	**1,066**

meaning that fewer passengers used the bus service in these time zones. The 16:00 time zone has the longest *DT*, and the number of passengers per bus is at its peak. These extreme values for dwell time (*DT*) are indicated with arrows in the figure.

The average value of moderate dwell time (*Ave mDT*) was then used as a parameter for the fitness function in the subsequent process of optimizing the GA algorithm. *Ave mDT* was calculated using (1):

$$Ave\ mDT = \sum_{t=1}^{t=mTZ} DT_t / mTZ \tag{1}$$

where *Ave mDT* is the average value of the moderate dwell time, and *mTZ* is the number of time zones with moderate *DT* (i.e., excluding time zones with high and low values).

TABLE 8. Timetable for bus line B on weekdays, before optimization.

Departure time	Number of buses (NB_t)	
	Outbound	Inbound
5:00	1	0
6:00	4	1
7:00	4	6
8:00	3	3
9:00	3	3
10:00	4	3
11:00	4	4
12:00	3	4
13:00	4	4
14:00	4	4
15:00	4	3
16:00	5	4
17:00	3	5
18:00	3	3
19:00	3	3
20:00	4	3
21:00	1	4

TABLE 9. Vehicle allocation and operation rate (*Or*) for bus line B on weekdays.

Departure time	Outbound		Inbound	
	NB_t	Or_t	NB_t	Or_t
5:00	1	19%	--	--
6:00	4	29%	1	23%
7:00	4	37%	6	28%
8:00	3	32%	3	22%
9:00	3	33%	3	19%
10:00	4	25%	3	16%
11:00	4	23%	4	21%
12:00	3	23%	4	24%
13:00	4	27%	4	24%
14:00	4	30%	4	30%
15:00	4	21%	3	30%
16:00	5	32%	4	38%
17:00	3	34%	5	40%
18:00	3	29%	3	37%
19:00	3	21%	3	34%
20:00	4	11%	3	24%
21:00	1	11%	4	20%

Table 7 shows the values of NS_t, DT_t, and the *Ave mDT* for Bus Line B.

D. OPTIMIZATION WITH THE GA

As stated earlier, our optimization method seeks a balance between demand and supply for the bus service.

TABLE 10. Current and GA operation plan with *Ave mDT*, *mDT* and *f* (*Schedule*) (bus line B).

Current operation *DT*		Current operation plan				GA operation plan			
Time zone	*DT*	*NB*[c]	*Ave mDT*	*mDT* (*NB* × *Ave mDT*)	$f_t(Schedule)$ |(*DT*– *mDT*)|	*NB*[*]	*Ave mDT*	*mDT* (*NB* × *Ave mDT*)	$f_t(Schedule)$ |(*DT*– *mDT*)|
5:00	208	1	1,066	1,066	858	1	1,066	1,066	858
6:00	1,224	4	1,066	4,264	3,040	2	1,066	2,132	908
7:00	1,585	4	1,066	4,264	2,679	2	1,066	2,132	547
8:00	1,015	3	1,066	3,198	2,183	1	1,066	1,066	51
9:00	1,040	3	1,066	3,198	2,158	1	1,066	1,066	26
10:00	1,054	4	1,066	4,264	3,210	1	1,066	1,066	12
11:00	973	4	1,066	4,264	3,291	1	1,066	1,066	93
12:00	723	3	1,066	3,198	2,475	1	1,066	1,066	343
13:00	1,135	4	1,066	4,264	3,129	2	1,066	2,132	997
14:00	1,286	4	1,066	4,264	2,978	2	1,066	2,132	846
15:00	909	4	1,066	4,264	3,355	1	1,066	1,066	157
16:00	1,685	5	1,066	5,330	3,645	2	1,066	2,132	447
17:00	1,100	3	1,066	3,198	2,098	1	1,066	1,066	34
18:00	919	3	1,066	3,198	2,279	1	1,066	1,066	147
19:00	658	3	1,066	3,198	2,540	1	1,066	1,066	408
20:00	461	4	1,066	4,264	3,803	1	1,066	1,066	605
21:00	119	1	1,066	1,066	947	1	1,066	1,066	947
-		**30**	$f(Schedule)$		**44,668**	**22**	$f(Schedule)$		**7,426**

An appropriate trade-off between the operational cost and a reliable service with a moderate waiting time for passengers is therefore required.

The World Bank Group and PPIAF have established an acceptable average peak hour occupancy rate in the region of 85% to 95%. In their benchmark indicators, they state that if the average occupancy is 100%, this may mean that the service is just adequate; however, it is more likely to mean that there is unsatisfied demand, with passengers having to wait excessive periods before being able to board a bus [23].

Based on the dwell time per time zone (DT_t) and the number of buses per time zone (NB_t), we define the bus operation rate (Or). This is calculated for each time zone using (2):

$$Or_t = \frac{(DT_t/NB_t)}{Ave\ mDT} \times 100\,(\%) \tag{2}$$

where Or_t is the operation rate per time zone t.

As an example, we use actual data from Table 5 for time zone 1, with values of 208 s for *DT*, 1 for *NB*, and 1,066 for *Ave mDT*. The operation rate (Or_1) for time zone 1 is then calculated as follows:

$$Or_1 = (208/1)/1,066 \times 100(\%) = 19\%$$

For the daily operation rate Or of the bus schedule, the average operation rate per time zone (Or_t) is used.

Table 8 shows the actual numbers of bus vehicles allocated by the operating company during the period July–September for Bus Line B.

Table 9 shows the operation rate for Bus Line B, calculated using (2). The values for *DT, Ave mDT,* and *NB* used here appear in Tables 5, 7, and 8, respectively.

As shown in Table 9, the operation rates are far lower than the 85% level of the benchmark; this indicates that the service may be acceptable to passengers, but may be unsatisfactory in terms of operational cost.

In the rest of this section, we present the details of the implementation of the GA optimization process. As mentioned earlier, the average moderate dwell time per time zone ($Ave\ mDT_t$) and the values for the number of stops per time zone (NS_t) (i.e., the outcomes of the K-means clustering process) are used to calculate fitness values for each bus schedule solution in the GA population. The GA operators are then applied and iterated.

A GA optimization algorithm is then applied to improve the current operation rate (Or). Table 9 shows the operation rates (Or) for all time zones using the method described in (2).

Each individual in the GA (artificial genetic chromosome) represents the number of bus vehicles allocated to each time zone. For example, time zones (5:00–21:00) are counted as 1 to 17. Lower and upper limits of one and five are imposed as constraints on the minimum and maximum numbers of bus vehicles that can be allocated to each time zone slot.

$$\text{Lower} : \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1\},$$
$$1 \times 17 = 17 \text{ (Minimum number of vehicles)}$$
$$\text{Upper} : \{5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5\},$$
$$5 \times 17 = 85 \text{ (Maximum number of vehicles).}$$

The upper limit of five vehicles was based on the maximum number of bus vehicles operating in each time zone in the current schedule determined by the bus operator. In other words, a minimum of 17 and a maximum of 85 vehicles can operate on each line per day. The range of operating costs in terms of the number of vehicles per day, based on 200 cost units per vehicle, can be calculated as follows:

$$17 \text{ vehicles } \times 200 \leq \text{ operating cost } \leq 85 \text{ vehicles } \times 200$$

The GA starts with an initial population of randomly generated individuals within the constraints set out above.

- A GA individual (schedule):
  [1, 2, 2, 1, 1, 1, 1, 1, 2, 2, 1, 2, 1, 1, 1, 1, 1]
  A total of 22 vehicles per day is used in this schedule. Based on an assumption of 200 cost units per day for one vehicle, then the daily operating cost for the bus fleet would be 4,400 cost units.

The moderate dwell time per time zone (mDT_t) is calculated as shown in (3), using the number of vehicles and the average dwell time per time zone: NB_t and *Ave mDT*. Since the number of vehicles *NB* is directly connected to the operational cost, *mDTt* reflects the operational cost.

$$mDT_t = NB_t \times Ave\ mDT \tag{3}$$

where mDT_t is the moderate dwell time per time t.

The value of mDT_t was then used to evaluate the allocation of vehicles per time zone for the schedule generated using the GA. The fitness value for the GA schedule individual, labeled $f(Schedule)$, was calculated in terms of the dwell time (DT_t) and moderate dwell time (mDT_t) over all times zone in the day, as shown in (4):

$$f\ (Schedule) = \sum_{t=1}^{t=TZ} DT_t - mDT_t \tag{4}$$

where $f(Schedule)$ is the fitness value for a GA individual. Here, each individual represents a candidate schedule.

Table 10 compares the values of *Ave mDT*, *mDT* and $f(Schedule)$ for the schedule generated by the GA to those of the current operating schedule. In this table, the value of *Ave mDT* is taken as 1066, which was calculated using (1). The value of *mDT* is calculated for each time zone using (3). The GA operation plan shown in the table is the optimal in the GA population (the one with the lowest fitness value).

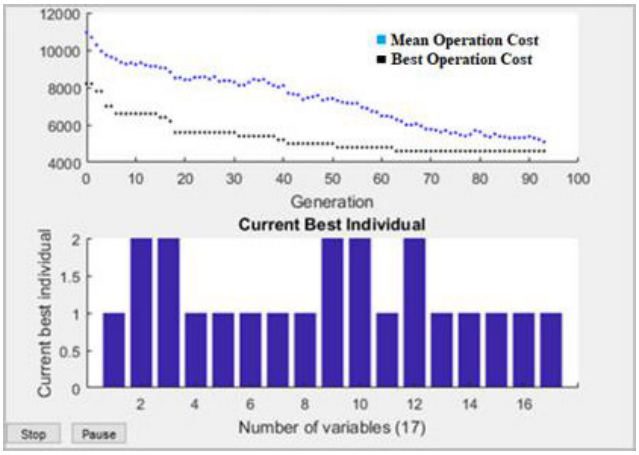


FIGURE 8. Operational cost (number of bus trips per day) and flow optimized with the GA for Bus Line B using MATLAB (outbound, weekdays).

TABLE 11. GA schedule for bus line B.

Time zone	Time zones # (1-17)	NB (Current)	NB (Proposed by GA)
5:00	1	1	1
6:00	2	**4**	**2**
7:00	3	4	2
8:00	4	3	1
9:00	5	3	1
10:00	6	4	1
11:00	7	4	1
12:00	8	3	1
13:00	9	4	2
14:00	10	4	2
15:00	11	4	1
16:00	12	5	2
17:00	13	3	1
18:00	14	3	1
19:00	15	3	1
20:00	16	4	1
21:00	17	1	1
Total vehicles	**-**	**57**	**22**

The numbers of vehicles per day (*NB*) are 30 in the current operating plan and 22 in the GA plan, while the fitness value is 44,668 for the current operation plan and 7,426 for the GA plan. The fitness formula (f) is a composite of the dwell time and the number of vehicles per day, and its value reflects the distance of the solution from the moderate dwell time. The *mDT*, as mentioned earlier, is considered to be a favorable operation pattern that gives a suitable balance between the interests of the passenger and the bus operator.

MATLAB was used to implement the GA algorithm in the optimization process. Fig. 8 shows the best individual

TABLE 12. Operation rates for current and GA–optimized schedules.

Parameters	Outbound								Inbound							
	weekday				weekend				weekday				weekend			
	Current		Proposed by GA		Current		Proposed by GA		Current		Proposed by GA		Current		Proposed by GA	
	NB	*Or*	*NB*	*Or*	*NB*	*Or*	*NB*	*Or*	*NB*	*Or*	*NB*	*Or*	*NB*	*Or*	*NB*	*Or*
Line A	30	59%	26	71%	22	61%	22	62%	30	67%	30	67%	22	82%	23	78%
Line B	57	25%	22	69%	45	44%	27	73%	57	27%	22	66%	45	31%	22	61%

elements and the corresponding cost over 94 generations. The 17 output variables correspond to the time zones (1–17). The GA was set to terminate when the current best individual was unchanged over 30 generations. The population size was 100, and a stochastic uniform selection method was applied with 5% elite preservation. A crossover operation with a fraction of 0.8 and an adaptive feasible mutation were used.

E. RESULTS AND DISCUSSION

For Bus Line B, the results of GA optimization are shown in the proposed timetable in Table 11. The numbers of bus vehicles allocated to each of the 17 time zones are shown. For example, for the 6 a.m. time zone, the GA calculated that the optimum number of vehicles was two, meaning that the number of vehicles was reduced by half in this time zone.

The operation rates (*Or*) for the current schedule plan and the plan calculated using the GA are shown in Table 12, for the best individual vehicle allocations per time zone for both outbound and inbound trips for Lines A and B. The values for *NB* and *DT* for inbound and outbound trips are also shown. For Line A, the outbound weekday rate was improved from the current *Or* of 56% to 67%, while for inbound weekend trips, buses were added to reduce the operation rate (*Or*). For Line B, the outbound weekday rates were improved from 25% to 69% by the GA.

Tables 13 and 14 show summaries of the scheduling results from the GA compared to those of the current timetable, in terms of the reduction in number of buses operating per day (*NB*) and the operation rate (*Or*). The numbers of buses allocated per day in the original schedule are also shown. For Line A, Table 13 shows that the number of buses in the GA schedule is slightly reduced by 2.2% on average. For Line B, Table 14 shows that the number of buses in the new schedule is significantly reduced by 53.5% on average. The operating rates for Lines A and B are 69.6% and 67.5%, respectively. In the GA schedules, the operation rates for both lines have adequate values.

TABLE 13. Summary of GA-optimized bus schedule for line A.

Parameters	Bus line A			
	Outbound		Inbound	
	Weekday	Weekend	Weekday	Weekend
Current *NB*	30	22	30	22
GA *NB*	26	22	30	23
NB reduction	13%	0%	0%	-5%
Average reduction	2.2%			
GA *Or* (current)	69.6% (67.4%)			

TABLE 14. Summary of GA-optimized bus schedule for line B.

Parameters	Bus line B			
	Outbound		Inbound	
	Weekday	Weekend	Weekday	Weekend
Current *NB*	57	45	57	45
GA *NB*	22	27	22	22
NB reduction	61%	40%	61%	51%
Average reduction	53.5%			
GA *Or* (current)	67.5% (32.0%)			

The results presented above demonstrate that the problem of optimizing a schedule while controlling the cost and maintaining an adequate operating rate (and hence an acceptable dwell time) is effectively addressed by the proposed optimization model. Our model optimizes the process based on the dwell time and number of bus stops in each time zone, and adjusts the allocation of vehicles in each time zone. The number of buses operating throughout the day represents a measure of the operation cost.

V. CONCLUSION AND FUTURE WORK

This research has proposed a framework for optimizing timetables using a public bus operation dataset, K-means clustering, and a GA. The objective was to optimize the allocation of vehicles per time zone throughout the operational hours during the day. An optimization target was based on two elements representing the passenger demand and the needs of the service provider, and the dwell time and number of stops per hourly time zone were used as the defining elements. A moderate operation rate that represented a balance between high and low levels was used as a base to find a better allocation of bus vehicles per time zone, with a moderate dwell time and operation cost, using the optimization framework proposed here. The high operation

level represented a higher operational revenue due to the high volume of customers served per trip, but might involve more dwell time and compromised reliability of service from the passengers' point of view. On the other hand, a low operation level representing a lower dwell time with more comfort for passengers would generate lower revenue for the service provider due to the lower passenger volume. The K-means clustering algorithm was used to group data and to identify a suitable moderate level of operation. This moderate set of data was used as an objective fitness function for the GA, which was used to find the optimal allocation of bus vehicles per time zone within the operational constraints of the upper limit on the number of buses. Based on three months of actual operation data on public buses, the proposed optimization model yielded good results and generated new operation schedules with lower overall dwell times, higher revenue and lower overall costs in terms of the number of buses operated per day. This schedule can improve reliability while ensuring that passengers are comfortable by reducing the dwell time. The proposed optimization model provides a method of maintaining a high operation rate and low cost while avoiding overoperation.

One novel aspect of our optimization model is that it makes use of real operational data for an existing schedule; in addition, it is applied to only a small portion of this operation dataset, and can suggest clear and easy improvements. The model can be useful in improving working plans within the capacity of the current resources. In our experiments, although we used operational data collected over three months, containing approximately 4 million data elements, the dwell time and number of bus stops used in the optimization framework made up only a small fraction of the total dataset. The proposed model can also operate based on scalable ranges such as monthly or yearly datasets.

Our framework can be applied to datasets that are widely available and are regularly collected by modern tracking devices. This means that it is easy for operators to adopt this approach to create consistent improvements in their schedules, more effective allocation of resources and better service to passengers.

The design of our optimization framework is based on the availability of existing bus operational data, which are used to find the optimum patterns of operation. Our approach is based on the assumption that the time zones of bus operation and the traffic conditions are likely to remain the same as during the period of collection of the dataset. Although this may represent a limitation of the proposed approach, it means that it can be applied in most cases, since most public transportation systems are in operation. The improvements created using the optimization model of this study are easy to generate and are likely to be easy to adopt and integrate within the current constraints on operational systems. The tendency of keeping the majority of work in operation is not rare, with a need for improving the quality of the service with a better scheduling of resources that are available.

One limitation of this study is that the verification of the proposed vehicle schedule is based on the original dataset; an empirical verification would require collaboration with a bus operator to put these modifications into practice and validate the proposed scheme. This option was not available to us in the course of this research work.

ACKNOWLEDGMENT

This research utilized data on public buses in Okinawa Prefecture. The authors would like to thank the Ryukyu, Okinawa, Naha and Toyo bus operators who provided experimental data. They are also grateful to the Okinawa Industry Promotion Public Corporation for providing grants and opportunities for this research.

REFERENCES

[1] J. Wu, M. Yang, S. Rasouli, and C. Xu, "Exploring passenger assessments of bus service quality using Bayesian networks," *J. Public Transp.*, vol. 19, no. 3, pp. 36–54, Sep. 2016.

[2] M. Wagale, A. P. Singh, A. K. Sarkar, and S. Arkatkar, "Real-time optimal bus scheduling for a city using a DTR model," *Procedia - Social Behav. Sci.*, vol. 104, pp. 845–854, Dec. 2013.

[3] S. Carosi, A. Frangioni, L. Galli, L. Girardi, and G. Vallese, "A matheuristic for integrated timetabling and vehicle scheduling," *Transp. Res. B, Methodol.*, vol. 127, pp. 99–124, Sep. 2019.

[4] A. Ceder, "Urban transit scheduling: Framework, review and examples," *J. Urban Planning Develop.*, vol. 128, no. 4, pp. 225–244, Dec. 2002.

[5] Y. Wu, H. Yang, J. Tang, and Y. Yu, "Multi-objective re-synchronizing of bus timetable: Model, complexity and solution," *Transp. Res. C, Emerg. Technol.*, vol. 67, pp. 149–168, Jun. 2016.

[6] A. Ceder, *Public Transit Planning and Operation: Modeling, Practice and Behavior*. Boca Raton, FL, USA: CRC Press, 2016.

[7] F. D. Wihartiko, A. Buono, and B. P. Silalahi, "Integer programming model for optimizing bus timetable using genetic algorithm," *IOP Conf. Series, Mater. Sci. Eng.*, vol. 166, no. 1, 2017, Art. no. 012016.

[8] C. Bai, Z.-R. Peng, Q.-C. Lu, and J. Sun, "Dynamic bus travel time prediction models on road with multiple bus routes," *Comput. Intell. Neurosci.*, vol. 2015, pp. 1–9, Oct. 2015.

[9] N. Oort, D. Sparing, T. Brands, and R. M. P. Goverde, "Optimizing public transport planning and operations using automatic vehicle location data: The Dutch example," in *Proc. 3rd Int. Conf. Models Technol.*, Dresden, Germany, 2013, pp. 291–300.

[10] S. Kornfeld, W. Ma, and A. Resnikoff, "Optimizing bus schedules to minimize waiting time," *Oper. Res.*, vol. 7, pp. 321–392, Oct. 2014.

[11] N. S. Bais, N. Pitale, and S. Thorat, "Optimal schedule modeling for public transportation system," *Int. J. Sci. Res.*, vol. 4, no. 4, pp. 2053–2056, 2013.

[12] R. Liu and S. Sinha, "Modelling urban bus service and passenger reliability," in *Proc. Int. Symp. Transp. Netw. Rel.*, The Hague, The Netherlands, Jul. 2007, pp. 1–20.

[13] Y. Shima, R. Abdul, F. Ali, and R. Alayman, "Optimisation of a public bus schedule using a neural network and K-means," *J. Theor. Appl. Inf. Technol.*, vol. 98, no. 19, pp. 3211–3221, 2020.

[14] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, Dec. 2017.

[15] X. Lin, "A road network traffic state identification method based on macroscopic fundamental diagram and spectral clustering and support vector machine," *Math. Problems Eng.*, vol. 2019, pp. 1–10, Apr. 2019.

[16] M. A. Mondal and Z. Rehena, "Identifying traffic congestion pattern using K-means clustering technique," in *Proc. 4th Int. Conf. Internet Things, Smart Innov. Usages (IoT-SIU)*, Ghaziabad, India, Apr. 2019, pp. 1–5.

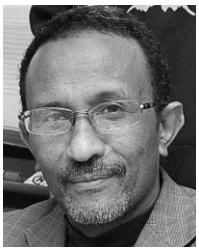
[17] D. Whitley, "A genetic algorithm tutorial," *Stat. Comput.*, vol. 4, no. 2, pp. 65–85, Jun. 1994.

[18] D. Peña, A. Tchernykh, S. Nesmachnow, R. Massobrio, A. Feoktistov, I. Bychkov, G. Radchenko, A. Y. Drozdov, and S. N. Garichev, "Operating cost and quality of service optimization for multi-vehicle-type timetabling for urban bus systems," *J. Parallel Distrib. Comput.*, vol. 133, pp. 272–285, Nov. 2019.

[19] M. Fadaei and O. Cats, "Evaluating the impacts and benefits of public transport design and operational measures," *Transp. Policy*, vol. 48, pp. 105–116, May 2016.
[20] D. Sun, Y. Xu, and Z.-R. Peng, "Timetable optimization for single bus line based on hybrid vehicle size model," *J. Traffic Transp. Eng.*, vol. 2, no. 3, pp. 179–186, Jun. 2015.
[21] T. Galba, Z. Balkić, and G. Martinović, "Public transportation bigdata clustering," *Int. J. Elect. Comput. Eng. Syst.*, vol. 4, no. 1, pp. 21–26, 2013.
[22] M. Hidetoshi, "Three distance calculation methods using latitude and longitude," *Oper. Res., Manage. Sci.*, vol. 60, pp. 701–705, Oct. 2015.
[23] World Bank Group. *Urban Bus Tool Kit. Evaluate Your Bus System: Benchmarks and Indicators/Peak Occupancy Rate*. Accessed: Mar. 24, 2021. [Online]. Available: https://ppiaf.org/sites/ppiaf.org/files/documents/toolkits/UrbanBusToolkit/assets/1/1c/1c13.html

YASUKI SHIMA received the bachelor's degree from the Department of Management and Information Science, Meio University, Japan, in 2002, and the master's degree in management and information science from Meio University, in 2006, with a focus on information design. He is currently pursuing the Ph.D. degree with the Institute of IR4.0, UKM, Malaysia. His research interests include big data analysis and visualization. He is a Core Member of an Research and Development Project pursued by the Institute of IR4.0 and Meio University.

RABIAH ABDUL KADIR is currently a Research Fellow with the Institute of IR4.0, University Kebangsaan Malaysia (UKM). Her research interests include intelligent computing and computational linguistics. She is actively working on big data analytics, semantic knowledge repositories and extraction, intelligent medical systems, and intelligent learning management systems. She is also an Active Researcher and regularly participates in international and regional conferences and academic events.



FATHELALEM ALI received the B.Sc. degree in mechanical engineering from the University of Khartoum, in 1988, and the master's degree in electrical and information engineering and the Ph.D. in information engineering from the University of the Ryukyus, Japan, in 1997 and 2000, respectively. He is currently a Professor of information engineering with the Department of Management and Information Sciences, Meio University, Japan, where he joined as a Faculty Member, in 2000. He works in a multidisciplinary research and education environment and actively involved in collaborative research with several universities and institutes in Japan, Malaysia, the US, France, and Sudan.