International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

Performance of Public Transport Appraisal using Machine Learning

R. Thiagarajan [1], Dr. S. Prakash kumar[2]
[1]Ph,D Research Scholar/PG and Research Department of Computer Science
Maruthu Pandiyar College
Vallam Post,Thanjavur-613 403.
(Affiliated to Bharathidasan University,Tiruchirappalli,Tamilnadu,India)
Email id: thiagarajan.nvr@gmail.com

[2]Assistant Professor/ PG and Research Department of Computer Science
Maruthu Pandiyar College
Vallam Post,Thanjavur-613 403.
(Affiliated to Bharathidasan University,Tiruchirappalli,Tamilnadu,India)
Email id: drsp1974@gmail.com

Abstract: Public passenger transport holds immense significance in the overall transportation system. Forecasting the movement of public transport has emerged as a crucial problem in transport planning due to its practical implications. Recently, there has been a lot of significant attention in Intelligent Transportation Systems (ITS), introducing various advancements and innovative applications to develop conditions for public transit that are safer, more effective, and fun. To fully leverage the potential of ITS applications and deal with road situations proactively, it becomes crucial to have a reliable method for predicting traffic flow. This opens up opportunities for ITS applications to anticipate and address potential challenges in advance. Enhancing the efficient functioning of Public Transport (PT) networks is a primary objective for urban area authorities, and the proliferation of location and communication devices has led to an abundance of operational data. Applying appropriate Machine Learning (ML) methods can help identify patterns in the data to improve the Schedule Plan. This research focuses on heterogeneous information that influences the prediction value, aiming to predict the required transport demand for specific routes and the arrival time of public transport. Utilizing DBSCAN clustering with SARIMA Algorithm, real-time passenger demand forecasting is extensively promoted to enhance dynamic bus scheduling and management. Furthermore, this paper compares the accuracy of the proposed Prophet Model with traditional time series models like ARIMA and SARIMA. The aim is to provide precise and robust passenger demand predictions, enabling more effective planning and management of PT services.

Keywords*: Public Transport, Automatic Passenger Counting, Automatic Vehicle Location, Dwell times*

Introduction: The dependability of public transport (PT) is a significant issue in modern metropolitan cities and achieving a balance between resource usage and revenue while delivering high-quality services requires effective operational planning. Modern technologies like Radio-frequency Identification (RFID) readers, Global Positioning System (GPS) antennas and 3G connection devices have recently been added by major PT operators to their fleets. These technologies enable the collection of real-time data, including Automatic Passenger Counting (APC) and Automatic Vehicle Location (AVL) which is transmitted to a central server.

The ITS have gained prominence since their adoption at the World Congress in Paris, 1994. In order to inform travellers and improve the efficiency and safety of public transportation networks, ITS makes use of electronics and communication technology, computer. The capacity of ITS to support efficient and secure road transport movement is one of the technology's key benefits.

According to recent studies, smart card data has the potential to be a useful tool for managing and planning transportation. Every passenger's smart card stores important commute information, including trip dates, hours, places of origin, destinations and travel distances. Leveraging this demand information from smart cards can enable transport authorities to optimize the entire transport network. Subsequently, smart card data has only been used sparingly for such purposes in research so far.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

In order to successfully plan, administer, and assess the transport service, public transportation companies must conduct a verification of the number of passengers. While traditional manual methods for passenger counts are costly and produce small samples, the advent of Automatic Data Collection (ADC), including APC, Automatic Fare Collection (AFC) systems and AVL systems , has given more accurate and comprehensive data.

In the context of public transport, this literature review reveals that there is a growing focus on exploiting ITS data for strategic planning, considering issues of sustainability and efficiency. AFC data is a more useful tool than conventional survey gathering techniques for examining the factors behind travel behaviour because of its high spatio-temporal resolution.

IoT devices like smart cards in public transportation systems have unique serial numbers that record transaction details. Combining smart card data with GPS tracing from On-Board Units (OBUs) can help estimate crowding in buses. Parametric techniques like ARIMA have been commonly used for transportation demand forecasting, but their linear assumptions limit their ability to capture non-linear relationships. Non-parametric methods such as Neural Networks (NN) have gained popularity for their adaptability, non-linearity, and function mapping capability.

Clustering methods, particularly k-means and DBSCAN, have been applied to ticketing data to study individual travel regularity and retrieve recurrent travel patterns, respectively. These methods offer valuable insights into passenger behavior and patterns in public transportation systems.

The adoption of ITS technologies and the intelligent analysis of data hold great potential for improving the reliability and efficiency of public transport systems, leading to enhanced services and a better travel experience for commuters.

Methodology: The research methods presented in this work involves a comprehensive approach to automatically identify passenger demand and vehicle running times by leveraging data from both APC and AVL systems. Optimising the quantity of scheduled vehicles and the total number of people they accommodate each day is the main goal. The suggested approach is only effective on days when route travels perform similarly with regard to round-trip durations over the day. In such cases, these trips are assigned to the same schedule with various routes. The flowchart illustrating the working process of this method is depicted in Figure 1.

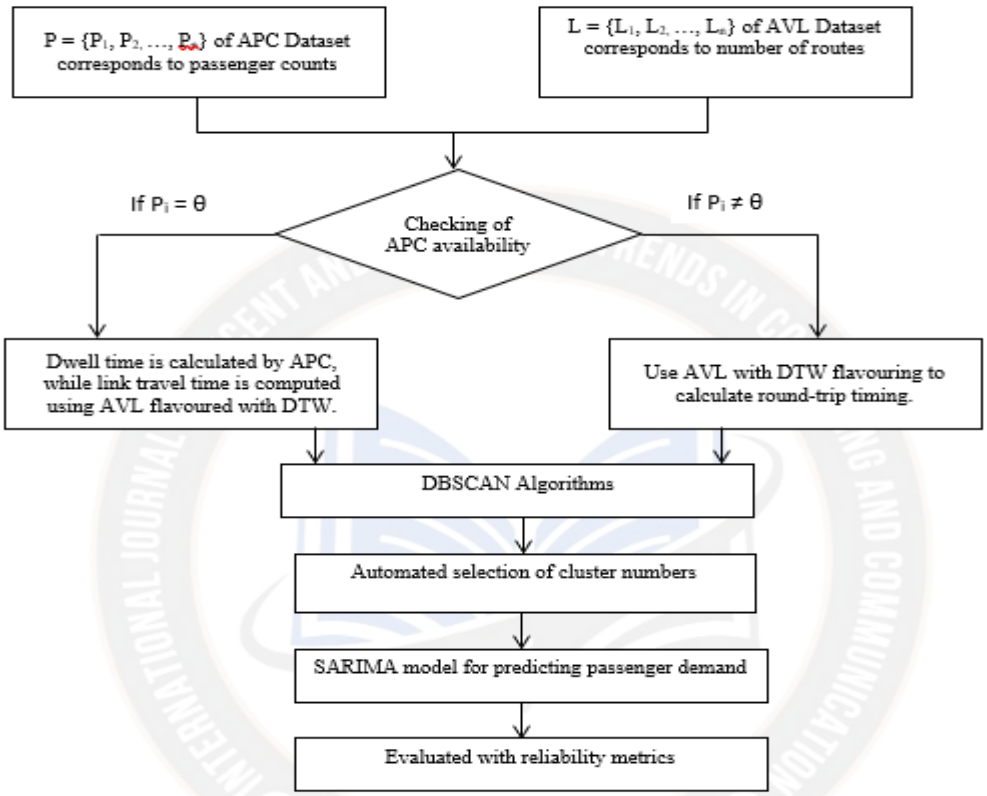

Figure 1 DBSCAN cluster proposal model with SARIMA

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

Let the route of interest be denoted as L = {R1, R2, R3, ..., Rn}. Starting with the original AVL or APC dataset, the approach extracts the operating durations and boarding or alighting information for each route R ϵ L. Based on the available APC data, everyday assessments are created, and if data is missing for a specific route, a suggested procedure is employed.

The next steps involve creating a distance matrix among the days and clustering them using Euclidean flavoured Dynamic Time Warping (DTW) and DBSCAN Model. Unlike previous approaches, for permitted number of schedules S ⊂ N that is user-defined, clustering is carried out, rather than a single predefined value of s. The process is repeated for all routes, and the best number of schedules s ϵ S is selected based on a two-stage evaluation metric.

For each trip (Pi ≠ θ), the dwell periods pause and connect travel times obtained from AVL data are added to calculate the round-trip time. When using multilayer APC data, the calculated journey time can be slightly adjusted for fluctuating demand. Modelling is done by decomposing the dwell duration at every location into many components. Overall, this methodology aims to optimize schedule coverage for the routes of interest by considering running times, boardings/alighting, and demand patterns derived from the available data.

$$\delta_{k,j} = max\left(\alpha \times a_{k,j} \times \beta \times b_{k,j}\right) + alct \quad (1)$$

Alighting and boarding times for each passengers are represented by the constants α and β respectively, while the time allotted for activities like door opening and closing at each stop is indicated by alct. Additionally, the numbers a_(k,j) and b_(k,j) show how many passengers boarded and debarked during journey k at stop j, respectively. k∈{1,2,…,t} and j∈{1,2,…,c}. are used here. The values of α, β, and alct are estimated using a linear regression approach using the provided values for dwell times (AVL) δ_(k,j), a_(k,j), and b_(k,j) (APC).

The DBSCAN algorithm characterises clusters as dense regions separated into lesser dense sections. Its two global parameters are the smallest number of points MinPts and the greatest density reach distance ε. A point is referred to as a "core point" (i_c) if it has at least MinPts (density) in a radius of ε, as indicated in equation 2.

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

If a point is inside ε the area ε of a core point i_c but has less points than MinPts, it is referred to as a "border point" (i_b). A point is referred to as a "noise point" if it no longer to form a border or a core point. The core points i_c and their corresponding border points i_b is merged to define a cluster when they are close together (within distance ε). Separate applications of the technique are made for mining temporal and spatial sequences to provide a more thorough explanation of DBSCAN. Using a two-level DBSCAN technique, it constructs regular Origin-destinations depending initially on past alighting stops and then on boarding stations. This distinct application improves the resilience of the clustering technique overall and yields findings that are helpful for later passenger segmentation. The sequence of the two levels does not influence the results.

Selection of Schedules through Automated Process and selecting the optimal number of clusters is a challenging task in data analysis. To address this, SARIMA is utilized to compute an entropy-based probabilistic score aiming to maximize the optimal s from a set of values, i.e., S, through maximising the entropy between data from various clusters while minimising the entropy across samples from the same cluster. However, taking into account the limitations of each application area, this optimisation problem might not produce an appropriate solution for real-life scenarios. As a result, dependability measurements are frequently used to solve these problems. In this context, establishing a dependability metric where m is viewed as a problem of the linear combination of numerous factors is an improvement over SARIMA. Two significant limitations are linked to these factors:

1. A gain in the offered service's punctuality and a decrease in the entropy on the clusters' created must be used to offset a hike in the specified programmed number cost.
2. The output of the cluster must represent a common sequence (for example, pairing Saturdays and Sundays for a period of five months of the year).

Equation 3 provides an expression for these variables.

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

The SARIMA standardised value is denoted by the variable "nsarima(s,R)".

Greater values of nsarima in equation 2 denote an improvement in the accuracy of an appropriate timetable specified for such splitting. The total count of clusters is indicated by the initial term in equation 2. To maximise punctuality, more schedules can be created, but this ought to be done only when it is absolutely essential. In order to meet the model's requirements, a trade-off must be made between the potential benefits of adding extra schedules being added

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

and the associated cost of making the model harder to interpret. We may now determine the standardised number of clusters, indicated by S, using the metric that has been calculated for all pairings (R, s). Let R's normalised route count be represented by η(R). As indicated in equation 4, K denotes the number of clusters based on the weighted average of s ϵ S.

$$\lceil \sum_{R \in L} \quad \sum_{s \, \epsilon \, S} \quad \frac{\frac{m(s,R)^2 \times s \times \eta(R)}{\psi}}{\sum_{R \in L} \quad \eta(R),\psi} = \sum_{s \, \epsilon \, S} \quad (s,R)^2 \rceil \tag{4}$$

The dataset for this study was gathered from a significant Swedish city bus company. Four high-frequency routes are represented in the dataset, namely RA1, RA2, RB1, and RB2, which belong to two bus lines, RA and RB. Line RA serves residential areas and connects them to a significant shopping place and a centre for public transport. The city's southern regions are connected to the city centre by line RB, which also passes across a large hospital, a logistics area, and a transportation hub.

The data used in this research covers the period between August 2019 and January 2020, representing the first half of the year. For each time, two schedules: one for workdays and the other for weekends and holidays were established. In order to remove trips having over eighty percent of the missing link journey times, a trip trimming technique was used before starting the study. For the remaining samples, data imputation was performed using an interpolation procedure. Additionally, dwell durations were reduced based on the 98th percentile to eliminate erroneous measurements, and the Automatic Passenger Counting (APC) data was utilized as obtained.

Result and Discussion

The experiments in this research were conducted using Python, utilizing the DBSCAN and SARIMA implementations from the Scikit-learn package. The SARIMA model ensures the normality of the calibration vectors by performing a transformation to stabilise the time series' variance by bringing the processed data closer to a Gaussian distribution. Among various transformation methods, the Box-Cox power transformation was chosen and widely used in fields such as APC and AVL. The parameters K, γ, and φ were set to values that provided satisfactory results: $2 \leq s \leq 7$ for K, 0.25 for γ, and 0.4 for φ. The values of γ and φ were determined by iterative parameter tuning using training data subset.

The suggested approach was applied to the provided dataset, yielding a unique schedule planning process, which was evaluated in terms of its impact on schedule reliability through a simulation procedure. The study provides a high-level summary of the collected data, including specifics for each route, such as the number of trips (NT), round-trip duration (RTT), daily trips (DT), number of stops, and passenger capacity like total on-board passengers. It is possible that any change to the schedule's coverage will result in either of the following outcomes:

1. RB changes from one form of coverage to another during a transition period amongst those currently in place,
2. Adopting an entirely new schedule.

Assuming A and D are different days with different schedules and coverage, and RB be a subset of RA to assess the impact of such changes. In order to migrate from coverage in site A to coverage in location D, it is necessary to determine whether the time period RB would gain from adopting an identical schedule as D rather than the initial one.

Table 1 Statistics analysis for Route RA and RB

Bus routes	**No. of stops**	**NT**	**RTT in Sec**	**DT**	**Loads**
RA - I	30	15290	3010±04	125±35	98±51
RB - I	27	15255	2616±26	116±22	80±33
RA - II	30	15155	2842±51	125±16	98±31
RB - II	27	15350	2688±47	119±40	81±48

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

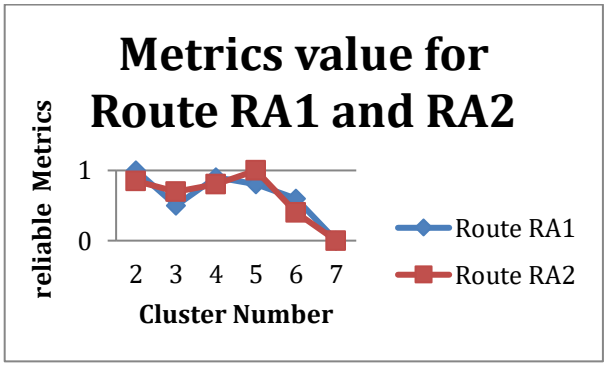

Figure 3 DBSCAN cluster computing performance metrics for route RA and s ϵ S = {2,3, …,7}

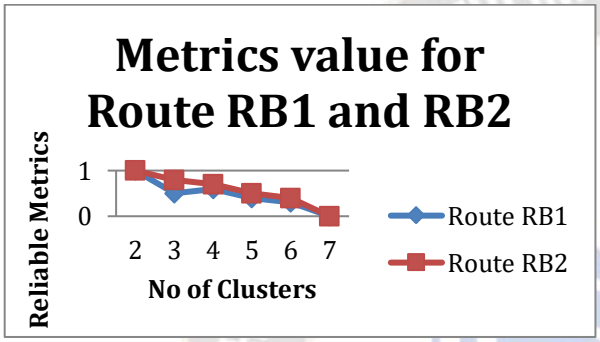

Figure 4 DBSCAN cluster computing performance metrics for route RB and s ϵ S = {2, 3, …, 7}

The suggested framework works well in linear time, processing the 15,000 trips in our example study in about 550 seconds on a single-core CPU. According to Figure 3, the ad-hoc metric used to evaluate the efficiency of splitting is calculated for a given value of k. The results showed a consensus around S = 3 as the most appropriate number of schedules. Figure 4 depicts the negative influence of the parameter f(s,R)^2, demonstrating a distinct pattern of the calculated value decreasing with increasing s value. Despite the empirical suggestion from the charts, the weighted voting schema ultimately converges to a consensus value of S = 3 instead of 2.

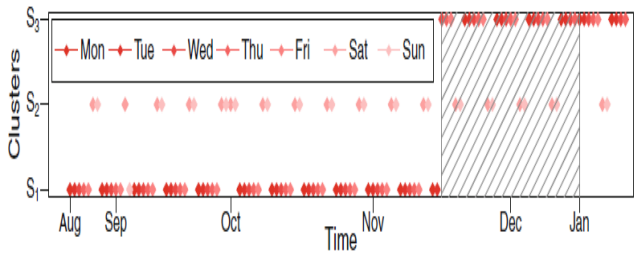

Figure 5 Constant clustering with Auto-selection results as S = 3

After assigning the initial schedule data, a simulation is run to link trip time and give RB a new timetable, transitioning from the one in place in A to the one used in D. The research observed a change in coverage from Summer to Winter on workdays, as depicted in Figure 5. The suggested Winter schedule is notably different from the existing one, suggesting that it should be implemented four weeks sooner than the existing schedule, particularly from the middle of November to the middle of December. For the purpose of conducting an impact analysis using simulation-based data, this time period was selected as the research subject of the study.

The SARIMA model's clustering findings unequivocally show the considerable potential gains of putting this alteration into practise. The excessive simplification of dwell duration calculation and different limits in daily PT operations, however, may have an impact on these benefits, which are essentially theoretical in nature. The proposed model's RMSE and MAE performance must be evaluated using valid metrics in comparison to the existing ARIMA model in order to assess the specific implications of the proposed changes.

The available original data is then simulated based on the necessary normalisation and weighting, connecting the travel and dwell times generated from the timetable provided by AVL/APC data. OTP (On-Time Performance) is used as a measure of the ability of transport services to be on time, with timetables describing when vehicles are scheduled to arrive at specific stops, a common feature in almost all transportation systems.

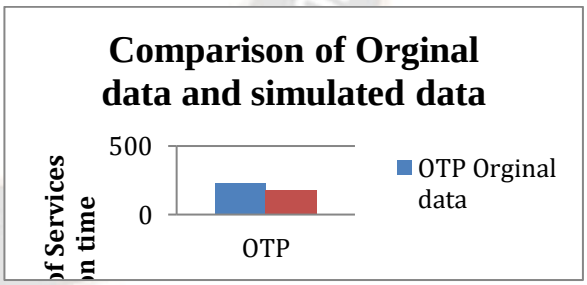

Figure 6 No of service on time with constant clustering in Auto-selection results as S = 3

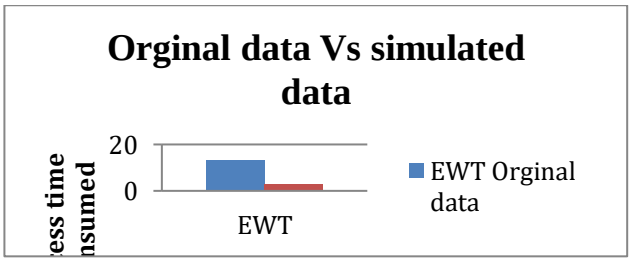

Figure 7 Excess time consumed with constant clustering in Auto-selection results as S = 3

The on-time accomplishment of public transport (PT) is accurately represented in Figure 6 by the AVL-associated parameters, which show the original data taken into account before to model process and the data after model processing depending on OTP operations. Figure 7 also shows how long passengers end up waiting at their stops.

Table 2 compares the suggested SARIMA model and the current ARIMA model for dwell duration of passenger recognition. It is clear that ARIMA's Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values are significantly larger than those of the suggested SARIMA model.

Table 2 Performance of Model accuracy

Error Rate Parameter	ARIMA	SARIMA
RMSE	23.55	13.70
MAE	14.81	9.38

Conclusion

In this paper, a unique method for improving schedule coverage for AVL or APC on PT networks has been presented. In order to increase ridership and cost effectiveness, PT reliability must be enhanced. Our key contribution lies in utilizing reliability metrics to enhance clustering using DBSCAN and SARIMA, enabling the selection of the optimal number of schedules with a focus on sequence mining and probabilistic reasoning to model sufficiency, interpretability, and reliability.

To evaluate the accuracy of our proposed method, we compared it with the existing ARIMA model. However, the new coverage must be deployed on-field in order to accurately analyse the effect of the recommended modifications and evaluate their efficiency, utilizing reliable metrics of PT operations. Thus, the proposed method determines the specific locations with minimal passenger movement, as well as find out benefits by estimating maximum flow of passengers and exhibits improved responsiveness to fluctuations in passenger movement of buses.

References

[1] Moreira-Matias, L., Mendes-Moreira, J., Freire de Sousa, J., Gama, J.: Improving mass transit operations by using avl-based systems: a survey. IEEE Trans. Intell. Transp. Syst. 16(4), 1636–1653 (2015).

[2] 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.

[3] Utsunomiya, M., Attanucci, J., Wilson, N., 2006. Potential uses of transit smart card registration and transaction data to improve transit planning. Transp. Res. Rec. J. Transp. Res. Board 1971, 119–126.

[4] Munzinga, M. A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago de Chile, Transportation Research Part C 24, 9-18.

[5] 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. J. Transp. Geogr. 41, 21–36.

[6] 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.

[7] Arana, P., Cabezudo, S., Peñalba, M., 2014. Influence of weather conditions on transit ridership: A statical study using data from Smartcards. Transportation Research Part A 59, 1-12.

[8] W. Hu, Z. Feng, Z. Chen, J. Harkes, P. Pillai, M. Satyanarayanan, Live synthesis of vehicle-sourced data over 4G LTE, in: ACM MSWIM, 2017, pp. 161–170.

[9] W. Bao, D. Yuan, Z. Yang, S. Wang, B. Zhou, S. Adams, A. Zomaya, SFog: Seamless fog computing environment for mobile IoT applications, in: ACM MSWIM, 2018, pp. 127–136.

[10] D.L.L. Moura, A.L.L. Aquino, A.A.F. Loureiro, Towards data VSN offloading in VANETs integrated into the cellular network, in: ACM MSWIM, in: MSWIM '19, 2019, pp. 235–239.

[11] N. Nya, B. Baynat, Performance model for 4G/5G heterogeneous networks with different classes of users, in: ACM MSWiM, 2017, pp. 171–178.

[12] B. Olivieri, M. Endler, DADCA: An efficient distributed algorithm for aerial data collection from wireless sensors networks by UAVs, in: ACM MSWiM, 2017, pp. 129–136.

[13] N. Aljeri, A. Boukerche, Movement prediction models for vehicular networks: An empirical analysis, Wirel. Netw. 25 (4) (2019) 1505–1518.

[14] N. Aljeri, A. Boukerche, A probabilistic neural network-based road side unit prediction scheme for autonomous driving, in: ICC 2019 - 2019 IEEE International Conference on Communications (ICC), 2019, pp. 1–6.

[15] Zhao, J., Rahbee, A., Wilson, N.H.: Estimating a rail passenger trip origin-destination matrix using automatic

data collection systems. Comput.-Aided Civ. Infrastruct. Eng. 22(5), 376–387 (2007)

[16] 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 (pp. 25–31).

[17] 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.

[18] M. Yu, D. Zhang, Y. Cheng, and M. Wang, "An RFID electronic tag based automatic vehicle identification system for traffic iot applications," in Proc. Chin. Control Decision Conf. (CCDC), May 2011, pp. 4192–4197.

[19] X. Cheng et al., "Electrified vehicles and the smart grid: The ITS perspective," IEEE Trans. Intell. Transp. Syst., vol. 15, no. 4, pp. 1388–1404, Aug. 2014.

[20] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results," J. Transp. Eng., vol. 129, no. 6, pp. 664–672, Nov. 2003.

[21] M. C. Tan, S. C. Wong, J. M. Xu, Z. R. Guan, and P. Zhang, "An aggregation approach to short-term traffic flow prediction," IEEE Trans. Intell. Transp. Syst., vol. 10, no. 1, pp. 60–69, Mar. 2009.

[22] S. Clark, "Traffic prediction using multivariate nonparametric regression," J. Transp. Eng., vol. 129, no. 2, pp. 161–168, Mar. 2003.

[23] Y. Tang, W. H. Lam, and P. L. Ng, "Comparison of four modeling techniques for short-term aadt forecasting in Hong Kong," J. Transp. Eng., vol. 129, no. 3, pp. 271–277, May 2003.

[24] H. Zhang, "Recursive prediction of traffic conditions with neural network models," J. Transp. Eng., vol. 126, no. 6, pp. 472–481, Dec. 2000.

[25] Ma, Xl, Wu, Y.J., Wang, Yh, Chen, F., Liu, Jf: Mining smart card data for transit riders' travel patterns. Transp. Res. C: Emerg. Technol., 36(0), 1–12 (2013).

[26] Agard, B., Morency,C.,Trépanier, M.:Mining public transport user behaviour from smart card data. In: The 12th IFAC Symposium on Information Control Problems inManufacturing (INCOM), pp. 17–19 (2006).

[27] Morency, C., Trépanier, M., Agard, B.: Analysing the variability of transit users behaviour with smart card data. In: The Ninth International IEEE Conference on Intelligent Transportation Systems, Toronto, Canada, September (2006).