Identification of Passenger Demand in Public Transport Using Machine Learning

R. Thiagarajan
Research Scholar, PG and Research Department of Computer Science, Marudupandiyar College (Arts & Science), Thanjavur.
Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.
E-mail: thiagarajan.nvr@gmail.com

Dr.S. Prakashkumar
Assistant Professor, PG and Research Department of Computer Science, Marudupandiyar College (Arts & Science), Thanjavur.
Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.
E-mail: drpk400077@gmail.com

Received November 05, 2020; Accepted December 15, 2020
ISSN: 1735-188X
DOI: 10.14704/WEB/V18SI02/WEB18068

Abstract

An essential aspect of the transport system is public passenger transport and the Public Transport (PT) movement prediction is significant issues faced in the transport planning area because of its operational importance. In recent years, Intelligent Transportation Systems (ITS) have received a growing amount of interest. There are many advances and innovative applications that have been introduced for a safer, highly efficient, and even congenial environment from PT. A reliable and efficient system of traffic flow prediction is required for accomplishing these applications that build an event with the application of ITS implementations to resolve the potential road situation in advance. However, the PT network efficiency plays the main role for all urban authority areas in which the advancement of both communication and location devices are randomly increasing the data availability generated over the operational platform. In order to recognize trends useful for improving the Schedule Plan, adequate Machine Learning (ML) approaches need to be implemented. Therefore, this paper focused in heterogeneous data that affect the prediction value which is utilized for predicting the demand transport required in the particular route and arrival time of public transport using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) with Seasonal Autoregressive Integrated Moving Average (SARIMA) algorithm to analyze the forecasting of the real-time passenger demand dynamically endorsed the growth of the dynamic bus management and scheduling. Moreover, the accuracy of proposed SARIMA Model is compared with traditional hybrid model such as Gaussian Mixture Model (GMM) with ARIMA model for providing an efficient and robust prediction of PT based on passenger demand.

Keywords

Public Transport (PT), Automatic Passenger Counting, Automatic Vehicle Location, Dwell Times.

Introduction

In modern cities, the effectiveness of PT is a main problem for delivering some service quality necessities like planning a good organization is most important when keeping a healthy relationship through resource use and revenues obtained. At present, most of the PT operators provide their vehicles fitted with antennas such as Global Positioning System (GPS), radio frequency identification, and communications systems. It has capable of transmitting the position of the vehicle Automatic Vehicle Location (AVL) and its driver ship as Automatic Passenger Counting (APC) to the central server [1]. At the World Congress in Paris in 1994, the Intelligent Transportation System (ITS) is introduced. For increasing the safety and effectiveness of road transport systems, electronic and communication technology provide the ITS has been using a computer for traveler information. The primary benefit of the ITS seems to be a safe movement of road transport services.

The smart card file capability has been used as a tool to transport planning, and management is illustrated in several recent studies [2][3]. An especially interesting dimension of the study here is that demand estimation has been provided smart cards record for each passenger's transportation characteristics like traveling day and date, source and endpoint, travel times, etc. Indeed it may have ability to maximize the network of transport as entire if the transport authorities and regulators could take advantage of this knowledge on demand. Few studies have resorted to their input data to the use of smart cards. The most groundbreaking instance can analyze smart card data to predict demand [4-7]. Similarly, the effective method to boost the traffic environment is to build an innovative and reliable transport system that can support us in improved coordinate transport services, distribute the flow of traffic until it is overflowing, and also provide better road entertainment. One of the most popular of these systems is the ITS [8–10]. It is a dynamic structure that incorporates several progressive technologies, such as communicating transportation systems. In the meantime, ITS will boost traffic performance, simplicity traffic congestion, growth of road capacity, decrease traffic accidents, and environmental smog by the compelling benefit of the advancement of 5 G system, extensive on-road sensors, etc.[11–14].

In PT authorities to prepare, operate, and analyze their transit systems, the evaluation of passenger numbers is of vital importance. Currently, passenger counts are collected manually through passenger surveys or human ride checkers contain a small sample. Formerly, the evaluations of travelers may be biased and inaccurate. For the latter, the precision of the manual counts by ride checkers seems to be questionable, because the automated counting systems of the first era have considered earlier considered to be more precise. There are three categories based on Automatic Data Collection (ADC) systems in PT such as APC, Automatic Fare Collection (AFC) systems, and AVL [15].

The organizations of this paper listed as follows, section 2 describes the associated survey regarding technique based ML, section 3 describes the proposed methodology based on passenger demand identification and transportation schedule, and section 4, result and discussion, section 5 discusses the conclusions.

Literature Review

Irrespective of passenger increases in urban areas, there is an essential in incremental of PT in those areas are need to be considered with historical data. It is not possible to deal with the problem of designing effective public transit systems that are subject to organizational and resource constraints. The use of ITS data for strategic planning for public transport has only gained interest when sustainability is becoming an urgent topic in modern times. On several fronts, various ITS applications have made it possible to collect information such as the performance of PT, demand patterns, and ridership [16, 17]. Compared to conventional survey collection methods, the AFC data collected over an extended span provides a beneficial environment for understanding the primary processes of travel behavior. In the public transit system [18], a specific serial number may be used to identify the smart card. The details of the transaction are registered and every time a smart card can be taped. The OBUs can record the physical location of the vehicle at various times, typically with GPS tracking devices. The combination of the two data will help us to measure the congestion on the bus [19].

In particular, since the 1970s, Autoregressive Integrated Moving Average (ARIMA) has been one of the prevalent parametric techniques to predict transportation demand. In predicting short-term traffic data are speed, traffic flow, occupancy, and travel time, the ARIMA model has utilized widely. It is also to forecast traffic flow and international air passenger flow with the seasonality characteristics and trends in traffic data [20],[21]. ARIMA achieves well and strongly in modeling with stationary and linear time series, as indicated in Brooks.

ARIMA models are restricted because they require linear relation between time variables to prevent the structure of non-linear relation from being captured. Several approaches have been used to predict the transport demand for non-parametric techniques [22] [23]. In the non-parametric technique, the frequently used model for implementation is the NN model. It is due to its characteristics like nonlinearity, adaptability, and arbitrary capacity for mapping functions [19]. Basically, without prior knowledge of the relationships between input and output variables, NN can solve complicated non-linear problems [19]. In recent studies, NN-based traffic and transport forecasting models have been implemented like Kalman filter-based multilayer perceptron, multilayer perceptron NN, time-delay NN [24], the radial base function NN, dynamic NN, state-space NN, SVM, etc.

The clustering techniques are a unique way to obtain knowledge about mobility patterns. It was first implemented in [25] to perform cluster analysis on ticketing data. By combining transactions smartcard belonging into daily profiles, each representing time slots while at least one boarding was made by the cardholder using k-means clustering to classify clusters of related day's related to boarding times. In weekly travel activity, a similar analysis is carried out to investigate group behavior, Hierarchical Agglomerative Clustering (HAC) and k-means have applied to bus trips are accumulated into profiles summarizing passenger weekday behaviors [26]. In [27], to retrieve recurrent travel patterns, DBSCAN is applied to individual journey chains, and also based on regularity, k-means++ has been used to cluster passengers.

Methodology

It is a combined methodology of identifying the demand of passenger and running time of vehicle has been proposed automatically by considering both the number of scheduled vehicles and their everyday coverage of passengers using APC and AVL. The working theory is established for those days that should be allocated to the same schedule where route trips have such a similar behavior with different routes in terms of round trip times during the day. Figure 1 has illustrated the workflow of this proposed method.

Let the route of interest is mentioned in term of $L = \{R1, R2, R3, \ldots, Rn\}$. Initially, the time of running and the boarding at every stop have extracted from their original AVL and APC dataset for each $R \varepsilon L$. Secondly, yet there is no APC information available for a particular path, the regular profiles are created by the originally suggested procedure has been used. Otherwise, in accordance with APC data generated using a biased dwell time model based on the demand for peaks and valleys. The output is acquired by the access travel times

which are calculated via AVL data. However, from its initial AVL and APC dataset, the running times of every R ε L and the boarding and alighting at each stop are extracted. It is effectively made in a regular profile when no APC data is obtainable in a particular route. The method is therefore generally suggested biased dwell time model has to consider peak and valley demand which has been created in accordance with APC data. Therefore, the output is computed through the AVL data which is additional to the link travel times

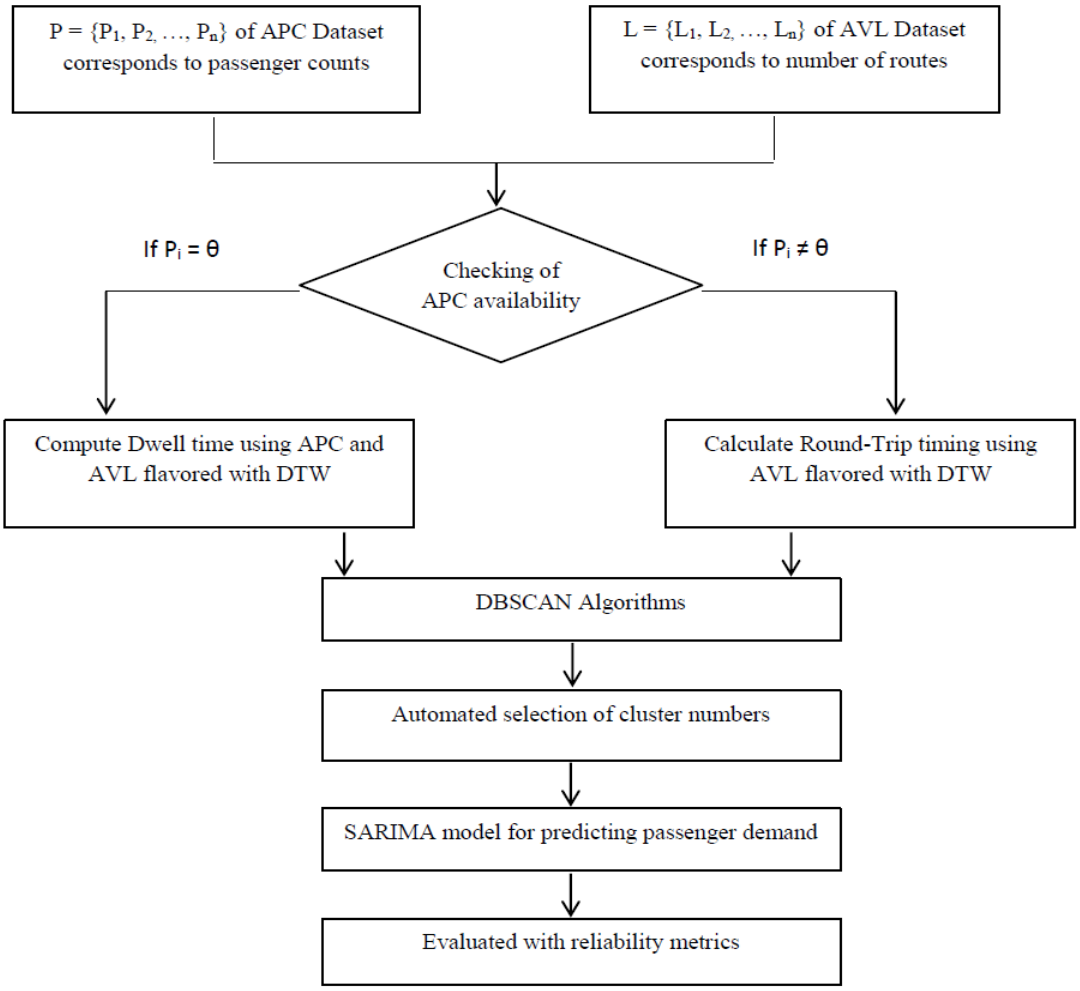

Figure 1 Proposed model of DBSCAN cluster with SARIMA

Similarly, distances matrixes between the days are generated by two steps also cluster them. During the latter, the first task is performed using a Euclidean flavored Dynamic Time Warping (DTW) is solved by the DBSCAN model. Subsequently, a set of a permissible number of schedules is made by clustering S ⊂N, i.e. ∀s ϵ S rather than a single pre-defined s value. For all routes, the steps discussed above are repeated. Additionally, to identify the best number of schedules for placement, the best possible has been selected as s ε S. This is achieved using a two-stage method, while for every pair (R, s), ∀R ϵ L, s ϵ S are the

metric is designed to estimate the clustering result. The previously computed metrics are dwell time APC and domain-oriented weighted mean can be used to identify S. Finally, to measure the proposed Schedule Coverage, a DBSCAN Clustering method is constructed using the clustering pieces obtained for s = S.

The time for each trip is obtained when $Pi \neq \theta$ by including the dwell times at each stop and the travel times of the link is defined in the AVL data. While using multilevel APC data, the demand for peaks and valleys may be expected to be expressed as slight increases or decreases in the calculated round-trip time. With the related datasets (Li, Pi), let R be a route of interest where 't' represents the trip counts and 'c' represents the number of bus stops. This process commences through modeling multiple factors of the dwell time at each stop by decomposition. It could be measured as follows:

$$\delta_{m,n} = \max\left(\alpha \times a_{m,n} \times \beta \times b_{m,n}\right) + alct \qquad (1)$$

Where α and β represents constants denoting the time of alighting and boarding per passenger, respectively, alct represents the time assigned for each stop, such as opening and closing doors. At the same time, $a_{(m,\ n)}$ and $b_{(m,\ n)}$ seem to be the passengers which correspondingly arrive and board at the j stop during k trips. Let $m \in \{1,2, \ldots ,t\}$ and $n \in \{1,2, \ldots ,c\}$ use the available dwell time values (AVL) $\delta_{(m,\ n)}$ and $a_{(m,\ n)}$ and $b_{(m,\ n)}$ (APC) values whereas α, β, alct values can be performed by linear regression method.

DBSCAN Algorithm

Clusters are identified as dense regions using DBSCAN algorithm which are separated by regions of a lower point density. The maximum reach distance density ε and the minimum number of points (MinPts) have two global parameters used in the algorithm. A point, if it has at least MinPts inside a radius ε, where expressed in equation 2, can be called a "core point" i_c.

$$\left|N_{\varepsilon(i_c)}\right| \geq \text{MinPts} \qquad (2)$$

If it has fewer points than MinPts inside ε but it can lie inside the ε range of a core point, such a point may be called a "border point" ib. If it is neither a core nor a border point, such a point is considered a "noise point". For a more comprehensive definition of DBSCAN, a cluster is described by merging the core points ic which are not greater than ε distance apart with their related border points ib. The spatial and temporal patterns are applied separately in the algorithm for mining in which a two-level DBSCAN application can extract the standard Origin-Destinations (ODs): the first one based on the historical alighting stops and the second one based on the boarding stops. Without modifying the conclusions, the order of the two levels is exchangeable. DBSCAN application improves the robustness of the

overall clustering algorithm, and for later passenger segmentation, the results of each stage are useful.

Automated Selection of Number of Schedules

A complex problem in data processing is the selection of the best number of clusters. SARIMA is used to compute a probabilistic score based on entropy when maximized over a set of values, i.e. S, by decreasing the entropy among samples of a similar cluster and exploiting the entropy among samples of dissimilar ones which aims to return the optimum s. However, the limitations in which each application domain encloses, such optimization problems could not lead to a good solution. Consequently, to resolve such problems, reliability metrics are also built.

In this context, the advancement from SARIMA is to set up reliability metric, i.e. m is considered as an issue of the multiple factors linear combination. These factors are involved with two major constraints namely

- Increases in the defined number cost will naturally be compensated by a benefit in the reliability of the service provided by dropping the critical entropy on the generated clusters.
- The output of the cluster can model a frequent pattern.

Such aspects can be expressed in equation 3.

$$m(s,R) = (nsarima(s,R) - f(s,R)^2) + \left(q(s,R) - \hat{\sigma}(s,R)\right), \text{s } \epsilon \text{ S, R } \epsilon \text{ L} \quad (3)$$

Where,

$\text{nsarima}(\text{s}, \text{R})$ = the normalized value of SARIMA

According to equation 2, the initial term describes the number of clusters, while the maximum value of SARIMA is gained by the timeliness of the effective timetable specified for such partitioning. Simultaneously, when the requirement is profited by an increase with the scheduled number for increasing such reliability must be achieved. Therefore, by the number of schedules and the related cost of reducing its interpretability also the model requirement gets a trade-off between potential benefits. Now compute a consistent number of S clusters based on the available metric computation for all pairs (R, s). Let η(R) indicate the normalized number of Route R journeys. A weighted average of s ϵ S has been utilized to describe the consistent number of clusters K. Also express S ϵ N as shown in equation 4.

$$\left[\sum_{R\in L}\sum_{\text{s}\,\epsilon\,\text{S}}\frac{m(s,R)^2\times s\times \eta(\text{R})}{\psi}\Big/\sum_{R\in L}\eta(\text{R}), \psi = \sum_{\text{s}\,\epsilon\,\text{S}}(s,R)^2\right] \quad (4)$$

In this analysis, the dataset is collected from a huge Swedish urban bus operator. This completed dataset used data from four RA1 / RA2 / RB1 / RB2 higher-frequency routes (maximum scheduled headway of 15 min between 7:00 and 20:00) from two RA and RB bus lines. Line RA connects suburban areas with a PT Centre and the main shopping areas as well. Likewise, the RB ties the south region of the city to the middle of the city, which is crossed by a PT Centre, medical Centre, and a transportation Centre. This analysis covers half of the year 2019, which is seen between August 2019 and January 2020. For each time, two schedules are defined such as workdays, weekends, and holidays. A trip was carried out as pre-processing by eliminating trips with much more than 80% of the missing travel times for links. Conversely, the interpolation approaches are indicated based on remaining samples of performed data for imputation. Hence, 98% percentile can be used for pruning the dwell times to eliminate inaccurate measurements and APC data can be obtained and utilized.

Result and Discussion

The analysis was carried out using the python and DBSCAN performs the model-based clustering, also SARIMA implementation of sklearn which import the DBSCAN libraries from the Sklearn package cluster. The normality of calibration vectors was guaranteed by this SARIMA model, a transformation was carried out to create the transformed data roughly Gaussian and maintain the variance of the time series. However there are many families of transformations that can be used for such a function, the transformation of Box-Cox power has been established in different fields, namely APC and ALV. There are three parameters for this framework: K, φ and γ also the values were set to $2 \leq s \leq 7$, $\forall s \epsilon S$, 0.25 and 0.4. On the other hand, φ was selected from three possible values such as 0.4, 0.5, 0.6 using an iterative parameter performed on a small subset of the training data. The value of γ can be set only for the duration of at least four weeks. A novel scheduled plan process was suggested as detailed by applying the proposed methodology to the existing dataset. The effect on the activities of the agency in terms of schedule consistency has been evaluated through a simulation process. The results of this research provide an overview of the subsequent dataset for the complete per path which consists of

- Number of Trips (NT)
- Daily Trips (DT)
- Round Trip Time (RTT)
- Number of Stops
- Load (total number of passenger on board)

Any adjustment in the coverage of the schedule would result in two scenarios:

(i) A group of RB days would move from one to other coverage for those already in place.

(ii) An entirely new schedule will take place.

Let A and D are the two classes of days with dissimilar covers and, therefore, assigned separate timetables where RB is RA are illustrated in table 1. The aim is to test whether RB will benefit from using the same D timetable instead of its original one for the time. The timetable has been assigned to RB that may change from one in location A to the one used in location D.

Table 1 Statistics as per Route RA and RB

Bus routes	NT	No. of stops	DT	RTT in Sec	Loads
RA1	15295	31	126±36	3012±05	99±50
RA2	15160	31	124±17	2845±52	97±32
RB1	15250	26	115±21	2617±27	81±34
RB2	15345	26	118±39	2689±45	80±49

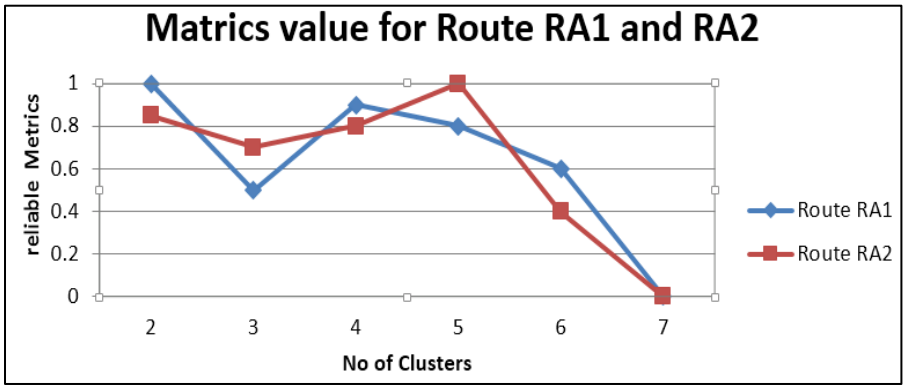

Figure 3 DBSCAN cluster computing with quality metrics for route RA and s ϵ S = {2,3, …,7}

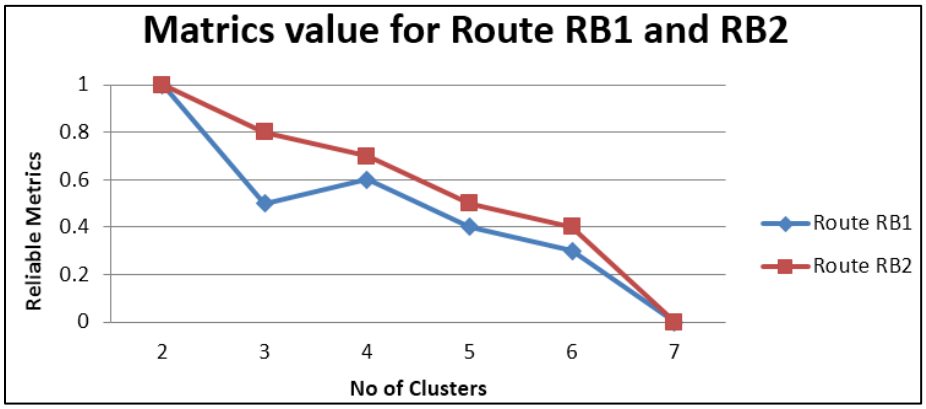

Figure 4 DBSCAN cluster computing with quality metrics for route RB and s ϵ S = {2, 3, …, 7}

Usually, this architecture runs in linear time, in which the 15 k trips were processed in ~550 sec by a single-core processor. Figure 3 shows the measured ad-hoc metric values intended to measure the consistency of the dividing given by each k value. These concepts resulted in S = 3 becoming consensual. The term fs, R2 are seen in Figure 4 as there is a strong trend to decrease the calculated score with the rise of s. However, as the charts may empirically indicate, the proposed weighted voting scheme ends up finding a majority about S = 3 and not 2.

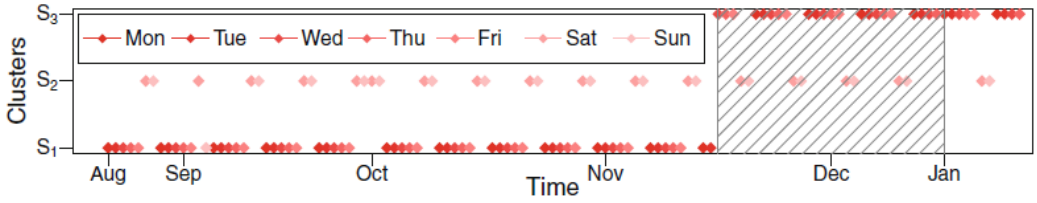

Figure 5 Automated selection of consistent clustering results as S = 3

The original data time table is allocated, simulation is needed that will first link travel time and it need to assign RB a schedule that will shift from the A to D. This study noted that the shift in coverage indicated by the highlighted area is seen in Figure 5 on the working days from the summer to winter season. It varies substantially by indicating that the winter timetable from one in place should be in place of four weeks longer than it is from mid-November to mid-December. However, as a research case study, the affected period was used to perform as an impact study of simulation-based data. Therefore, the findings of the clustering obtained specifically outline the high potential benefits of such a transition in the SARIMA model. Additionally, these benefits are mostly statistical limits that can be biased by the various restrictions of regular PT operations and also by the oversimplification of the measurement based on dwelling time. To assess the exact effect of the proposed changes that can be carried out by accurate metrics and compared to the current ARIMA model for measuring the MAE and RMSE performance of the proposed model, the on-field implementation of new coverage was therefore mandatory.

Based on the weighting and normalization required the link travel times and the dwell times produced by such a timetable from the obtainable AVL and APC data, the available original data becomes simulated. The OTP is a test of transport providers' determination to be on schedule. There are timetables for nearly all transit networks, specifying when vehicles are to arrive at planned stations.

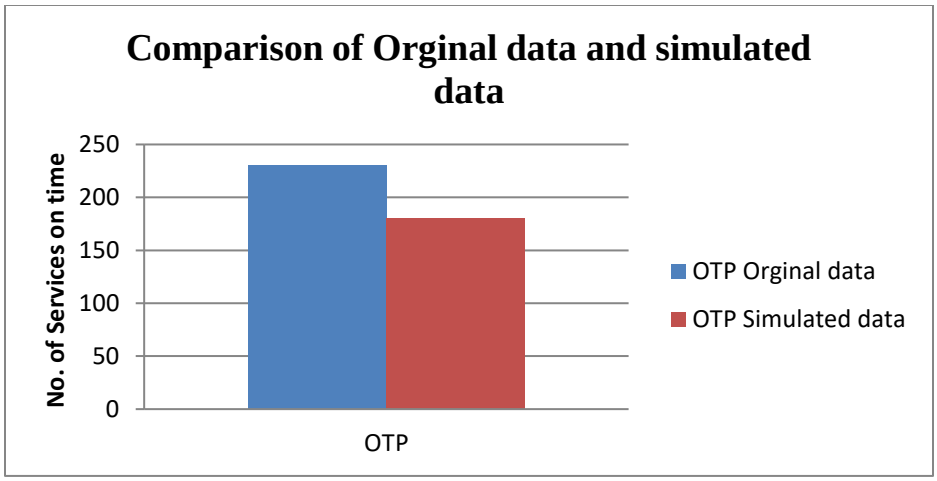

Figure 5 Automated selection of consistent clustering results as S = 3

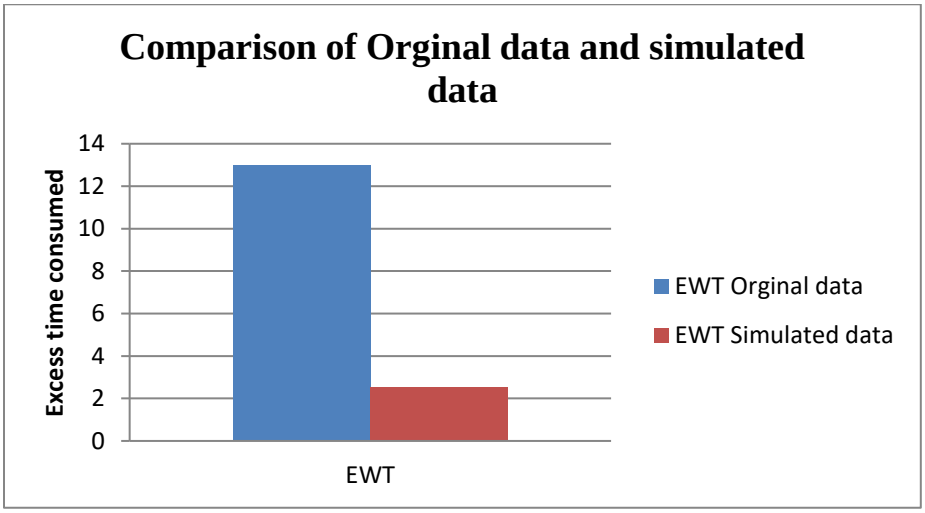

Figure 6 Automated selection of consistent clustering results as S = 3

Figure 6 has illustrated the AVL associated parameter in which original data which is considered before processing of model and after processing of model based on OTP services which get illustrated an exact result of on-time services of PT. Similarly, figure 7 has illustrated the Excess Waiting Time of the passenger in their stops. Moreover, the dwell time of the passenger identification accuracy of proposed SARIMA is compared with existing ARIMA is illustrated in table 2 that MAE of ARIMA is comparatively higher, and similarly RMSE value is also higher than proposed SARIMA.

Table 2 Model accuracy performance

Parameter	ARIMA	SARIMA
MAE	14.82	9.36
RMSE	23.57	13.71

Conclusion

This paper introduced a novel combination technique as a hybrid ML method for improving the PT network schedule coverage which is exclusively on data from AVL and APC. The overall aim is to increase the efficiency of PT and subsequently with their ridership and cost-effectiveness. The major contribution of this research is about passenger demand identification with effective metrics. The metrics of proposed DBSCAN and SARIMA has been evaluated are considerably better than GMM with ARIMA clustering to select the best schedule number. The selection of schedule number plays the main role in identifying the better demand of PT for the passenger. The corresponding routes are implemented based on adequacy, interpretability, and reliability which can be illustrated by sequence mining and probabilistic reason. In areas with lower passenger flow, the proposed model shows better advantages when measuring peak passenger flow and is additionally adaptable to changes in bus passenger flow.

References

Moreira-Matias, L., Mendes-Moreira, J., de Sousa, J.F., & Gama, J. (2015). Improving mass transit operations by using AVL-based systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *16*(4), 1636-1653.

Pelletier, M.P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*(4), 557-568.

Utsunomiya, M., Attanucci, J., & Wilson, N. (2006). Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation research record*, *1971*(1), 118-126.

Munizaga, M.A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, *24*, 9-18.

Tao, S., Rohde, D., & Corcoran, J. (2014). Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. *Journal of Transport Geography*, *41*, 21-36.

Tao, S., Corcoran, J., Hickman, M., & Stimson, R. (2016). The influence of weather on local geographical patterns of bus usage. *Journal of transport geography*, *54*, 66-80.

Arana, P., Cabezudo, S., & Peñalba, M. (2014). Influence of weather conditions on transit ridership: A statistical study using data from Smartcards. *Transportation research part A: policy and practice*, *59*, 1-12.

Hu, W., Feng, Z., Chen, Z., Harkes, J., Pillai, P., & Satyanarayanan, M. (2017). Live synthesis of vehicle-sourced data over 4G LTE. In *Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems*, 161-170.

Bao, W., Yuan, D., Yang, Z., Wang, S., Zhou, B., Adams, S., & Zomaya, A. (2018). sFog: Seamless fog computing environment for mobile IoT applications. In *Proceedings of the 21st ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 127-136.

Moura, D.L., Aquino, A.L., & Loureiro, A.A. (2019). Towards Data VSN Offloading in VANETs Integrated into the Cellular Network. In *Proceedings of the 22nd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 235-239.

Nya, N., & Baynat, B. (2017). Performance model for 4G/5G heterogeneous networks with different classes of users. *In Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems*, 171-178.

Olivieri, B., & Endler, M. (2017). An algorithm for aerial data collection from wireless sensors networks by groups of UAVs. *In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 967-972.

Aljeri, N., & Boukerche, A. (2019). Movement prediction models for vehicular networks: an empirical analysis. *Wireless Networks*, *25*(4), 1505-1518.

Aljeri, N., & Boukerche, A. (2019). A Probabilistic Neural Network-Based Road Side Unit Prediction Scheme for Autonomous Driving. *In ICC 2019-2019 IEEE International Conference on Communications (ICC)*, 1-6.

Zhao, J., Rahbee, A., & Wilson, N.H. (2007). Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering*, *22*(5), 376-387.

Chapleau, R., Trépanier, M., & Chu, K.K. (2008). The ultimate survey for transit planning: Complete information with smart card data and GIS. In *Proceedings of the 8th international conference on survey methods in transport: Harmonisation and data comparability*, 25-31.

Pelletier, M.P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*(4), 557-568.

Yu, M., Zhang, D., Cheng, Y., & Wang, M. (2011). An RFID electronic tag based automatic vehicle identification system for traffic IOT applications. *In Chinese Control and Decision Conference (CCDC)*, 4192-4197.

Cheng, X., Hu, X., Yang, L., Husain, I., Inoue, K., Krein, P., & Wang, F.Y. (2014). Electrified vehicles and the smart grid: The ITS perspective. *IEEE Transactions on Intelligent Transportation Systems*, *15*(4), 1388-1404.

Williams, B.M., & Hoel, L.A. (2003). Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of transportation engineering*, *129*(6), 664-672.

Tan, M.C., Wong, S.C., Xu, J.M., Guan, Z.R., & Zhang, P. (2009). An aggregation approach to short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, *10*(1), 60-69.

Clark, S. (2003). Traffic prediction using multivariate nonparametric regression. *Journal of transportation engineering*, *129*(2), 161-168.

Tang, Y.F., Lam, W.H., & Ng, P.L. (2003). Comparison of four modeling techniques for short-term AADT forecasting in Hong Kong. *Journal of Transportation Engineering*, *129*(3), 271-277.

Zhang, H.M. (2000). Recursive prediction of traffic conditions with neural network models. *Journal of Transportation Engineering*, *126*(6), 472-481.

Ma, X., Wu, Y.J., Wang, Y., Chen, F., & Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, *36*, 1-12.

Agard, B., Morency, C., & Trépanier, M. (2006). Mining public transport user behaviour from smart card data. *In: The 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)*, *39*(3), 399-404.

Morency, C., Trépanier, M., & Agard, B. (2006). Analysing the variability of transit users behaviour with smart card data. *In IEEE Intelligent Transportation Systems Conference*, 44-49.