Journal of Industrial and Systems Engineering
Vol. 8, No. 4, pp72-86
Autumn 2015

JISE

Proposing an approach to calculate headway intervals to improve bus fleet scheduling using a data mining algorithm

Seyyed-Mahdi Hosseini-Motlagh[1]*, Peyman Ahadpour[1], Abdorrahman Haeri[1]

[1]*School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran*

motlagh@iust.ac.ir, peyman_ahadpour@ind.iust.ac.ir, ahaeri@iust.ac.ir

Abstract

The growth of AVL (Automatic Vehicle Location) systems leads to huge amount of data about different parts of bus fleet (buses, stations, passenger, etc.) which is very useful to improve bus fleet efficiency. In addition, by processing fleet and passengers' historical data it is possible to detect passenger's behavioral patterns in different parts of the day and to use it in order to improve fleet plans. In this research, a new approach is developed to use AVL data to investigate relationship between headway change and passenger downfall rate. For this purpose, a new method is developed that is called Intelligent Headway Selection (IHS) approach. The aim of this approach is finding similar days from passengers' behavior perspective in the dataset and by focusing on unusual patterns of each group, headway changes effects on passenger downfall rate is being studied. In this approach, in the first step, each day is classified into specific time periods (like half of hours) and the passengers' behavior pattern is detected for each day during the specified time periods. Then, in the K-Means algorithm, Euclidian distance measure is replaced with Dynamic Time Warping (DTW) algorithm to enable the K-Means to compare time series. The modified K-Means algorithm is used to compare days in the dataset and categorize similar days in the same clusters. Then, headway – passenger per minute plot is created for each time period to detect unusual patterns. Then, a Headway Interval Detection Procedure (HIDP) is developed to use these unusual patterns to find suitable headway values for each time period. Afterwards, these plots merged and the final headways are calculated.

Keywords: Headway, AVL, Dynamic Time Warping (DTW), Data mining, k-means algorithm, Bus scheduling.

1- Introduction

Public transportation is one of the main alternatives for intercity travels in metropolises. Public transportation efficiency has many benefits for city, people, environment and etc. In other hand there are many challenges to increase efficiency and effectiveness of public transportation. Bus fleets are the most popular transportation alternative especially in Iran. Because of this popularity, bus fleet design, plan and management always involves much complication. Bus fleet design consists of five major steps.

*Corresponding author.
ISSN: 1735-8272, Copyright c 2015 JISE. All rights reserved

1. Transportation network design: in this step the transportation network including bus lines, stations and line routes will be defined.
2. Calculate the efficient frequency for each path (calculate efficient headways): after defining bus lines, it's necessary to calculate headway for each line. Headway is the time interval between two sequent bus departures from the first station
3. Designing the time tables: after defining headway, time table for every line will be created. These tables describe bus travel details for every line.
4. Bus Scheduling: finally, it's necessary to calculate how many buses are required for the designed network. Then, a plan is prepared to determine how each bus should serve in transportation network. This step will use techniques like dead head trips (the inter network trips to move a bus from one line to another one) to serve designed network with minimum number of buses.
5. The Crew Scheduling: this step includes scheduling of drivers and supervisors.

In bus fleet network design, output of each step is an important input for next one, but among all of them, headway is the key step for all other steps. Station headway can be defined as the maximum passengers' waiting time for bus arrival. Headway is not same in different times of day and also different days a week for every line. Therefore, headway acts as a basis to design time tables, line frequency and bus scheduling. In other words, if suitable headway for each bus line couldn't be found, next steps of bus fleet design will lose their accuracy. In other hand headway has a real impact in other network factors like passenger satisfaction, fleet costs, and load factor. Therefore, providing the balance between efficient headway, passengers' satisfaction and fleet costs is one of the most important and complicated issues in bus fleet management problems. Therefore, calculation of appropriate headway is one of the most important aspects in last decade's public transportation research.

The three main contribution of this paper are *a*) proposing a methodology based on AVL historical data, forassessing the effect of change in headway on passenger behavior (Intelligent Headway Selection). *b*) developing a visual-based methodology to determine the proper headway interval for each bus line and then developing related computerized algorithm based on the methodology (Headway Interval Detection Procedure) and *c*) representing a flexible managerial package for headway instead of specific value to enable decision makers to choose the suitable headway values based on their criteria, priorities and managerial requirements.

1-1- AVL Systems

Considerable improvements in information and communication technologies such as GPRS and GPS caused rapid growth in using AVL systems in metropolises. In these systems, every vehicle is equipped with a GPS module which gather vehicle's location information from satellites and GPRS module (or 3G module in some cases) that send gathered data to a central server. The central server has critical functions that process all data to provide valuable reports for fleet managers to make the appropriate decisions. Based on different requirements of the cities, AVL systems developed with different architecture. But all of them has a unique and important function which is online data gathering. This function enables AVL systems to process huge databases of buses, stations, lines, passengers and their effect on each other. AVL central server uses the raw data in three main subsystems:

1. Real time passenger information (RTPI): This subsystem uses AVL online data to inform passengers about fleet network state.
2. Fleet management: This subsystem uses AVL online data to inform fleet managers about buses and drivers, and to help them to monitor network performance.
3. Traffic management: This subsystem supposes that every bus is a traffic node and the data in the traffic center will be updated with the most recent traffic states.

Furthermore, these raw data will be used as an input in further processes. These processes use raw data to improve fleet monitoring, planning and scheduling using time space analysis. In other words, this data will be processed to determine bus fleet efficiency and design new plans to improve fleet performance. In other hand, AVL systems has so much dependency on

data quality (it’s absolutely inefficient with wrong data). Therefore,data accuracy and consistency is critical in AVL systems’ functionality.

1-2- Literature review

For a long time researchers focused on the bus scheduling aspect of the bus fleet planning problems (fourth step in bus fleet design). For example, Lobel (1999) defined the scheduling problem as a linear programming problem and solved it with a cut and branch method. Rahin and Kwan (1999) developed an object-oriented solution based on VAMPIR algorithm that improves the solution by running repetitive rounds. Mesquita and paixao (1999) used a search tree method to find an exact result. They used modified version of cut and branch method to solve the considered problem. Banihashemi and Haghani (2000) defined a scheduling problem in an actual condition in multi trip state. In addition, Banihashemi and Haghani (2002) applied a heuristic algorithm to improve the proposed method. Freling and Wegelmans (2001) used auction method in single path and fixed vehicle state to solve scheduling problem. Huisman et al.(2004) used dynamic programming for bus scheduling. In addition, they removed the fixed time trip assumption from this problem and used a heuristic algorithm to solve it. Yan (2007) solved the scheduling problem in a competitive environment. He formulated a nonlinear integer programming model and implemented it in one of Taiwan cities. Chang-Sheng(2010) considered all of the stockholders requirements and expectations on the model.. He used simulated annealing to solve his model.

On the other hand, some researchers focused on headway planning as one of the most important aspects of bus scheduling. Adebisi (1986) developed a mathematical model to determine the headway variants in fixed bus routes. Oudheusden and Zhub (2000) used a linear programming method with two heuristic algorithms to solve this problem. Yang (2008) used artificial intelligent algorithms to find out the suitable headway value. Matias(2010) used historical data to find number of headways which is needed for eachday.In addition,He used a method to update this value in certain periods of time. Sun et al. (2008) used genetic algorithm to find out the suitable headway in bus rapid transit (BRT). Hairong and Changsha (2009) modeled the time table scheduling as mixed integer nonlinear programming. This research improved the genetic algorithm to solve this model.Yu et al. (2011) used genetic algorithm for regular bus fleets too. Liu et al. (2010) developed an exploratory hazard based analysis to determine the first discharge headway.Li et al. (2013) developed an expected value model for optimizing the multiple bus headways.

With growth in communication technologies, AVL systems have been applied in metropolises and as a result, they gathered huge amount of robust data about bus fleets, passenger behaviors, lines and their influence on each other. Many researches use data mining algorithmson historical data of bus scheduling problem. Some of the researches used data to develop a framework for fleet performance determination. Cheng et al. (2004) used APC (Automatic Passenger Counter) data to predict bus arrival time for each station. Geneidy et al. (2011) developed a method to determine transit performance using AVL and APC data. Mandelzys and Hellinga (2010) proposed a methodology that used AVL-APC data for identifying bus stops that are not meeting performance standards for scheduling. Checn et al. (2013) used AVL real time information to develop a strategy based on dead head trips. That strategy used backup vehicles to prevent headway irregularity. Barbin et al. (2013) analyzed AVL raw data to measure service level of bus regularity at each bus stop. Literature review indicates that the historical data are rarely used for headway value detection in most of the former researches while are mostly used in bus scheduling area. Furthermore, in the most of the researches about headway optimization, the final result is a single value for headway. But there are many factors in public transportation planning that affect the headway value and the priority of these factors constantly changes. Therefore, instead of presenting a single value for headway, proposing a managerial package with enough flexibility to support adaptation with different circumstances is a necessity.

In this research, data mining is used as a suitable approach to analyze the historical data and identify the logical link between different headways' value and passengers' behavior. Therefore, the impact of headway on passengers' behavior could be found out. In addition,

this paper proposes a procedure to provide a flexible managerial package which can help decision makers to choose different values, instead of one fix headway, in different circumstance. This research consists of four major sections. After introduction, in the second section, the proposed approach by this research will be described using CRISP-DM methodology. In third and fourth section, the proposed approach is deployed on a real case study and the results are examined in details. Finally, the managerial applications of the approach and future researches are described.

2- Intelligent Headway Selection Approach

In this research CRISP-DM methodology is used to provide a structured approach to solve the problem. CRISP-DM proposed a process-based approach to perform data mining projects. CRISP-DM is independent from business and technology. Therefore, it can be used in any industry. The main steps of the CRISP-DM as follows.

2-1- Modeling

This stage of CRISP-DM, details of the proposed approach is stated and data mining techniques are specified to analyze dataset and solve business problem.

In this research the main focus is on finding the most suitable headways (and not a single value for it) using historical data that generated in AVL systems. In this paper, suitable headway results in least passenger downfall rate. For this purpose, an idea about similar days was developed. Inthis regard, the focus is on finding similar days (from passenger behavior perspective) and put them in the same groups. Similar days in this research mean similarity in whole day pattern and not necessarily equality in each part of the days. After finding similar days, passenger patterns for each group will be compared. In this case it is expected to detect days with similar patterns in each time period (which is determined in the previous phase). However, in real applications there are always unusual patterns (in headway or passenger behaviors) which will help decision makersto investigate the relationship between headway changes and passenger behaviors. Therefore, modeling phase of the proposed approach consists of three general steps as follows.

2-1-1- Creating daily chains

Initially, passengers' behavior should be observed to find the patterns of each day. For this purpose, it could be assumed that each day is a chain of specific periods in preparation phase. For example,each day could be assumed as a chain of 24 time interval so that each interval isone hour. For each of the periods, the averages of headway and number of passengers are calculated in the preparation phase. Therefore, if passenger's behavior could be determined in each of the periods, it can be considered as a day to create a chain for it.

To achieve this purpose, the approach used K-Means algorithm on time periods and categorize them into different clusters based on passengers' behavior similarity. In other words, the dataset is a collection of time periods with corresponding headway and number of passengers for the period.

K-Means is one of the most popular clustering algorithms in data mining which is used to cluster N observation (in this research record) into K clusters so that each observation is dedicated to the cluster with the nearest mean. It is possible to use different distance measures such as Euclidean distance measure initially, the appropriate number of clusters (k) should be determined. In addition, the start point is very important in K-Means so that different start points can result different clusters.

After clustering the time periods, the center of each cluster can be used to label it. For each cluster center, passenger per minutevalue needs to be calculated. Afterwards, clusters will be ordered based on their value of clusters center to label them. Since each time period has its label, it can be used to build the daily pattern. To achieve that, each time period will be replaced with its label in the day. After that, each day is a chain of labels that shows the passenger pattern for the entire day.

2-1-2- Similar Days Detection

After creating a daily pattern for each day of the dataset, these patterns could be used to compare them with the others. In this phase, the K-Means is used again to cluster these patterns into similar groups. In this case, data records are time series and it is not possible to use Euclidian distance as a distance measure. Dynamic time warping is an algorithm to measure similarity between two time series which vary in x or y. Therefore, DTW is used instead of Euclidean distance measure.

For this purpose it's necessary to choose an Implementation of DTW with suitable time complexity. In general computing DTW requires O (n^2) but there is an implementation which computes it with O (n) called Fast DTW. Therefore in this research instead of DTW, FastDTW is used to modify K-Means.

The modified version of K-Means is capable of comparing different time series and clustering them based on their similarities.

2-1-3- Create headway - passenger / min plot

In this case, similarity is defined as the compliance between similar time periods of the days. Based on the defined similarity, it is aimed to detect specific patterns that present a useful insight about relationship between headway changes and passengers' behavior and their downfall rate.

For this purpose, a plot will be drawn that contains headway value (X-axis) and passenger / min (Y-axis). These plots would create a clear understanding of passenger behavior in different headways. Figure 1 shows steps of the modeling phase in the proposed approach.

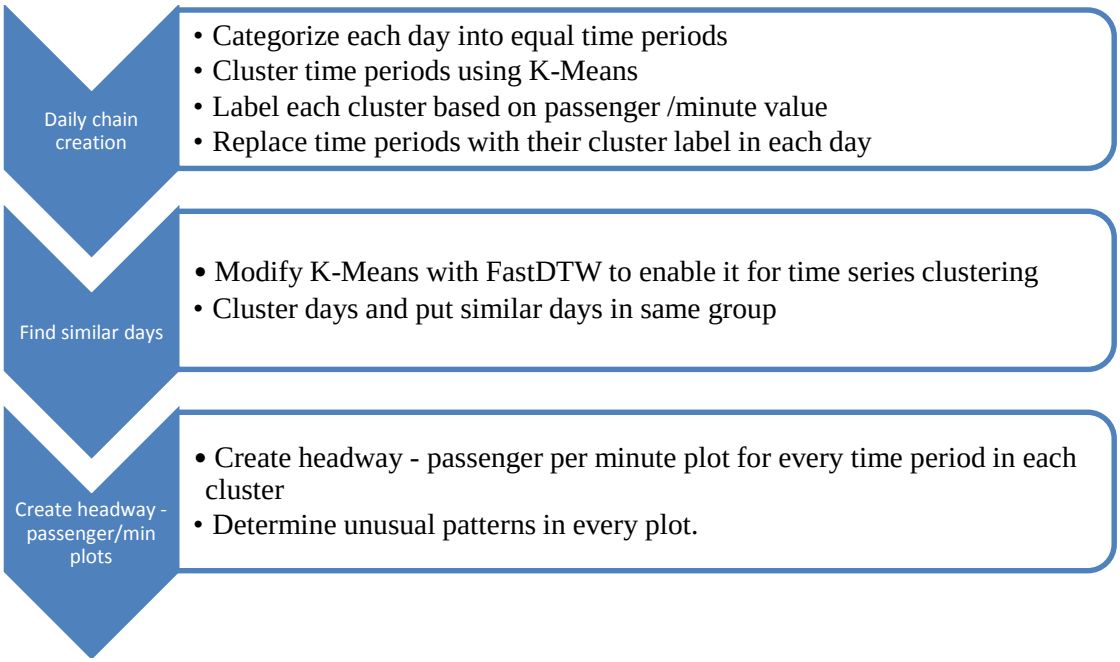

Figure 1. Steps of modeling phase in proposed approach

2-2- Headway Interval Detection Procedure

This stage of CRISP-DM is aimed to analyze the detected patterns. In addition, in this phase it's needed to assess the results to find whether they meet the business requirements that are specified in the first phase (business understanding).

The drawn plots show relationship between headway changes and number of boarded passengers' per minute (especially for time table stations).

The obtained plot pattern for each time period such as *t* has a different shape, but there is a repetitive pattern in all of them. In each plot, there is a time interval for headway that has the maximum value of passengers per minute. This interval can be very small (depending on the data set size) and also if data noises are ignored it would have almost zero slopes. To

determine the mentioned interval for each time period, an algorithm with the following steps is proposed:

1. Finding the point with maximum value of passenger / minute and name it as (D_m).
2. Defining (n) as number of adjacent points plus (D_m). Default value for n is 2.
3. Defining ($t = \infty$) as suitable headways error and ($f = D_m$) as final headways.
4. Creating different combinations with (D_m) and it (n-1) adjacent and calculate headways error for each combination with following formula.

$$e = \frac{Sum\ of\ the\ distance\ of\ adjacent\ points}{n}$$

5. Choose the lowest value of e in different combinations of (n) and name it as (e_n)
6. If value of (e_n)is lower than or equal (t), put (t=e_n), (f = the combination with value e_n), ($n = n+1$) and jump to step 4 again. Else terminate the algorithm and print value of (f).

Figure 2 shows the flow chart of this algorithm.

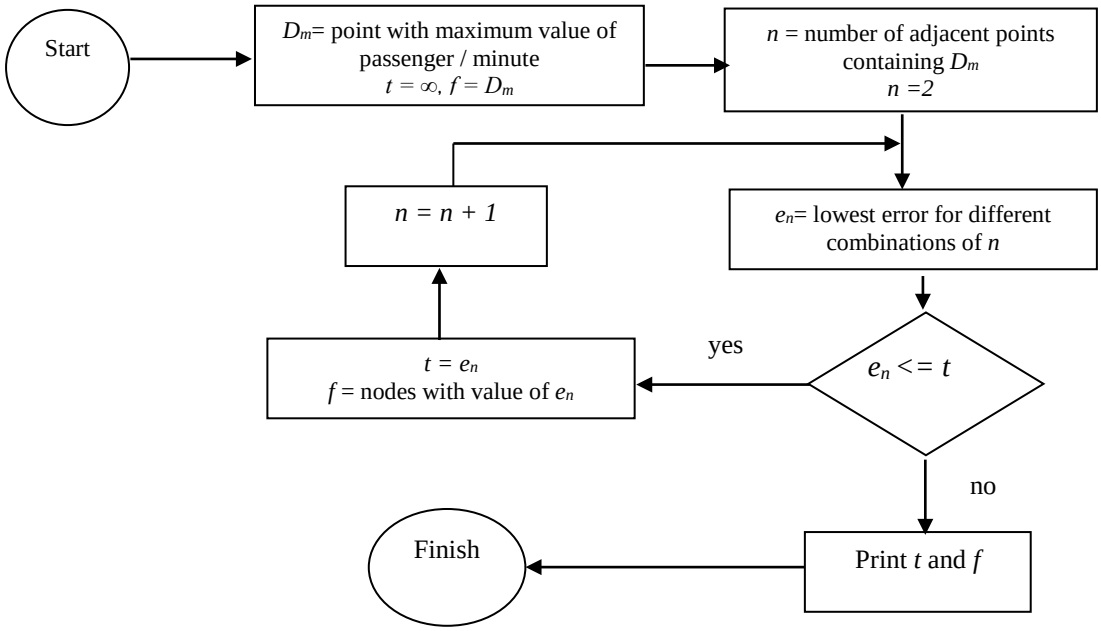

Figure 2. Flowchart of headways select algorithm

The final result of this algorithm shows the suitable headways for that time period in the selected cluster.

3. Case Study

In this section, the proposed approach is applied in an Iranian metropolis named Mashhad.

3-1- business understanding

Mashhad has one of the biggest bus fleets in Iran. Therefore, it requires an accurate plan for the buses that serves for this fleet. In Mashhad's fleet management, headway values are set by experience. Therefore, it is expected to have a lot of problems with these values and also there are a lot of complaints from passengers about long waiting times and extremely crowded buses. The bus fleet managers increase the numbers of buses several times but this decision usually do not have the expected effectiveness.

3-2- Data Understanding

In Mashhad's AVL system, there are about 3000 buses and they send their GPS data every 20 seconds with GPRS. In addition, each bus collects GPS data every 5 seconds and saves it in its internal memory. At the end of the day, when the buses return to the terminals, all their data will be transferred to terminal repositories and send them to the main server through terminal networks. There are two types of data in this AVL system: *a*) Online data which every bus collect and send every 20 seconds and *b*) offline data which every bus collects every 5 seconds and send them all at end of the day. In fact central system uses the online data for daily monitoring and offline data to calibrate and accurate the previously sent data. In addition, AVL uses offline data to create more detailed reports about system performance. Therefore, there is a hugeamount of data in this system that can be used for this study.

Mashhad's database consists of a lot of tables which is used in different parts of the system. But most of them are obtained through post process of the basic table called Track Point. Track point table consist of spatial data which each bus send in online or offline status. These data forms the basic data for Mashhad's AVL system and most of the other data is obtained through processing on these data.Table (1) shows the important fields of Track point table.

3-3- Data Preparation

The first step of the process is to choose the data and make sure that the data is valid and reliable. It is required to check data and any invalid, outliers and in some cases, suspicious data should be removed. In some cases data removal could cause a gap in dataset and cause some problems in the results.

Table 1. The mainfields of Track Points

Field Name	Description
Point	A geographical point of the buss in the time of data sending
Speed	Bus Speed in the time of sending data
Direction	Direction of the bus
Time	The time that GPS module gathered data
Saved Time	The time that this data received in server and saved in database
Passenger On Board	The number of the passengers which boarded till this track point data. The difference between 2 track points in this field that shows the number of passengers which boarded between them
Vehicle	The bus which sent this data
Line	The line which the bus working in (one bus can work in different lines in different time of day)

For example, when invalid track points were deleted, a gap in the bus arrival time chain has been made that create fault in headway calculation. In such cases, it's recommended to stimulate data in these gaps and fix wrong data. In this case, if there were more than three wrong data in any half hour, the entire record was removed from the datasets (our analysis is based on half hours of every day which will be explained in next sections).

Mashhad's AVL data preparation includes three steps. In the first step, a dynamic algorithm is developed to choose a station for headway determination. In the second step, an algorithm is developed to find out the suitable time interval for those stations. The suitable time interval means the interval that chosen station has enough data with the desired quality. Finally, in the third step Mashhad's data is converted to appropriate dataset for the proposed approach.

3-3-1- Choosing appropriate station

To determine the suitable headway values for a bus line, an appropriate station with the much population in the considered line should be selected. If the suitable headway value can be found for the station, these headways are usable for the entire line.

A dynamic algorithm is used to find the most populous stations in the city. The target station is a station which has the most population of passengers among other stations during entire of day. With this definition of populous, the suggested algorithm first will find the most populous zones in the city. In each cycle, the populous zones were narrowed down into some sub-zone and these cycles goes on until candidates 'populous stations were found. The output of the algorithm in Mashhad's datais two stations that match considered criteria and one of them is selectedfor the next step.

3-3-2- Finding appropriate time interval for the selected station

The next step of data preparation aimed toprepare time period for data with least data fault. There are many potential faults in AVL systems. GPS modules could become faulty, GPS satellite can become in accessible for minutes, buses could go to dead zones and GPRS fails to send data and so on. Any of these faults could make data loss in AVL system. In this regard, proper time period needs to be found which means that in that period of time all of the buses in this line send proper data to the server. It's obvious that there is no perfect period but still the period of time with least data loss could be selected. In this phase, each day stations' data will be checked and the days which have so many data losses will be removed. Afterwards, the time period that had the least missing days in it will be selected. The suitable time period was between 17 Aug 2014 and 13 Dec 2014. There were 116 valid days in this period. Table (2) shows the dataset of this step.

Table 2. chosen station sample data

Bus Arrival Time	Bus Departure Time	Passengers On Board
8/17/14 9:19:00	8/17/14 9:19:59	23
8/17/14 9:27:43	8/17/14 9:27:43	5
8/17/14 9:37:45	8/17/14 9:37:45	16
8/17/14 9:47:24	8/17/14 9:47:24	21
8/17/14 9:57:52	8/17/14 9:58:24	29
8/17/14 10:08:22	8/17/14 10:08:22	15
8/17/14 10:18:48	8/17/14 10:18:48	18
8/17/14 10:28:09	8/17/14 10:28:09	4
8/17/14 10:38:07	8/17/14 10:38:07	12
8/17/14 10:48:18	8/17/14 10:49:20	21

3-3-3- Converting Mashhad's data to suitable dataset for proposed approach

The final step of data preparation phase is to cast AVL system data to provide suitable data input for the proposed approach in this research. In this research, each day is categorized into 29 half hours (bus fleet working time is not 24 hours). These half hours are supposed as input data for comprehensive analysis. These half hours will be used to create passenger behavioral patterns for each day. For this purpose, the averages of headway values and number of passengers for each half hour should be calculated. In the table (2), the difference between each row's arrival time and previous row departure time will result in the headway value of each row. After this calculation, the average of headway passengers for each half hour, a day

was found which result in 3624 rows of data. A sample of the final result of data preparation phase is shown in table (3).

Table 3 . An example of average of headway and boarded passengers for each half hour in each day

Half of hour Start	Half of hour End	Boarded Passenger Average	Headway average
10/23/14 6:30:00	10/23/14 7:00:00	15.5	9.196000099
10/23/14 7:00:00	10/23/14 7:30:00	14.66699982	7.382999897
10/23/14 7:30:00	10/23/14 8:00:00	12	13.32800007
10/23/14 8:00:00	10/23/14 8:30:00	13	8.079000473
10/23/14 8:30:00	10/23/14 9:00:00	14.5	12.2329998
10/23/14 9:00:00	10/23/14 9:30:00	20.33300018	9.949999809
10/23/14 9:30:00	10/23/14 10:00:00	18.33300018	10.01099968
10/23/14 10:00:00	10/23/14 10:30:00	15.5	10.30000019
10/23/14 10:30:00	10/23/14 11:00:00	15	14.85000038
10/23/14 11:00:00	10/23/14 11:30:00	9.333000183	9.522000313
10/23/14 11:30:00	10/23/14 12:00:00	10.66699982	10.5170002
10/23/14 12:00:00	10/23/14 12:30:00	5.333000183	9.92800045

3-4- Modeling

3-4-1- Daily chains creation

As mentioned in the previous section, first step of the proposed approach is to find out and cluster days with similar passengers' patterns. For this purpose,finding daily pattern of each day is necessary. Figure (3) shows the scatter plot for headway-passenger in different half hours in different days which was created in the previous step. The X-axis shows headway value and the Y-axis represents the number of passengers.

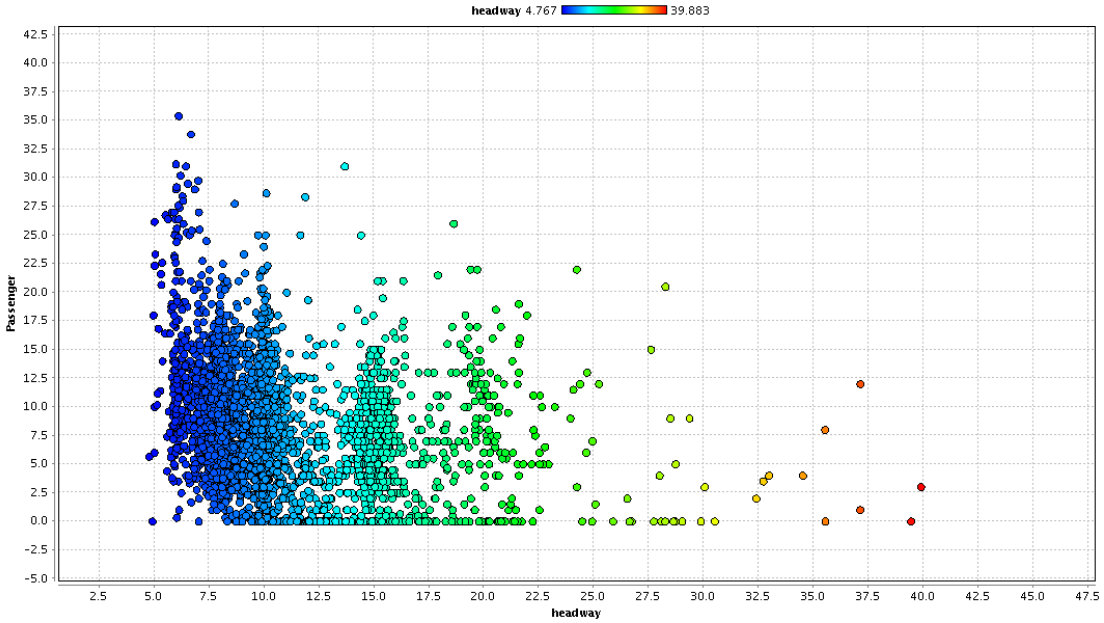

Figure 3 . Scatter plot of headway-passenger for every half hour

This plot indicates that there are similar behaviors in different half hours and different days. In other words, these data could be put in different groups based on passenger arrival rate per minutes for each half hour. K-Means algorithm is used with different K values (K=3, 4,..., 11). For each K, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) have been calculated to choose suitable number of clusters. Table (4) shows the SSE, AIC and BIC for each value of K (for all of them the lower number is better).

Table 4 . AIC and BIC values for different K in k-means

	K = 3	**K = 4**	**K = 5**	**K = 6**	**K = 7**	**K = 8**	**K = 9**	**K = 10**	**K = 11**
AIC	79.226	76457	74377	74159	70615	71091	72417	73720	703905
BIC	78258	77158	76362	73766	71301	72564	71658	72611	72821

Based on the BIC and AIC evaluation for different values of K, it is shown that the suitable number of K is seven Result of clustering are shown in the table (5). For each cluster center, the "Passenger per Minute" value was calculated by allocatingthe passengers to headway. This value shows the average arrival rate for the considered cluster. Each cluster is labeled with this number.Therefore,cluster with the biggest passenger per minute is labeled with 0 and the cluster with least value is labeled with 6 as it is shown in table (5).

Each day consists of 29 half hours thateachhalf houris replaced with its cluster label and, as a result, each day in data set is shown with a chain of clusters.

Table 5 . Clustering result for half hours

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
Cluster Size	1082	42	114	387	465	618	916
Center Point Passenger Average	9.335	2.988	23.272	10.158	2.367	14.27	4.527
Center Point Headway Average	8.3	28.3	8.106	16.666	15.014	8.467	8.798
Passenger per Min	1.125	0.106	2.87	0.610	0.158	1.68	0.515
Cluster Label (Value)	3	7	1	4	6	2	5

Table (6) shows this procedure for a sample day (Sep 19 2014).

Table 6 . Day pattern for Sep 19 2014

6:30 – 7:00	7:00 – 7:30	7:30 – 8:00	8:00 – 8:30	8:30 – 9:00	…..	20:30 – 21:00
Cluster 3	Cluster 4	Cluster 4	Cluster 3	Cluster 3	……	Cluster 4
4	6	6	4	4	……	6

The full day pattern for 19 Sep is stated as follows:

4 – 6 - 6 - 4 – 4 - 6 – 4 - 4 – 4 – 4 – 4 – 4 – 6 – 4 – 6 – 4 – 6 – 4 – 4 – 4 – 7 – 6 – 4 – 4 – 4 – 4 – 6 – 4 – 6

By using this idea a daily pattern for each day in our dataset could be created.

3-4-2- Finding the similar days

Now the behavior pattern of each day is extracted and it can be used to find and categorize similar days in the dataset. For this purpose, the K-Means algorithm can be used again to cluster the days. But data set of the days is not simple points and so popular distance measurement algorithm like Euclidean distance can't find proper distance between the data. As shown in the table (4) each day has a time series which shown passenger arrival pattern of that day. As it mentioned previously, in this condition DTW as an appropriate distance measure has been applied. Therefore, to cluster days, the distance measurement method in the K-Means algorithm is changed and Euclidean distance is replaced with the with the Fast DTW algorithm. The rest of the algorithm is exactly the same as the default versionof K-Means. This algorithm used with different k values again (k=3 ...9) and AIC and BIC measures have beencalculated for different values of *K* parameter as shown in table (7).

Table 7 . AIC and BIC values for different K in clustering days

	K= 3	**K= 4**	**K = 5**	**K= 6**	**K = 7**	**K = 8**	**K = 9**
AIC	118406	134513	146832	150285	140701	148713	145871
BIC	120309	135109	147842	151681	142581	150951	146510

Based on the AIC and BIC values, the suitable number of K is three that have been explained as follows.

1. The first cluster (Cluster A) consists of 22 days which most of them are off days. Among these 22 days, 17 days were Fridays (which is weekend in Iran), four of them are religious holidays. Finally, one of them was a regular day and it is assumed as an outlier which is caused by data faults.
2. The second cluster (Cluster B) includes 30 days. Most of the days in this cluster were those before holidays. In addition, there were some of wednesdays in this cluster. In Iran some schools and most of the companies (especially in public sector) are close in the Thursday too and so this had some effect on Wednesday's patterns. In addition, there were two regular days in this cluster.
3. The third Cluster (Cluster C) contains 64 days. The most significant attribute of this cluster is containing regular work days. There was not any suspicious data in this cluster.

3-4-3- Headway - Passenger per Minute plots creation

After clustering similar days, headway's effect on passenger downfall rate could be considered for more investigation. Similarity in days is based on the pattern of the whole day and not necessarily similarity of eachhalf of hours' pattern. Figure 4 shows headway – passenger per minute plot for time interval 7:00 – 7:30 in the cluster A. Each point in the plots represents one day in the cluster A.

This plot shows that in most of the days of cluster A, the passenger per min value has similar (not equal) values in different days. But there are some unusual patterns which are important for behavior analysis.

Based on the above explanation, for eacha time period of a cluster, the plot (figure (4)) has been drawn.

3-5- Applying Headway Interval Detection Procedure

It is shown a repetitive pattern on the obtained plots. If the noises are not considered, in all the plots there is a specific pattern (which can be too small) that has the absolute maximum with zero slopes. For example, in figure 2, this specific pattern is between headways with value 14.8 and 15.3. To find the headway interval in figure 4, the proposed algorithm is applied. The point with the maximum number of passenger/min in figure 4 is *A*. Then, for different values of (*n*), the error for different adjacent of *A* is calculated that has been shown in table 8. This table shows the appropriate value of *n* is three and the best adjacent points for *A*, are *B* and *C*.

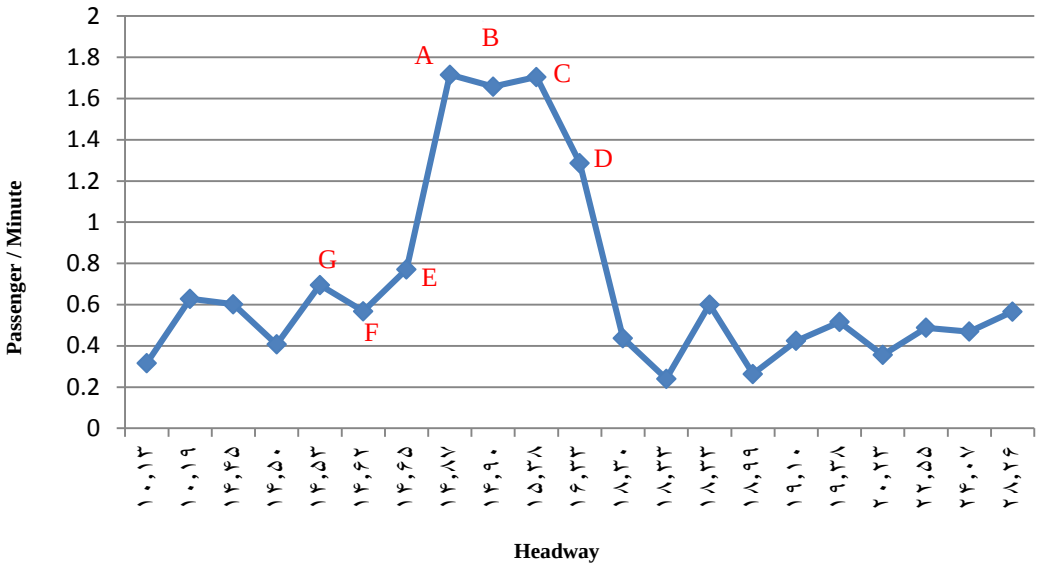

Figure 4. headway – passengers per minutes plot for time interval between 7:00 and 7:30

Table 8 . Result of $e_n$for different values of n in figure 4

Adjacent Points	A , B	A ,B ,C	A ,B ,E	A ,B ,F	A ,B ,C,D	A,B,C,E	A,B,E,F	A,E,F,G
$\mathbf{e_n}$	0.035	0.033	0.34	0.39	0.125	0.262	0.307	0.308

Results of the proposed algorithm for headway values show an important finding. If headway is set between this time intervals, on-board passenger rate per minute could be intensified to its maximum and so the minimum downfall of the passengers would have been happened. In other words, this plots shows when the headway value is bigger than specified time interval, the passengers will prefer to use other transportation modes.In addition, when the headway value is lower than this time interval, fewer numbers of passengers will be boarded. Therefore, the historical pattern in data shows that there is a time interval for headway in each time period.

In this approach, 29 headway intervals will be found which each of them is for one time period. But in the real world it is not possible to use 29 different headways for each day. But most of these headway intervals have overlap with each other. In other words, many of them could be merged together and create one common headway interval for the whole day.

Figure (5) shows an example of this state for cluster *C*. Five time periods has been chosen for cluster *A* and efficient headway interval (which is driven from the headway-passenger per Minute plots) is specified for each of them. Obviously, these intervals have a common interval between 14.1 and 14.7 (which is determined with two dotted lines). Therefore, instead of using five headways for the selected time periods of cluster C, one headway interval (14.1 – 14.7) can be used.

Based on the overlaps, the optimum number of headways can be identified. The table (9) describes the optimum number of headways and headway intervalsfor each cluster. The first column of table (9) shows the cluster label and the second column describes the headway intervals. The first row of this table is the final time periods for each headway, and the second row show the efficient headway interval for the time period (by minutes).

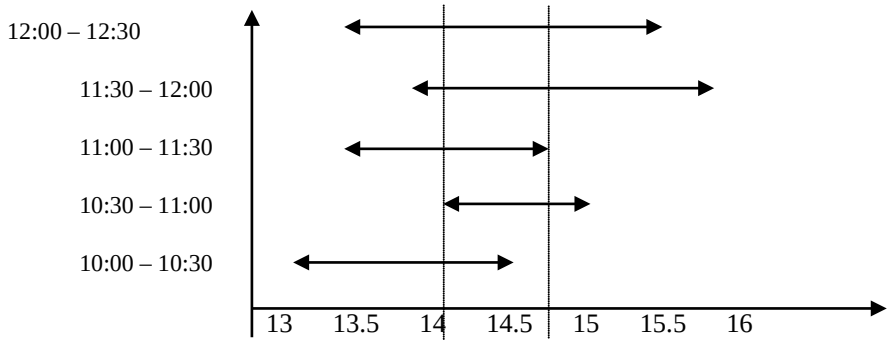

Figure 5 . headway interval overlaps in cluster C day's morning

Table 9 . The final headway values for Mashhad's dataset

Cluster A	**Time period**	7:30 – 12:30		12:30 – 17:00		17:00 – 20:30	
	Headway interval	14.1 – 14.7		19.2 – 20.8		16.4 – 18.6	
Cluster B	**Time period**	7:30 – 11:30		11:30 – 15:30		15:30 – 18:30	18:30 – 20:30
	Headway interval	12.3 – 13.5		16.2 – 17		8.4 – 9.5	14.1 – 16.8
Cluster C	**Time period**	7:30 – 9:00	9:00 – 12:00	12:00 - 15:00	15:00 -17:00	17:00 – 19:00	19:00 – 20:30
	Headway interval	5.1 – 6.9	10.5 – 12.2	15.1 – 15.9	9.8 – 11.2	6.4 – 8.1	16.1 – 17.3

4- Conclusion

Bus headway detection was one of the most challenging issues in the public transportation systems. There are a lot of factors and complications in this problem and any fault in headway determination will increase passengers waiting time and passengers downfall rate. Moreover, big growth in information and communication technologies results in considerable improvements in AVL systems that enables them to gather and store huge amount of data. The data sets include useful information about passengers, buses, stations and lines. These data are widely used in bus scheduling problems, but it is not used for headway determination.

Therefore, in this study historical data of bus transportation were used todevelop an intelligent procedure to determine headway interval based on the passenger downfall rate.Initially, a Similar Days Detection approach was proposed which was able to find similar days of the data set based on passenger behaviors. This approach used K-Means and replaced Euclidian distance measurement with Fast DTW that enables it to compare time series. Afterwards, a Headway Interval Detection Procedure is developed to use specific patterns in each cluster and find appropriate headway values for each time period and the optimum number of headway intervals.

References

Lobel, A., 1999. Solving large-scale multiple-depot vehicle scheduling problems. In: Wilson, N.H.M. (Ed.), Computer-Aided Transit Scheduling. Lecture Notesin Economics and Mathematical Systems, vol. 471. Springer-Verlag, pp. 193–220.447–458.

Kwan, R.S.K., Rahin, M.A., 1999. Object oriented bus vehicle scheduling – the BOOST system. In: Wilson, N.H.M. (Ed.), Computer-Aided Transit Scheduling.Lecture Notes in Economics and Mathematical Systems, vol. 471. Springer-Verlag, pp. 177–191.

Mesquita, M., Paixao, J.M.P., 1999. Exact algorithms for the multi-depot vehicle scheduling problem based on multicommodity network flow typeformulations. In: Wilson, N.H.M. (Ed.), Computer-Aided Transit Scheduling Lecture Notes in Economics and Mathematical Systems, vol. 471. Springer-Verlag, pp. 221–243.

Banihashemi, M., Haghani, A., 2000. Optimization model for large-scale bus transit scheduling problems. Transportation Research Record 1733, 23–30.

Freling, R., Wagelmans, A.P.M., Paixao, J.M.P., 2001. Models and algorithms for single-depot vehicle scheduling. Transportation Science 35 (2), 165–180.

Haghani, A., Banihashemi, M., 2002. Heuristic approaches for solving large-scale bus transit vehicle scheduling problem with route time constraints.Transportation Research 36A, 309–333.

Haghani, A., Banihashemi, M., Chiang, K.H., 2003. A comparative analysis of bus transit vehicle scheduling models. Transportation Research 37B, 301–322.(Eds.), Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems, vol. 430. Springer-Verlag, pp. 115–129.

Huisman, D., Freling, R., Wagelmans, A.P.M., 2004. A robust solution approach to the dynamic vehicle scheduling problem. Transportation Science 38 (4),447–458.

Shangyao Yan,2007, Intercity Bus Scheduling Model Incorporating Variable Market Share, Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions,Volume: 37, Issue: 6,921 – 932

Zhu Chang-sheng,2010, the research in public transit scheduling based on the improved genetic simulated annealing algorithm, Computational Intelligence and Natural Computing Proceedings (CINC), Second International Conference on, Volume 2, 273 – 276

Zhiwei Yang,2008, Research on Bus Scheduling Based on Artificial Immune Algorithm, Wireless Communications, Networking and Mobile Computing. WiCOM '08. 4th International Conference on, 1 - 4

Matias, L ,2010, Validation of both number and coverage of bus schedules using AVL data , Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference, 131 - 136

O. Adebisi, 1986,a mathematical model for headway variance of fixed-route buses,Transportation Research Part B: Methodological, Volume 20, Issue 1, Pages 59–70

P. Furth, B. Hemily, T. Muller, J. Strathman, Uses of archived AVL–APC data to improve transit performance and management: review and potential, Transport. Res. Board (2003).

Ming-Jun Liua, b, Bao-Hu Mao, Shao-Kuan Chen, Li-Ping Gao, Quan-Xin Sun ,2010, An

Exploratory Hazard-based Analysis of the First Discharge Headway,Procedia - Social and Behavioral Sciences, 6th International Symposium on Highway Capacity and Quality of Service, Volume 16, Pages 536–547

Chuanjiao SUN, Wei ZHOU, Yuanqing WANG,2008, Scheduling Combination and Headway Optimization of Bus Rapid Transit,Journal of Transportation Systems Engineering and Information Technology, Volume 8, Issue 5, Pages 61–67

J. Patnaik, S. Chien, A. Bladikas, Using data mining techniques on apc data to develop effective bus scheduling, J. System. Cybernet. Inform. 4 (1) (2006)
86–90.

Bin Yu, Zhongzhen Yang, Xueshan Sun, Baozhen Yao, Qingcheng Zeng, Erik Jeppesen ,2011,Parallel genetic algorithm in bus route headway optimization,Applied Soft Computing, Volume 11, Issue 8, Pages 5081–5091

Yanhong Li, Wangtu Xu, Shiwei He,2013,Expected value model for optimizing the multiple bus headways, Applied Mathematics and Computation, Volume 219, Issue 11, Pages 5849–5861

Yang Hairong, Changsha, 2009, Optimal Regional Bus Timetables Using Improved Genetic Algorithm,Intelligent Computation Technology and Automation. ICICTA '09. Second International Conference on (Volume: 3), 10-11 Oct. 2009, 213 – 216

Dirk L. van Oudheusden, William Zhub, 1995, Trip frequency scheduling for bus route management in Bangkok, European Journal of Operational Research, Volume 83, Issue 3, Pages 439–451

Sho-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao, 2012, Data mining techniques and applications – a decade review from 2000 to 2011, Expert Systems with Applications, Volume 39, Issue 12, Pages 11303,11311

Tak-Chun Fu, 2011, A review on time series data mining, Engineering Applications of Artificial Intelligence, Volume 24, Issue 1, Pages 164-181

B. Barabino, M. Di Francesco, S. Mozzoni, Regularity diagnosis by automatic vehicle location raw data, Public Transport 4 (3) (2013) 187–208

Mandelzys M, Hellinga B (2010) Identifying causes of performance issues in bus schedule adherence with automatic vehicle location and passenger count data. Transp Res Rec 2143:9–15

Ruan M, Lin J (2009) An investigation of bus headway regularity and service performance in Chicago bus transit system. Paper presented at the Transport Chicago, annual conference

M. Chen, X. Liu, J. Xia, S. Chien, A dynamic bus-arrival time prediction model based on apc data, Comput.-Aid. Civil Infrastruct. Eng. 19 (5) (2004) 364– 376.

Q. Chen, E. Adida, J. Lin, Implementation of an iterative headway-based bus holding strategy with real-time information, Public Transport 4 (3) (2013) 165–186.

A. El-Geneidy, J. Horning, K. Krizek, Analyzing transit service reliability using detailed data from automatic vehicular locator systems, J. Adv. Transport. 45 (1) (2011) 66–79.