IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 26, 2018, accepted May 3, 2018, date of publication May 17, 2018, date of current version July 6, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2837498

Systematic Approach to Analyze Travel Time in Road-Based Mass Transit Systems Based on Data Mining

TERESA CRISTÓBAL, GABINO PADRÓN, ALEXIS QUESADA-ARENCIBIA, FRANCISCO ALAYÓN, AND CARMELO R. GARCÍA

Institute for Cybernetics, University of Las Palmas de Gran Canaria, 35017 Las Palmas, Spain

Corresponding author: Carmelo R. García (ruben.garcia@ulpgc.es)

This work was supported in part by the Spain Ministry of Economy and Competitiveness and in part by the State Program for Research, Development and Innovation Oriented to the Challenges of Society.

ABSTRACT Road-based mass transit systems are an effective means to combat the negative impact of transport that is based on private vehicles. Providing quality of service in this type of transit system is a priority for transport authorities. In these systems, travel time (TT) is a basic factor in quality of service. This paper presents a methodology, based on data mining, for analyzing TT in a mass transit system that is planned by timetable. The objective of the methodology is to understand the behavior patterns of TTs on the different routes of the transport network, as well as the factors that influence these patterns. To achieve this objective, the methodology uses clustering techniques to process the GPS data provided by the vehicles of the public transport fleet. The results that were obtained when implementing this methodology in a public transport company are presented as a use case, demonstrating its validity.

INDEX TERMS Road-based mass transit systems, travel time, intelligent transportation systems, data mining, pattern clustering, global positioning system.

I. INTRODUCTION

According to the International Energy Agency, there were an estimated 900 million passenger light-duty vehicles on our roads worldwide in 2015, a figure that is projected to grow to 2 billion by 2040 [1]. There is widespread agreement that road transport systems based on the use of private vehicles have a negative impact. This impact includes degradation of the environment, health and safety on roads, aspects that are particularly pronounced in densely populated areas. The World Health Organization estimates that approximately 3 million people die every year due to health problems caused by pollution [2]. One way to mitigate the negative impacts associated with this type of transport system is to develop efficient public transport systems that provide quality of service. Intelligent Transport Systems (ITS) are an effective means to meet this challenge. Therefore, in modern societies and in the new paradigm of the smart city, ITS has a fundamental role to play.

For public road transport systems to be an alternative to transport based on the use of private vehicles, they must provide quality of service to make them attractive to the general public. In the context of road-based mass transit systems, one of the most important factors that affect quality of service is timetable adherence. Adherence means punctuality in the frequency between services or scheduled stop times. For it to be reliable, information is required on travel time (TT) behavior according to time and space parameters on the transport network. This paper presents a methodology based on data mining for analyzing TT behavior in a context of mass transit systems planned by timetable, based on the GPS data provided by the vehicles of the public transport fleet. The proposed methodology provides information on the TT behavior of the lines according to the time of year, time of day and section of the route. It also evaluates quality of service based on punctuality, according to criteria and metrics that are widely used by transport agencies and the academic community working in this field. In addition, it provides a useful framework for making TT forecasts. Thus, transport planning may be geared towards efficiency and quality of service and it becomes possible to provide reliable information to the public transport user. Specifically, the proposal consists of using classification techniques to study TT behavior in mass transit systems planned by timetable; the originality of this approach lies in the fact that existing works on this topic have mostly looked at planning by frequency. It should also be pointed

out that the required data are commonly used by transport companies and agencies, and therefore do not require the deployment of infrastructure other than that which already exists in mass transit systems. The proposal is consistent with the current ITS paradigm: continuous observation of what happens in the transport network, continuous processing of the data produced by these observations and continuous improvement of the services provided, in order to make transport systems more efficient, safe, sustainable and adapted to the needs of users [3]. Moreover, the proposed methodology may be used to facilitate implementation of traffic control strategies that prioritize public transport vehicles [4], a key aspect of the smart city paradigm. The objective is to have better knowledge of travel time behavior that will enable subsequent studies to focus more on the routes and thus introduce measures to reduce this time or its variability.

The rest of this article is organized into five more sections. The second section lists works related to the proposed methodology. The methodology is described in the third section. Next, the results of a use case implementing the methodology in the study of the travel time of a bus line of a public transport company are presented. The fifth section is a discussion of the results, and the final section draws the conclusions.

TABLE 1. List of abbreviations.

DW	Dwell time
EXP	Journey on which GPS reading was taken
GMM	Gaussian mixture models
KNN	K-nearest neighbors regression
LAT	Latitude according to GPS reading
LIN	Line on which GPS reading was taken
LON	Longitude according to GPS reading
OT	Observed time of arrival at stop
QUA	GPS reading quality indicator
RT	Nonstop running time between two consecutive bus stops
RTD	Relative deviation in travel time for the section
RTV	Run time variation
SCAFC	Smart Card Automated Fare Collection system
ST	Scheduled arrival time
STO	Stop Node Identifier
SVM	Support vector machines
TD	Difference between the observed arrival time and the scheduled arrival time
TDB	Transport database
TIM	Time at which GPS reading was acquired, expressed in Coordinated Universal Time (UTC)
TT	Travel time
VEH	Vehicle
VEL	Vehicle speed at the GPS coordinates
VJ	Vehicle Journey
VJT	Date and time that VJ began, expressed in UTC
VLR	Vehicle location record

A. LIST OF ABBREVIATIONS

Table 1 contains a list of abbreviations used throughout this paper.

II. RELATED WORKS

In order to achieve the objectives of efficiency and quality of service in public transport, a fundamental requirement is to understand the mobility needs and habits of people. Based on this information, the three basic processes on which public transport is based—transport network design, service planning and operations control—may be carried out with guarantees. In [5] a global review was conducted of the methods used to design and schedule a transport network, and in relation to quality assessment, [6] provides an exhaustive review of the methods used to analyze behavior and to evaluate the main parameters that affect it. Technological advances, especially in mobile communications, sensors and computing, have enabled Intelligent Transport Systems to be developed that adapt transit systems to the needs of their users, to be more efficient and to provide greater quality of service. A common feature of this type of system is that it provides information on what happens in the transport network by performing an analysis of its time–space behavior from large amounts of data [7]. In this context of the massive use of data, data mining is a field that is increasingly used in transport engineering. A review of the literature in which data mining has been used to solve some of the problems in transport systems is presented below. Depending on the data source used, these works may be classified into two groups: those based on data related to the movements of travelers in the transport network and those that use data related to the location of the vehicles in the transport network. In both groups there are works that address the three main problems that need to be addressed to achieve efficiency and quality of service.

The works that use data associated with the movements of travelers include studies that: seek to acquire information about the profiles and usage habits of transport network users [8]; measure the use of the network infrastructure by travelers [9]; make predictions about travel times and develop personalized information services for the user [10], [11], based on records generated by the use of the Smart Card Automated Fare Collection system (SCAFC). In [12], socio-demographic factors are also taken into account: location of shopping centers, sports areas, residential areas, etc. The works that propose techniques to obtain mobility patterns of mass transit system users may be grouped into two categories according to the analysis carried out in [13]: those based on statistical methods capable of supplying a self-explanatory model when treating them as the result of a stochastic process, and those that use neural networks. In order to predict total demand in transit systems, based on time series of trips completed during certain time intervals, the use of statistical models is proposed in [14] and neural networks are used in [15]; as an example of mixed procedures, a process to select the generated functions before applying the neural network is introduced in [16]; in [17], the result of two different models of networks is analyzed using time-dependent parameters (trend, cycle and periodicity) in the observed demand data; and a new hybrid optimization algorithm is developed in [18], with set theory and neural network techniques, to predict the volume of passengers by road. As an example of other methods, in [19] the space–time behavior of travelers in a metro network based on the use of cards is studied using clustering techniques.

The location data of public transport vehicles have been used mainly to improve the design of the transport network, to evaluate the quality of service and to make travel time forecasts. The following works are examples of how different issues are tackled with these data: [20] proposes a methodology to evaluate the road network from the point of view of travel time stability through statistical distribution functions. In [21], by means of clustering techniques developed by the authors, the impact of demand and traffic on operational performance is analyzed. By gathering information on passengers boarding and alighting from vehicles, [22] looks at how to avoid overcrowding, which, together with delays in arrival times, can dissuade people from using public transport services; and in [23], diagnostic diagrams of service reliability are generated to determine how the variability of service attributes affects the behavior of travelers. In [24], a methodology for improving the design of the transport network is proposed: it detects the stop, classifies it, generates routes and estimates stop times by processing the vehicle GPS data using clustering techniques. Reference [25] proposes a new metric to evaluate the punctuality of buses using vehicle location data. In [26], the causes of scheduling irregularities are analyzed. In the context of road-based mass transit systems planned by frequency, in [27] and [28], location and passenger movement data are processed using the Gaussian Mixture Models (GMM) clustering technique and ad-hoc metrics with the aim of selecting the best cluster to evaluate quality of service taking into account the day coverage.

With regard to travel time forecasts by processing location data using machine learning techniques, a wide range of studies have been conducted on this subject. In [29], neural networks are used, and classification techniques are used in [30] with k-nearest neighbors regression (kNN), and in [31], k-means and v-means clustering. There are also a considerable number of proposals that tackle travel time predictions using state models and time series. For example, state models, more specifically Kalman filters, are used in [32] and time series in [33] and [34]. Lastly, a hybrid model using Support Vector Machines (SVM) and Kalman filters is proposed in [35].

III. METHODOLOGY

This paper was developed in the context of road-based intercity or long-distance mass transit systems. In this type of system, TT is an important criterion for providing quality of service. Firstly, because travelers want their journeys to last as little as possible, and secondly, because travelers expect punctuality. To achieve these objectives, accurate travel time estimates are needed, and the problem that arises when making these estimates is that travel time depends on factors such as traffic conditions, the travelers who board or alight from the vehicle at each stop on the route, the weather conditions at the time of the bus journey, etc.

This section describes a methodology for systematically analyzing travel time to improve quality of service and identify the factors that affect the travel time on each journey. Identifying these factors makes it possible to obtain TT behavior patterns for the different routes of the transport network and acquire a more precise knowledge of how this time varies, and thus:

- Improve the design of the transport network. Once the factors that affect the routes are known, the routes can be redesigned to reduce travel time.
- Plan more reliably. If the variations in travel time are known, more accurate estimates may be made.
- Control operations more efficiently. If the factors that affect travel time on a route and how this time varies depending on when the route is traveled are known, this greater understanding will enable real-time measures to be adopted that guarantee quality of service.
- Improve quality of service. If travel times are reduced and the reliability of service planning increased, quality of service will improve.

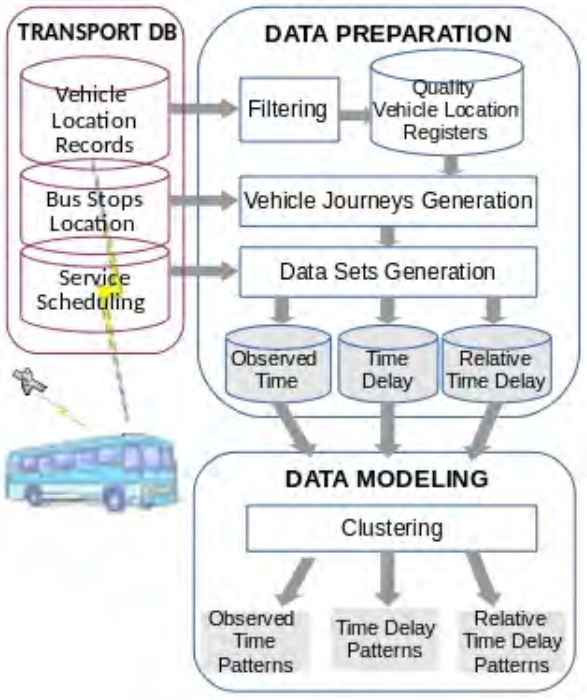


FIGURE 1. General diagram of the methodology.

The methodology used, as illustrated in Fig. 1, was inspired by the process-oriented methodology called CRossIndustry Standard Process for Data Mining (CRISP-DM) [36]. This study mainly followed two stages of this methodology: data preparation and modeling. In the first—data preparation—the data that form the basis of this study were merged and complemented. The second phase—modeling—incorporated modeling tools, based on clustering techniques, which were used to identify the factors that affect the travel time of a route and to obtain its patterns of behavior.

A. FORMALIZATION

This section describes the formal model used to analyze TT. This formalization had two objectives. The first was to ensure

that the methodology can be applied to different types of road-based mass transit systems, and was achieved by applying standards related to conceptual models of public transport systems. The second was to ensure that the information provided is useful, and was achieved by using criteria and metrics for evaluating quality of service in public transport that are widely used by transport agencies and the academic community.

To achieve the first objective, the methodology was based on Transmodel (Reference Data Model for Public Transport) [37]. At the first level of formalization, the transport network was represented. At this level, the entities involved are the nodes and the arcs that physically link the nodes. The nodes represent places in the transport network where transport-related activity takes place: passengers boarding and alighting, schedule controls, ticket sales, etc. Each of these activities are attributes of the entity node, which is represented by n_i, with the subscript i being the identifier of the node. The set of nodes of the transport network is denoted by N, $N = \{n_i\}$. The N nodes are connected by arcs that represent the routes taken by vehicles and travelers, giving rise to a directed graph; this set of arcs is represented by A and gives rise to the physical graph of the transport network, represented by G, $G = (N, A)$. For the purposes of the methodology, the nodes of interest are the passenger boarding and alighting nodes, which will be given the generic name of *stop*, and the schedule control nodes. The set of nodes that fall within these categories is represented by P, where $P \subseteq N$. The arcs of interest are those that represent the routes followed by the buses carrying passengers; the set formed by arcs of this category is represented by W, where $W \subseteq A$. P. On this first level of formalization, the next entity is the route, which is defined as the path followed by the vehicles of the fleet, and comprises an ordered sequence of arcs. Each route is represented by r_i, where the subscript i is the identifier of the route. If route r_i has n arcs, then r_i is specified by n-tuple $(a_i, \ldots, a_n)$, where $a_i, \ldots, a_n \in W$. From the route entity, the line entity is defined: a set of very similar routes from the topological point of view (usually a round trip) represented by l_i, where the subscript i is the identifier of the line. The set of lines in the transport network is represented by L, $L = \{l_i\}$.

The second level of formalization is associated with service scheduling. This scheduling, represented by S, is organized into basic planning units, s_i, so $S = \{s_i\}$. Each s_i planning unit is defined as a set of ordered operations, in which the start and end times and the nodes at which the route starts and ends are specified. For the methodology, there are two aspects of interest at this level. The first is how to specify the calendar dates on which an s_i service must be performed, and the second is the minimum unit of time to be considered in the schedule plan. Specification of the calendar dates is done through a schema in which the different types of calendar day are described. Examples of the most common types are: day of the week, work day, public holiday, weekend, school period, etc. As for the smallest unit of time used for scheduling, this is usually a minute. Each of the completed operations of a line service—a route completed by the corresponding vehicle—is called a Vehicle Journey, represented by VJ. The *VJ* set of all the routes of a line l is represented by VJ_l; to express the *VJ* set of a line l completed in a period of time T the notation VJ_{lT} is used.

Of special interest are the criteria and metrics used to evaluate quality of service in mass transit systems where two types of schedule are distinguished: those based on frequency of service and those based on timetables. The first type is used in urban or short-distance transport [38]. The second type is used for intercity or long-distance transport, where TT and punctuality are two basic criteria for assessing quality of service. In general, the travel time of a VJ on route r, represented by TTr, is formally expressed as a function of two times: the dwell time at each stop, DW, and the time, RT, that it takes to cover each arc of the route:

$$TTr = \sum_{n=1}^{N_s} DW_n + \sum_{n=1}^{N_a} \mathrm{RT}_a \tag{1}$$

Where N_s is the number of stops on r_i, DW_n the time the vehicle remains stationary at stop n of the route (dwell time), N_a the number of arcs on the route and RT_a the travel time of arc a of the route.

The methodology was developed to analyze travel time in a context of an intercity or long-distance mass transit system. For this type of system, a metric used to evaluate punctuality is Run Time Variation (RTV) [39]. The calculation of this metric is expressed below:

$$\mathrm{RTV} = (N_p)^{-1} x \sum_{n=1}^{N_s} \frac{|OT_n - ST_n|}{OT_n} \tag{2}$$

Where N_s is the number of stops on route r_i, OT_n the observed time of arrival at stop n and ST_n the scheduled time of arrival at stop n. The value $(OT_n - ST_n)$ is the deviation from the scheduled arrival time at stop n, which is represented by TD_n. If this value is positive, it indicates a delay with respect to the planned time, a value of zero indicates that the arrival time at the stop is on schedule, and a negative value indicates the vehicle has arrived at the stop ahead of time.

Another aspect that the methodology takes into account is the cost incurred by non-adherence with VJ timetables [40]. In the case of timetabled bus lines, it is assumed that the traveler arrives at the stop moments before the vehicle is scheduled to pass by [41]. Therefore, the methodology assumes that the cost of non-adherence with a VJ timetable, represented by the variable *COST*, is the time cost for the travelers on that VJ of having to wait at the stop, represented by TI_n.

$$\mathrm{COST} = \sum_{n=1}^{N_p - 1} |OT_n - ST_n| \times TI_n \tag{3}$$

B. DATA PREPARATION PHASE

The objective of this phase is to obtain the basic data required to analyze travel time. These data are obtained from the

records stored in the transport database (TDB). For the purposes of this methodology, the entities of interest in the TDB represent the activities that are planned and carried out in the transport network, comprising the following data:

- Geographical location of line stops. These locations are given by their GPS coordinates: latitude and longitude.
- Estimated stop arrival times for each planned VJ. This information is necessary to evaluate the punctuality during the period of analysis.
- The data that represent relevant events that occurred during the VJ, especially periodic updates of the vehicle GPS coordinates during the line services and the data used to ensure integrity.
- The total number of passengers that board and alight at each stop on the route, obtained from traveler payment records.

TABLE 2. VLR data structure.

VEH	TIM	LIN	EXP	LAT	LON	VEL	QUA

The basic data for this methodology are supplied by the readings that indicate the location of the vehicle at a given moment in time. The location of each vehicle is acquired periodically and stored in a data structure named Vehicle Location Record (VLR), this structure is shown in Table 2. The set of all location records is represented by $\{VLR\}$. The subset of $\{VLR\}$, comprising the locations obtained for vehicles on line l routes, in the period T, is represented as $\{VLR\}_{l,T}$. The set $\{QVLR\}$ is obtained from $\{VLR\}$. $\{QVLR\}$ is an integral dataset that guarantees the reliability of the results obtained by the methodology. In this case, integrity means that all records in the dataset $\{QVLR\}$ comply with the following properties:

- They contain a GPS reading of good quality, meaning that it was obtained using the signal provided by at least three GPS satellites and that the reading is less than 10 seconds old. These data properties were obtained from the protocol data used by the vehicle's GPS receiver.
- They were obtained on a VJ that completed a route, meaning that it went through all the planned stops in a coherent fashion (having covered all the planned arcs).

All the data for the $\{QVLR\}$ dataset were acquired through a filtering process (see Fig. 1). The subset of $\{QVLR\}$, comprising the locations obtained for vehicles on line l routes in the period T, is represented as $\{\mathrm{Q}VLR\}_{l,T}$. From the $\{QVLR\}$ dataset, the arrival times at each of the stops on the route are obtained for each VJ; $\{OT\}$ represents the arrival time dataset.

From all the arrival times for all the VJs, the three datasets—$\{OT\}_{l,T}$, $\{TD\}_{l,T}$ and $\{RTD\}_{l,T}$—to be used in the next phase of the data mining project, the modeling phase, were constructed for each line l at each instant of time T.

The dataset $\{OT\}_{l,T}$. This set was used to obtain the behavior patterns of arrival times of the line analyzed during the time period T. Table 3 shows the structure of dataset $\{OT\}_{l,T}$.

TABLE 3. Structure of the data associated with each element of the dataset {*OT*}.

VJ	VEH	VJT	STO	OT

The dataset $\{TD\}_{l,T}$ was obtained by calculating, for each data record, the deviation of the recorded arrival time from the scheduled arrival time. This dataset was used to obtain the behavior pattern of the deviations of the arrival times from the schedule and is therefore an indicator of the cost of said deviations. Table 4 shows the structure of dataset $\{TD\}_{l,T}$.

TABLE 4. Structure of the data associated with each element of the dataset {*TD*}.

VJ	VEH	VJT	STO	DT

The dataset $\{RTD\}_{l,T}$ (Relative Time Delay), was obtained by applying the following transformation to each data record of $\{TD\}_{l,T}$: Let TD_n be the deviation in the recorded arrival time of a vehicle on a VJ at stop n, and let $DTn-1$ be the deviation at stop $n-1$ on that same VJ, thus defining the relative deviation in the arrival time at stop n on the VJ (denoted as $RTDn$) as $RTD_n = TD_n - TD_{n-1}$. A positive value for RTD_n means that the vehicle has traveled the section between stops $n-1$ and n in a time longer than planned, a value equal to zero means that the vehicle has traveled the section in the planned time and a negative value means that the vehicle has traveled the section in less time than planned. The dataset $\{RTD\}_{l,T}$ was used to identify the sections that cause the VJ to run early or late and to understand the pattern that the relative deviations follow in these sections. Table 5 shows the structure of dataset $\{RTD\}_{l,T}$, thus avoiding a cumulative effect in the TD.

TABLE 5. Structure of the data associated with each element of the dataset {R*TD*}.

VJ	VEH	DAT	STO	RDT

C. MODELING PHASE

The objective of this phase is to gain an understanding of the TT behavior of the VJ according to different variables. For a traveler on the route, TT is the time consumed in going from the origin stop to the destination stop of their journey. The ultimate goal is to understand the TT behavior at each and every stop on the route. Specifically, the aim is to understand how certain time-dependent factors, such as the type of calendar day and the time of day, affect the behavior of these times. Also, to understand how the deviations from the scheduled arrival times at stops, TD_n, develop depending on the section of the route. A final objective is to identify on which sections of the route deviations from the scheduled TT are generated according to the type of calendar day, and time of day.

The methodology employed clustering techniques to group the different TT data according to their similarity; this type of clustering technique was chosen because they are capable of handling large datasets that are frequently used in the context of transport, specifically the k-medoids algorithm [43], which is one of the most robust against noise. A medoid may be defined as the element of a group whose average dissimilarity to all elements in the group is minimal. It is the point located the closest to the center in the whole group.

Once a cluster solution has been obtained, the next step is to evaluate the validity of the solution. There are various ways of carrying out this evaluation, which may be broken down into three categories [44]. The first category consists of techniques based on external metrics, which measure the coincidence of the groups with previously generated labels for that class. The second category consists of techniques that use internal metrics, which measure the intrinsic information of each dataset. Finally, the third category consists of techniques that use relative metrics, which are based on the comparison of several different clustering solutions. For the purposes of this study, an internal index was chosen to measure the quality of the clusters in the first instance: the silhouette function [45]. This measures the consistency of the segments generated, based on the tightness and separation of its elements, and is computed by the following formula:

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & if\ a(i) < b(i) \\ 0, & if\ (a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, & if\ b(i) < a(i) \end{cases} \quad (4)$$

In Formula (4) $a(i)$ is the average distance from object i to the other objects within the cluster and $b(i)$ is the smallest average distance from i to all the objects of each of the clusters to which i does not belong.

The k-Medoids clustering technique was applied to datasets $\{OT\}_{l,T}$, $\{TD\}_{l,T}$ and $\{RTD\}_{l,T}$.

IV. RESULTS

This section presents the results obtained in a use case of the methodology. The use case consisted of analyzing the behavior over one year of the arrival times at the stops on a line of the public transport company Global Salcai-Utinsa. This is a company that operates on the island of Gran Canaria (Canary Islands, Spain) and is the main intercity transport company on this island; it has a fleet of 304 vehicles operating on a transport network with 2686 stops, 110 different routes and 2395 daily routes. Every year, its vehicles travel around 25,000,000 kilometers, transporting 20,000,000 passengers.

With regard to the tools used, in the data preparation phase, Oracle was used for the database system and Pentaho for integration and visualization. In the modeling phase, the RStudio framework was used; more specifically, the PAM function of the Cluster package [46], selecting the Euclidean distance as the metric for calculating the dissimilarities between the data and without determining the initial medoids.

TABLE 6. Scheduled VJs on the analyzed line service Monday to Friday (excl. public holidays).

VJ ID	Departure time	VJ ID	Departure time	VJ ID	Departure time	VJ ID	Departure time
1	06:30	9	10:40	17	14:40	25	18:40
2	07:10	10	11:10	18	15:10	26	19:10
3	07:40	11	11:40	19	15:40	27	19:40
4	08:10	12	12:10	20	16:10	28	20:10
5	08:40	13	12:40	21	16:40	29	20:40
6	09:10	14	13:10	22	17:10	30	21:30
7	09:40	15	13:40	23	17:40	31	22:15
8	10:10	16	14:10	24	18:10		

The line selected was number 210, and all the VJs of this line follow the same route. With regard to the route followed by the bus line, it should be noted that it starts in the city of Las Palmas de Gran Canaria, which is the island's main traffic hub, and ends in the city of Arucas, one of the largest population nuclei on the island. It crosses urban areas and non-urban areas, and some of its stops are near health, educational and commercial centers. Therefore, it is an illustrative use case of a bus line since the travel times of its different VJs are affected by different factors related to demand, the calendar, the time of day, traffic conditions, etc. Fig. 2(a) shows an aerial view of the line service with the location of each of its stops, with those considered significant for this study highlighted in red, as will be explained below. In Fig. 2(b) the same bus line is represented schematically, distinguishing between the different types of road that it transits. The route has 30 stops and one control point, and covers a distance of 23 kilometers. The period studied was the whole of 2015. In this period, the VJs were scheduled according to three types of calendar day: the first (type 0), was Monday to Friday, excluding public holidays, the second (type 1), Saturdays, and the third (type 2), Sundays and public holidays. Table 6 shows VJ schedule planning on the days pertaining to the first type (Monday to Friday, excluding public holidays). For this type of day, it may be seen that the first VJ started at 06:30, that between 07:10 and 20:40 a VJ started every 40 minutes, and that the last two VJs started at 21:30 and 22:15. Table 7 shows VJ schedule planning on the days pertaining to the second and third type (Saturdays, Sundays and public holidays). On these days, a VJ was scheduled for every hour between 08:40 and 20:40, with the last two VJs starting at 21:30 and 22:15.

TABLE 7. Scheduled VJs on the analyzed line service for Saturdays, Sundays and public holidays.

VJ ID	Departure time	VJ ID	Departure time	VJ ID	Departure time	VJ ID	Departure time
1	08:40	5	12:40	9	16:40	13	20:40
2	09:40	6	13:40	10	17:40	14	21:30
3	10:40	7	14:40	11	18:40	15	22:15
4	11:40	8	15:40	12	19:40		

As for the arrival time at each of the stops, the schedule provided for the same travel time for each stop regardless of the type of day and time of day of the VJ. The smallest

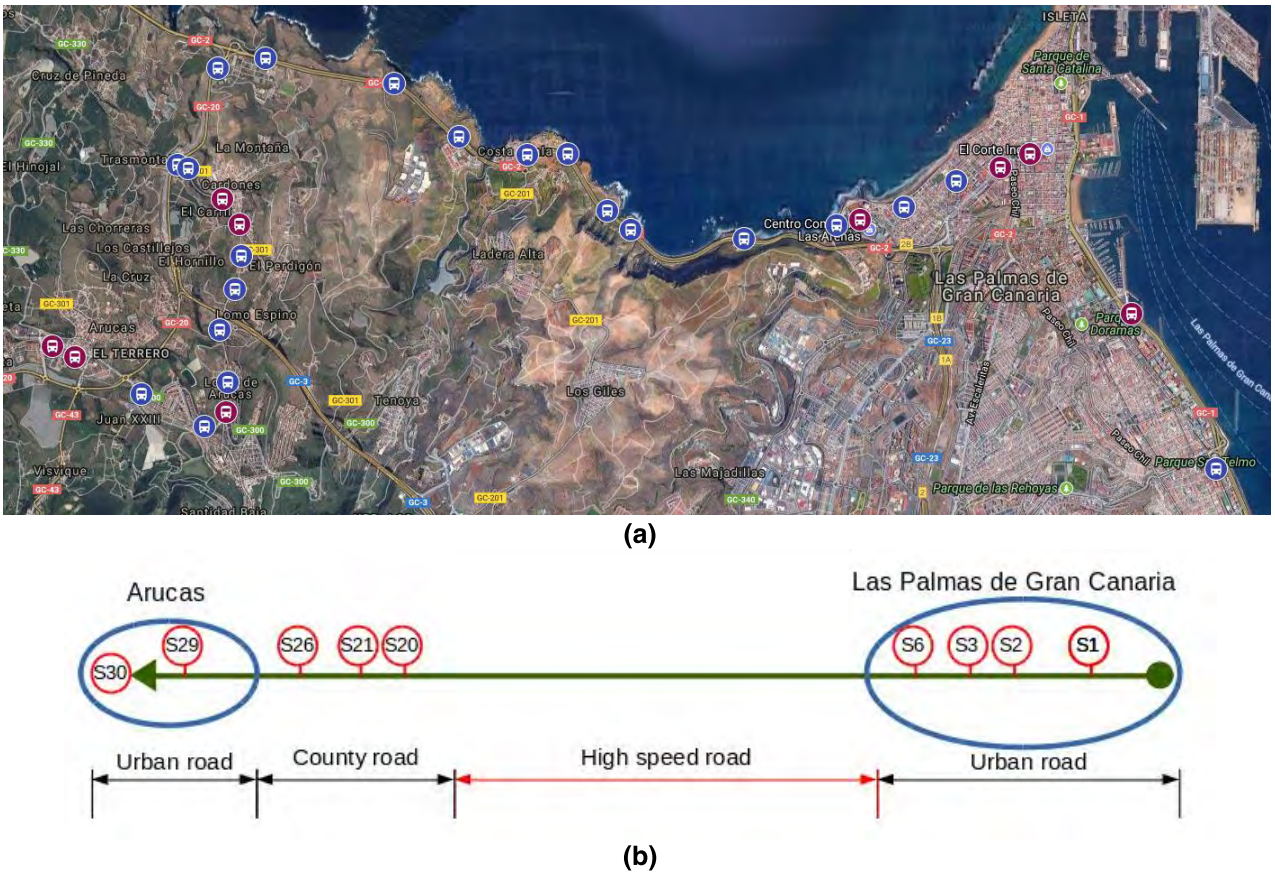


FIGURE 2. (a) Aerial view of the stops on line 210 (those considered significant in red). (b) Schema of the type of sections on the line.

TABLE 8. Arrival time at each stop on the line.

Stop	Arrival time	Stop	Arrival time	Stop	Arrival time
1	2	11	18	22	27
2	7	12	19	23	27
3	10	13	20	24	28
4	12	14	21	25	29
5	14	15	21	26	29
6	15	16	22	27	30
7	15	18	23	28	31
8	16	19	25	29	32
9	17	20	25	30	34
10	18	21	26		

TABLE 9. Number of users of the bus line that boarded or alighted at each stop in 2015.

Stop	Passengers	Stop	Passengers	Stop	Passengers
0	42960	10	13	21	8304
1	4856	11	13	22	1833
2	22772	12	2279	23	1530
3	9284	13	3472	24	704
4	3700	14	52	25	3301
5	2494	15	4446	26	10534
6	19571	16	1660	27	4047
7	251	18	2737	28	1496
8	14	19	210	29	9764
9	62	20	8301	30	42840

unit of time established in this schedule is a minute. Table 8 shows the planned arrival times for each of the 30 stops on the route. The first stop, stop 0, is not included in Table 8 since it is assumed that the vehicle starts the VJ at the scheduled time. The control point, labeled number 17, has also not been included in the table. Each stop on the line has been identified in the order of arrival following the set route; the stops correspond to the labeled points 0 to 16 and 18 to 30.

According to data from the TDB, the 10 stops that were most used by the passengers on this line in 2015 were: 0, 1, 2, 3, 6, 20, 21, 26, 29 and 30. Table 9 shows the number of passengers that board and alight at each of these stops. In Fig. 2(b) these stops are represented with numbered icons; the number indicates the order of that stop on the route, with the exception of the origin stop.

A. RESULTS OF THE DATA PREPARATION PHASE

Table 10 shows each dataset that was processed in the data preparation phase. These data refer to the year studied (2015). The total number of position readings obtained from the entire vehicle fleet after completion of all the VJs of all the lines defined in the transport network was 51,499,404.

TABLE 10. Number of elements of the datasets used in the methodology.

Set	Number of elements
$\{VLR\}_{2015}$: Total number of location readings obtained in 2015	51,499,404
$\{VJ\}_{210,2015}$: Number of VJs planned for line 210 in 2105	9675
$\{QVJ\}_{210,2015}$: Number of complete and coherent VJs completed on line 210 in 2015	6092
$\{QAVL\}_{210,2015}$: Number of location readings obtained on the VJs in dataset $\{QVJ\}_{210,2015}$	158,300
$\{OT\}_{210,2015}$: Observed arrival times processed in the modeling phase	6092
$\{TD\}_{210,2015}$: Deviations from the arrival times processed in the modeling phase	6092
$\{RTD\}_{210,2015}$: Relative deviations from the times processed in the modeling phase	6092

During this year, 9675 VJs were scheduled for the analyzed bus line. Of these 9675 scheduled VJs, when applying the filtering process, 6092 VJ were classified as complete and coherent from 158,300 location readings. From this integral set of location readings, the three datasets used in the modeling phase were created: $\{OT\}_{210,2015}$, $\{TD\}_{210,2015}$ and $\{RTD\}_{201,2015}$.

B. RESULTS OF THE MODELING PHASE

The objective of this phase is to obtain a pattern that describes the travel time behavior of the VJs on the analyzed bus line. This knowledge would help understand how the travel time varies depending on the type of calendar day and the time of day and the section of the route studied.

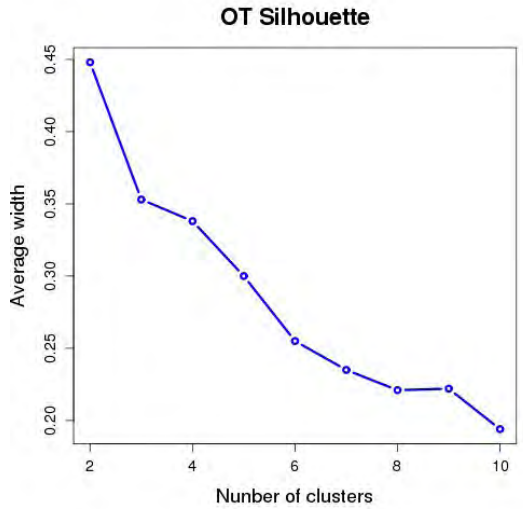


FIGURE 3. Value obtained with the silhouette function for each of the different clusters.

The first step of this phase consisted in modeling TT behavior in 2015. To this end, the k-Medoid clustering technique was applied to the dataset $\{OT\}_{210,2015}$. Nine clustering processes were carried out, generating from 2 to 10 clusters, evaluating in each process the segmentations created with the silhouette function. Fig. 3 contains the average value obtained in each of the nine different cluster groupings, showing that the consistency values decrease as the number of clusters used increases. As an example of the segmentations that were created, Fig. 4 shows the results obtained for the three groupings with the highest value in the metric: the results for two, three and four clusters. The vertical axis represents the arrival time from the start of the VJ, and the horizontal axis the stops analyzed. The three vertical lines represent the three sections into which the route was initially divided: urban, intercity, and urban section. Each graph shows the scheduled time (red line), the medoid of each resulting cluster group (blue line) and the elements classified in each cluster group (gray lines). As may be observed, using two clusters (Fig. 4(a) and 4(b)) the average cohesion value evaluated with the silhouette function is 0.45, and the cohesion values of each of the two clusters are 0.52 (Cluster 1) and 0.36 (Cluster 2). In the case of three clusters, the average value for the silhouette function was 0.35, with the cohesion values for each of the clusters 0.40 (Cluster 1), 0.44 (Cluster 2), and 0.21 (Cluster 3) (see Fig. 4(c), 4(d) and 4(e)). Finally, using four clusters in the group, the average cohesion value for the four clusters was 0.34, the value of each being 0.37 (Cluster 1), 0.36 (Cluster 2), 0.20 (Cluster 3), and 0.33 (Cluster 4) (see Fig. 4(f), 4(g), 4(h) and 4(i)). In this evaluation of the cluster groups, the two groups that produced the highest values were those obtained using two and three clusters. Although the group using two clusters produced the highest average consistency and cohesion value, the group of three clusters gave more precise information about arrival time behavior at the stops. If we compare both groups, we can conclude that the group using three clusters is a refinement of the result obtained with two clusters, and that three TT behavior patterns may be distinguished: Cluster 1 groups the VJs that arrive at the stops in a shorter time, Cluster 2 groups those that take more time than the VJs in Cluster 1, and Cluster 3 groups the VJs with the latest arrival times. In addition, the number of data records in each of the three clusters is significant; 1,777 in Cluster 1; 2,411 in Cluster 2; and 1,904 in Cluster 3. For the above reasons, the grouping of three clusters was taken as the reference for classifying TT behavior.

The second step of the modeling phase consisted of obtaining the behavior patterns for the deviations from the arrival times at the selected stops on the route. To this end, the reference grouping of three clusters was used. Therefore, three patterns were generated, which were defined as the difference function between the observed arrival time and the scheduled arrival time. Fig. 5 shows the data generated with this difference function grouped in each cluster: Cluster 1 Fig. 5(a), Cluster 2 Fig. 5(b) and Cluster 3 Fig. 5(c). The vertical axis represents the deviation of the arrival times from the scheduled times and the horizontal axis, the selected stops. The three vertical red lines represent the same as in Fig. 4. Each graph shows the medoid of the cluster group (blue graphs) and the deviations from the scheduled arrival time for each VJ from each cluster group (gray graphs).

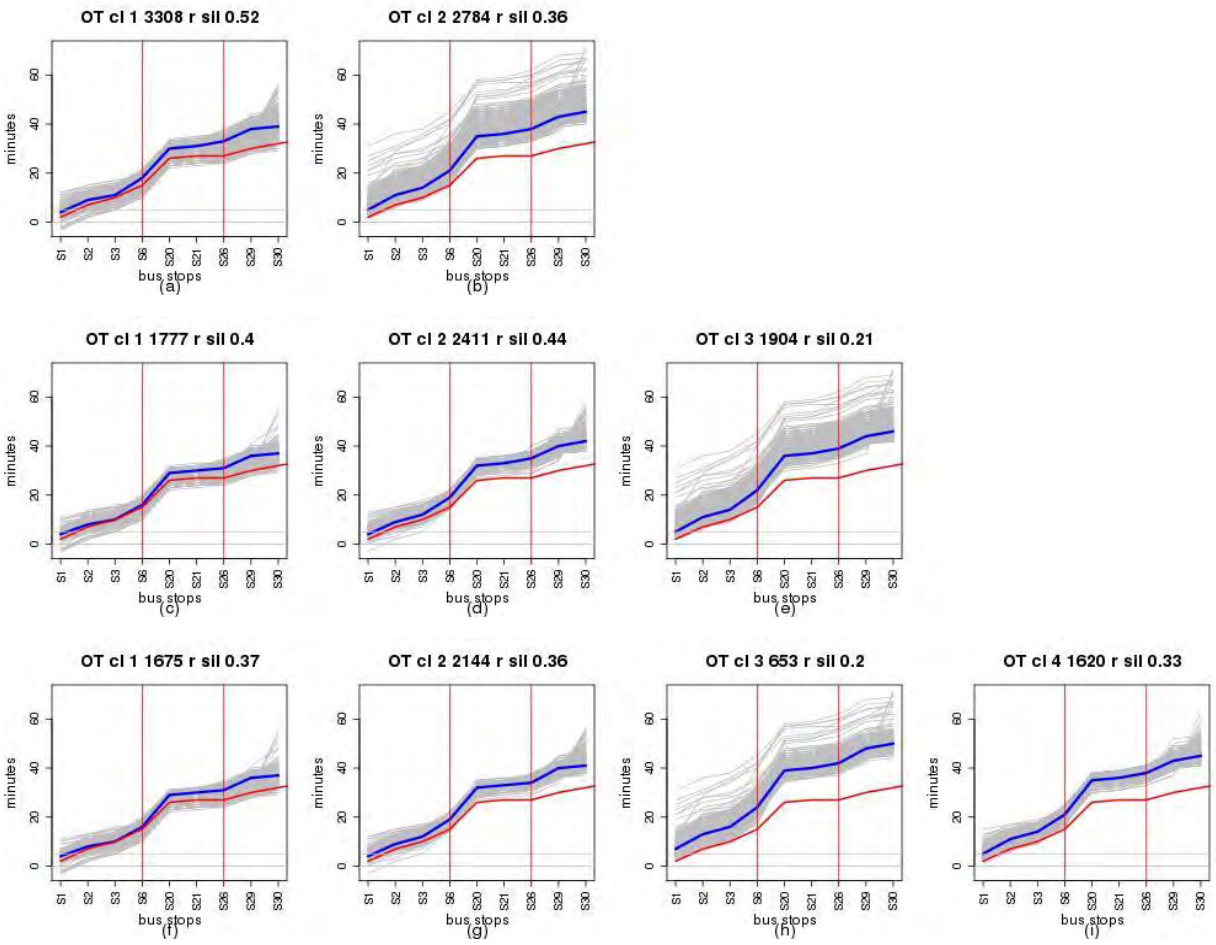


FIGURE 4. Result of clustering the dataset $\{OT\}_{210,2015}$, with two, three and four clusters using the *k*-Medoid technique.

The third step of the modeling phase identified the sections that generated delay and the sections in which there was a reduction in delays. For this, a group of three clusters was created with the dataset $\{RTD\}_{210,2015}$. The results are shown in Fig. 6. The vertical axis represents the relative deviation of the VJ at each stop, and the horizontal axis the stops analyzed. As in the previous two figures, the three vertical red lines represent the three sections into which the line route was divided. Each graph shows the medoid of each of the three resulting cluster groups (blue graphs) and the elements classified in each cluster group (gray graphs).

V. DISCUSSION

From the results obtained in the analysis of the arrival times at stops, it may be concluded that these times do not follow a single pattern, as was assumed in the bus timetable. From these results three behavior patterns were obtained. The first pattern relates to the VJs that reach the stops on the route in the least amount of time (Cluster 1, Fig. 4(c)), the second pattern, the VJs that take longer than the first cluster (Cluster 2, Fig. 4(d)), and the third, the VJs that take the most time to reach the stops (Cluster 3, Fig. 4(e)). The pattern of Cluster 1—the cluster with the greatest schedule adherence—is represented by its medoid, which indicates a deviation from the schedule that rarely exceeds 5 minutes, the time threshold considered tolerable according to studies carried out by various public transport agencies (see Fig. 5(a)). Nevertheless, it is noteworthy that a considerable number of VJs arrive before the scheduled time (in Fig. 4(c) the VJs below the red line representing the schedule and in Fig. 5(a) with TD the VJs with negative values). Non-adherence with schedules when arriving ahead of time is an event that should not occur on routes that are planned by timetables. Conversely, another behavior evinced by the results is that the greatest VJ delays accumulate on the final part of the route, specifically from stop 20 onwards—the first stop of the county road section as shown in Fig. 2(b). On this final part of the route, delays generally exceed 5 minutes, and in the clusters that show behavior patterns of greater deviation from the schedule, this may even exceed ten minutes. This fact is also relevant to VJ scheduling, since it implies that part of the time planned between the end of the delayed VJ and the next to be carried

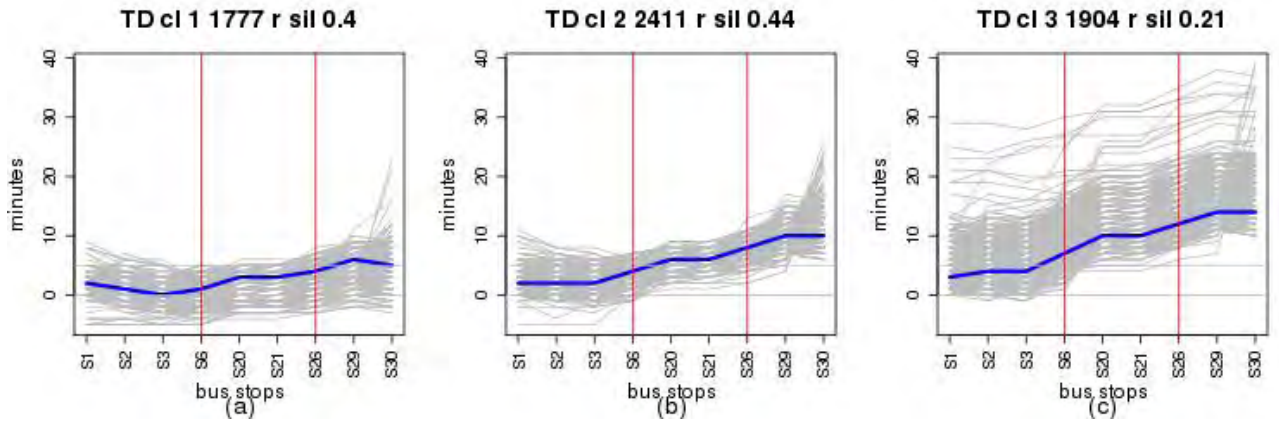


FIGURE 5. Result of the clustering process with 3 clusters applied to the dataset $\{TD\}_{210,2015}$.

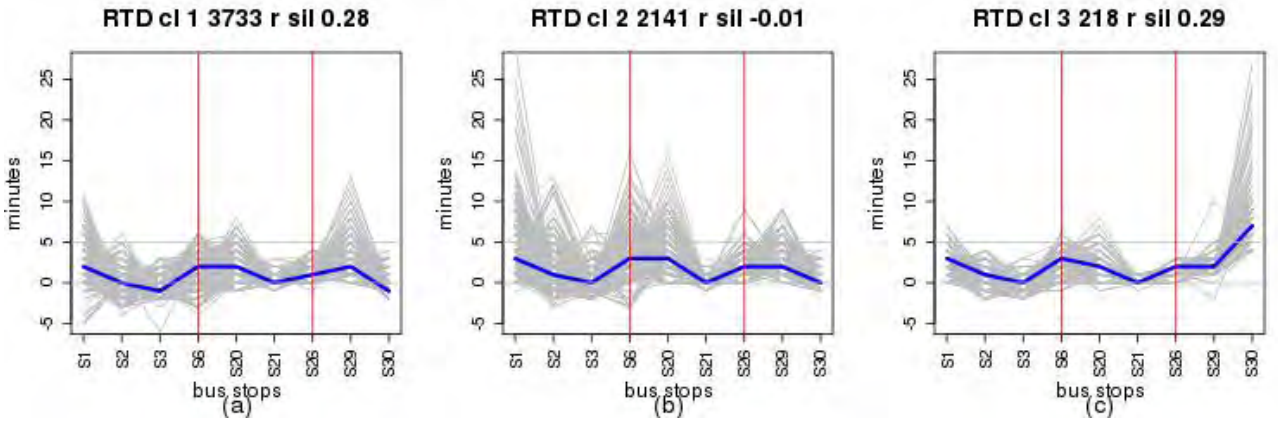


FIGURE 6. Result of the clustering process with 3 clusters applied to the dataset $\{RTD\}_{210,2015}$.

out by the same vehicle—a time interval planned so that the driver can rest and the passengers board the vehicle for its next VJ—is consumed by the delay and may result in the late departure of the next VJ to be made by that vehicle. Finally, another generalized behavior revealed by the results is that the deviations in arrival times at the stops are maintained or increase at the following stops on the route. This behavior can be clearly seen in the forms of the medoids in the three graphs of Fig. 5.

As has already been mentioned, it is clear from the results that the scheduling of arrival times assuming constant values at each stop on the route is not realistic. This statement is supported by the fact that the resulting clusters have a considerable number of samples and their medoids acquire different forms. The question that arises now is how to analyze the relationship between them and the type of day and time of day. To conduct this analysis, contingency tables have been used to represent the frequency with which these patterns occur on different types of day and times of day. Fig. 7 shows these tables for the grouping of three clusters. To analyze the relationship with the type of calendar day, two contingency tables were obtained; one with the months of the year (Fig. 7(b)) and another with the days of the week (Fig. 7(c)). To analyze the time of day, four contingency tables were obtained; one with the time of day at which VJs began on "Monday to Friday excluding public holidays" (Fig. 7(d) with VJs between 06:00 and 15:00 and 7(e) with VJs between 16:00 and 22:00); another with the time of day at which VJs began on Saturdays (Fig. 7(f)); and another with the time of day when VJs began on Sundays and public holidays (Fig. 7(g)). In the tables shown in Fig. 7(b) and 7(c) it may be seen that, in the month of August and on Sundays or public holidays, the most frequent pattern is Cluster 1: the VJs that takes the least amount of time to arrive at the stops. In the tables that associate the clusters with the time of day (Fig. 7(d), 7(e), 7(f) and 7(g)) it is clear that arrival time behavior varies depending on the type of day; the behavior is different from Monday to Friday, on Saturdays and on Sundays and public holidays. Moreover, for each of these types of day, the behavior varies according to the time of day. From the results it may be concluded that, in order to adapt a timetable to reality, different forecasts should be used that take into account the following:

- The time of year; August differentiated from the rest of the months of the year.
- The type of day, differentiating Monday to Friday excluding public holidays, Saturdays, and Sundays and public holidays.
- Time of day, differentiating time periods and type of day.

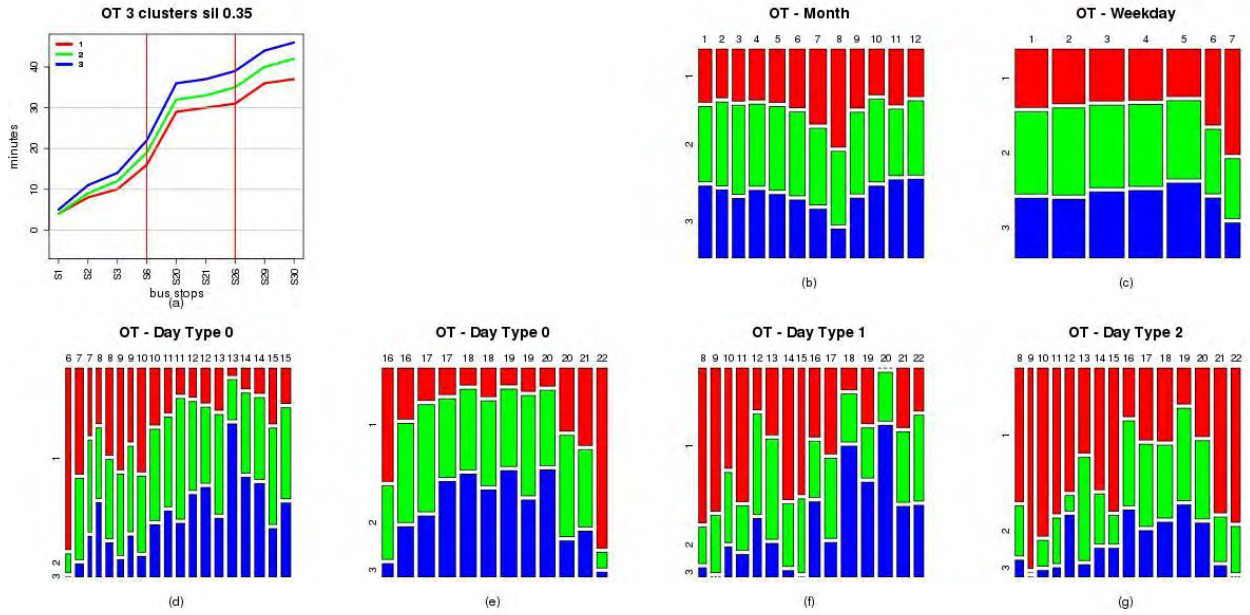


FIGURE 7. Graphs created with the grouping of three OT clusters. (a) Medoids; (b) contingency table of clusters with months of the year; (c) contingency table of clusters with days of the week; (d) contingency table of clusters with VJs on work days (type 0) until 15:00; (e) contingency table of clusters with VJs on work days (type 0) from 16:00 onwards; (f) contingency table of clusters with VJs on Saturdays (type 1); (g) contingency table of clusters with VJs on Sundays and public holidays.

In the analysis of the results obtained in the classifications of the dataset $\{RTD\}_{201,2015}$ formed by the deviations from schedule from one stop to the next on the route, the points of inflection in the medoids are of special interest. These points indicate a change in the behavior of the deviations from the planned schedule, as discussed in Section 3.3, in which the usefulness of this dataset was described. The possible changes are: a section in which the delay decreases; a section in which the delay is maintained; and a section in which the delay increases. In order to improve punctuality, the inflection points marking the beginning of a section in which delays are generated are particularly relevant, since once these sections have been identified, they can be studied to determine the causes of this behavior. At stops 3, 6, 21, 26 and 29 all the medoids have inflection points (see Fig. 6(a), 6(b) and 6(c)). Of these stops, those that begin a section in which a delay is generated are stops 3 and 21 in all the medoids, and stop 29 only in the medoid associated with Cluster 3. The section that begins at stop 3 ends at stop 6, the section that begins at stop 21 ends at stop 26, and the section that begins at stop 29 ends at stop 30. To study the possible causes of this behavior in these sections, it would be necessary to analyze the influences of the DW and RT times on the TT of these sections. If we consider the users of the stops located in these sections, these stops are not the most frequented on the route; this leads to the conclusion that the DW time is not the main cause of the slowness of the vehicle in these sections. To analyze the effect of the RT time on the TT of these routes, the information provided by the transport company's geographic information system was used and it may be observed that a factor that both routes have in common is that they run along single-lane roads in both directions and without any road signs that prioritize public transport vehicles. It could therefore be concluded that the reason for deviations from the schedule in these sections is due to the low speed of the vehicles owing to the conditions of the roads along which they travel. A source of valuable information to analyze the causes of the low speed of these vehicles is GPS readings indicating when vehicles are stationary in these sections, since these readings may follow a pattern that enables these causes to be identified, but this is a subject that falls outside the scope of this paper.

Finally, it should be noted that the proposed methodology enables information on TT behavior to be obtained without specialist knowledge, which would otherwise be necessary if traditional methodologies, based mainly on statistical methods, were used.

VI. CONCLUSIONS

This paper has presented a methodology for analyzing TT in a context of a road-based mass transit system planned by timetables. The methodology, based on data mining, uses the location data of vehicles from the public transport fleet as initial data. It enables the TT of the different scheduled routes to be systematically analyzed, guaranteeing the validity of the results by subjecting the data to validation processes. In addition, in order for the methodology to be suitable for implementation on the greatest possible number of mass transit systems, it has been formalized using standard data models and metrics. From the methodological point of view,

the proposal is based on the k-Medoids clustering technique, used to obtain the TT behavior patterns of the VJs, and the silhouette function, used to evaluate the consistency of the clusters.

In the modeling phase, three sets of input data were used. The first dataset, made up of the recorded arrival times at stops, was used to obtain the TT behavior patterns of the analyzed routes. The second dataset, containing the deviations from the scheduled stop times, was analyzed to understand the behavior of these deviations and to detect where the greatest cost is incurred in terms of quality of service. The third dataset, containing the relative deviations in arrival times at each of the stops, was used to obtain information about the TT behavior in the different sections of a route. This information enables the identification of the sections on the route in which scheduled TT deviations occur (late or early arrival). Once these sections have been identified, they may be analyzed individually to detect the places and causes of these deviations.

This paper presents a use case in which the TT of a transport line of a public transport operator was analyzed, using real data provided by the operator. The results have provided information about the TT behavior of this line according to different types of day and times of day. This information enables possible improvements in the scheduling of stops, making it more reliable and thus improving quality of service. It has also made it possible to identify the sections of the route in which the greatest schedule deviations occur.

ACKNOWLEDGMENT

This study was carried out with the collaboration of the public transport company Global Salcai-Utinsa and the Autoridad Única del Transporte de Gran Canaria (Transportation Agency of Gran Canaria)".

REFERENCES

[1] M. van der Hoeven. *World Energy Outlook*. Int. Energy Agency, Paris, France, Tech. Rep. WEO-2012, 2012, accessed: Feb. 15, 2018. [Online]. Available: https://www.iea.org/publications/freepublications/publication/world-energy-outlook-2012.html

[2] *WHO Releases Country Estimates on Air Pollution Exposure and Health Impact*. World Health Org., Geneva, Switzerland, 2016, accessed: Feb. 15, 2018. [Online]. Available: http://www.who.int/mediacentre/news/releases/2016/air-pollution-estimates/en/

[3] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[4] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[5] V. Guihaire and J. Hao, "Transit network design and scheduling: A global review," *Transp. Res. A, Policy Pract.*, vol. 42, no. 10, pp. 1251–1273, Dec. 2008.

[6] L. Moreira-Matias, J. Mendes-Moreira, J. F. D. Sousa, and J. Gama, "Improving mass transit operations by using AVL-based systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1636–1653, Aug. 2015.

[7] G. Dimitrakopoulos and P. Demestichas, "Intelligent transportation systems," *IEEE Veh. Technol. Mag.*, vol. 5, no. 1, pp. 74–84, Mar. 2010.

[8] B. Agard, C. Morenc, and M. Trépanier, "Mining public transport user behaviour from smart card data," in *Proc. 12th IFAC Symp. Inf. Control Problems Manuf.*, May 2006, pp. 399–404.

[9] N. Lathia, J. Froehlich, and L. Capra, "Mining public transport usage for personalised intelligent transport systems," in *Proc. 10th IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 887–892.

[10] N. Lathia and L. Capra, "Mining mobility data to minimise travellers' spending on public transport," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2011, pp. 1181–1189.

[11] N. Lathia, C. Smith, J. Froehlich, and L. Capra, "Individuals among commuters: Building personalised transport information services from fare collection systems," *Pervasive Mobile Comput.*, vol. 9, no. 5, pp. 643–664, Oct. 2013.

[12] B. Du, Y. Y. Yang, and W. Lv, "Understand group travel behaviors in an urban area using mobility pattern mining," in *Proc. IEEE 10th Int. Conf. Ubiquitous Intell. Comput., IEEE 10th Int. Conf. Auto. Trusted Comput.*, Dec. 2013, pp. 127–133.

[13] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 3, pp. 387–399, Jun. 2011.

[14] R. Xue, D. Sun, and S. Chen, "Short-term bus passenger demand prediction based on time series model and interactive multiple model approach," *Discrete Dyn. Nature Soc.*, vol. 2015, pp. 1–11, Mar. 2015.

[15] D. Celebi, B. Bolat, and D. Bayraktar, "Light rail passenger demand forecasting by artificial neural networks," in *Proc. Int. Conf. Comput., Ind. Eng.*, Jul. 2009, pp. 239–243.

[16] Y. Wei and M. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks," *Transp. Res. C, Emerg. Technol.*, vol. 21, no. 1, pp. 148–162, Apr. 2012.

[17] T.-H. Tsai, C.-K. Lee, and C.-H. Wei, "Neural network based temporal feature models for short-term railway passenger demand forecasting," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3728–3736, Mar. 2009.

[18] W. Deng, W. Li, and X. Yang, "A novel hybrid optimization algorithm of computational intelligence techniques for highway passenger volume prediction," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4198–4205, Apr. 2011.

[19] J. Zhao, Q. Qu, F. Zhang, C. Xu, and S. Liu, "Spatio-temporal analysis of passenger travel patterns in massive smart card data," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3135–3146, Nov. 2017.

[20] N. Uno, F. Karachi, H. Tamura, and Y. Iida, "Using bus probe data for analysis of travel time variability," *J. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 2–15, Feb. 2009.

[21] Y. Bie, X. Gong, and L. Zhiyuan, "Time of day intervals partition for bus schedule using GPS data," *Transp. Res. C, Emerg. Technol.*, vol. 60, pp. 443–456, Nov. 2015.

[22] C. Zhou, P. Dai, and R. Li, "The passenger demand prediction model on bus networks," in *Proc. IEEE 13th Int. Conf. Data Mining Workshops*, Dec. 2013, pp. 1069–1076.

[23] V. T. Tran, P. Eklund, and C. Cook, "Learning diagnostic diagrams in transport-based data-collection systems," in *Foundations of Intelligent Systems* (Lecture Notes in Computer Science), vol. 8502. Cham, Switzerland: Springer, Jun. 2014, pp. 560–566.

[24] F. Pinelli, F. Calabrese, and E. Bouillet, "A methodology for denoising and generating bus infrastructure data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 2406–2417, Apr. 2015.

[25] B. Barabino, M. Di Francesco, and S. Mozzoni, "Rethinking bus punctuality by integrating automatic vehicle location data and passenger patterns," *Transp. Res. A, Policy Pract.*, vol. 75, pp. 84–95, May 2015.

[26] S. Mozzoni, R. Murru, and B. Barabino, "Identifying irregularity sources by automated location vehicle data," *Transp. Res. Procedia*, vol. 27, pp. 1179–1186, Sep. 2017.

[27] J. Mendes-Moreira, L. Moreira-Matias, J. Gama, and J. F. de Sousa, "Validating the coverage of bus schedules: A machine learning approach," *Inf. Sci.*, vol. 293, pp. 299–313, Feb. 2015.

[28] J. Khiari, L. Moreira-Matias, V. Cerqueira, and O. Cats, "Automated setting of bus schedule coverage using unsupervised machine learning," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, Apr. 2016, pp. 552–564.

[29] R. Jeong and L. Rilett, "Prediction model of bus arrival time for real-time applications," *Transp. Res. Rec.*, vol. 1927, no. 1, pp. 195–204, Jan. 2005.

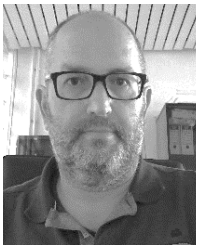
[30] H. Chang, D. Park, S. Lee, H. Lee, and S. Baek, "Dynamic multi-interval bus travel time prediction using bus transit data," *Transportmetrica*, vol. 6, no. 1, pp. 19–38, Oct. 2009.

[31] W. Lee, W. Si, L. Chen, and M. Chen, "HTTP: A new framework for bus travel time prediction based on historical trajectories," in *Proc. ACM 20th Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2012, pp. 279–288

[32] L. Vanajakshi, S. C. Subramanian, and R. Sivanandan, "Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses," *IET Intell. Transp. Syst.*, vol. 3, no. 1, pp. 1–9, Mar. 2009.
[33] W. Suwardo, M. Napiah, and I. Kamaruddin, "ARIMA models for bus travel time prediction," *J. Inst. Eng.*, vol. 71, no. 2, pp. 49–58, Jan. 2010.
[34] G. Chen, X. Yang, J. An, and D. Zhang, "Bus-arrival-time prediction models: Link-based and section-based," *J. Transp. Eng.*, vol. 138, no. 1, pp. 60–66, Jan. 2012.
[35] B. Yu, Z. Yang, K. Chen, and B. Yu, "Hybrid model for prediction of bus arrival times at next station," *J. Adv. Transp.*, vol. 44, no. 3, pp. 193–204, Jul. 2010.
[36] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," *J. Data Warehousing*, vol. 5, no. 4, pp. 13–23, Jan. 2000.
[37] *The CEN Public Transport—Reference Data Model*, CEN Standard CEN/TR 12896-9:2016, 2016.
[38] N. Paulley *et al.*, "The demand for public transport: The effects of fares, quality of service, income and car ownership," *Transp. Policy*, vol. 13, no. 4, pp. 295–306, Jul. 2006.
[39] J. Strathman, T. Kimpel, and K. Dueker, "Automated bus dispatching, operations control and service reliability," *Transp. Res. Rec.*, vol. 1666, pp. 28–36, Jun. 1999.
[40] M. Dessouky, R. Hall, L. Zhang, and A. Singh, "Real-time control of buses for schedule coordination at a terminal," *Transp. Res. A, Policy Pract.*, vol. 37, no. 2, pp. 145–164, Feb. 2003.
[41] P. G. Furth and T. H. J. Muller, "Service reliability and optimal running time schedules," *Transp. Res. Rec.*, vol. 2034, pp. 55–61, Dec. 2007.
[42] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.
[43] L. Kaufman and P. J. Rousseeuw, "Partitioning around medoids (program PAM)," in *Finding Groups in Data Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 1990, ch. 2, pp. 68–125.
[44] E. Aldana-Bobadilla and A. Kuri-Morales, "Clustering method based on the maximum entropy principle," *Entropy*, vol. 17, no. 1, pp. 151–180, Jan. 2015.
[45] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.
[46] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. (2017). *Cluster Analysis Basics and Extensions, R Package Version 2.0.6*. [Online]. Available: shttps://CRAN.R-project.org/package=cluster

TERESA CRISTÓBAL received the B.S. degree in computer science and the M.S. degree in master's degree in intelligent systems and numeric applications in engineering from the University of Las Palmas de Gran Canaria, Canary Islands, Spain, in 1990 and 2014, respectively, where she is currently pursuing the Ph.D. degree in enterprise, Internet and communications technologies.

Since 2012, she has been a Research Assistant with the Institute for Cybernetic, University of Las Palmas de Gran Canaria. Her research interest includes the development of intelligent transport systems for public transport and using data mining-based models for public information services. She is the author of eight articles.



GABINO PADRÓN was born in Caracas, Venezuela, in 1966. He received the B.S. and M.S. degrees in computer engineering from the University of Las Palmas de Gran Canaria, Canary Islands, Spain, in 1990, and the Ph.D. degree in computer engineering from the University of Las Palmas de Gran Canaria in 2015.

Since 1990, he has been a Professor with the Informatic and Systems Department, University of Las Palmas de Gran Canaria. He has authored three books and 25 articles. His research interests include passenger transport system focused on transport network planning, AVL systems, and global positioning system.

ALEXIS QUESADA-ARENCIBIA received the B.S. and M.S. degrees in computer science and the Ph.D. degree in computer science from the University of Las Palmas de Gran Canaria (ULPGC) in 1997 and 2001, respectively. He is currently a Doctor-Employed Teacher with the Computer Science and Systems Department, ULPGC, where he is also the Director of the University Institute for Cybernetics.

His main lines of research include the fields of cybernetics, robotics, artificial vision, and intelligent transport systems. He has authored 70 articles in international and national journals, co-authored five books, and edited 14, eight of them in international journals. He is an assessor of different international journal and conferences. He has taken part in over 30 international conferences and has participated in the organization of over 10. He has taken part in seven research projects, being the lead researcher in three of them. He has also participated in six investigation contracts as the lead researcher. Since 2004, he has been teaching Ph.D. course in the Cybernetics and Telecommunication Program, where he has been the Director since 2011. He has directed the development commission of the new doctorate program—Company, Internet and Communications Technologies—in which he currently teaches different activities. He has directed over 50 final degree projects (engineering, bachelor's, and master's degrees), has been part of several Ph.D. examining committees and directed a doctoral thesis; at present five Ph.D. students are under his tutelage. He has directed and has been a speaker in over 60 training courses.

FRANCISCO ALAYÓN was born in Las Palmas de Gran Canaria, Spain, in 1964. He received the B.S. and M.S. degrees from the University of Las Palmas de Gran Canaria, Canary Islands, Spain, in 1989, and the Ph.D. degree from the University of Las Palmas de Gran Canaria in 2007, all in computer engineering.

Since 1989, he has been a Professor with the Informatic and Systems Department, University of Las Palmas de Gran Canaria. He has authored over 50 articles and 20 inventions. He holds one patent. His research interests include passenger transport system focuses in transport network planning, communications systems, integration of the transport vehicle devices in the company's data network.

CARMELO R. GARCÍA was born in Las Palmas de Gran Canaria, Spain, in 1963. He received B.S. and M.S. degrees in Computer Science from the University of Las Palmas de Gran Canaria, Gran Canaria, Spain, in 1989, and the Ph.D. degree in computer science from the University of Las Palmas de Gran Canaria in 1995.

Since 1987, he has been a Professor with the Informatics and Systems Department, University of Las Palmas de Gran Canaria. He has authored one book, over 70 articles, and 20 inventions. He holds one patent. His research interests include ubiquitous computing, intelligent transport systems, and new technologies for education.