Australasian Transport Research Forum 2024 Proceedings
27-29 November, Melbourne, Australia
Publication website: https://australasiantransportresearchforum.org.au/

Train carriage passenger count estimation from ticketing data using a hybrid framework combining heuristics and machine learning models: Proof of Concept

Saeed Mohammadian[1], Zuduo Zheng[1], Tristan Miles[2]

[1]School of Civil Engineering, School of Civil Engineering, The University of Queensland, St. Lucia 4072, Brisbane, Australia
[2] Department of Transport and Main Roads

Email for correspondence (presenting author): Zuduo.Zheng@uq.com.au

1 Introduction

Accurate train passenger count estimations are critical for effective operational planning, optimizing resource allocation, improving service efficiency, and enhancing the overall passenger experience. Conventional methods rely on schedule-based transit assignments using large-scale historical ticketing data (Zhang et al., 2021). However, while these methods provide reliable and generally desirable accuracy at the network level, their predictive performance is less sufficient for granular, day-to-day estimations for specific trains. This is due to several factors, including reliance on historical ticketing data, which may not fully account for real-time factors such as fare evasion, mis-recorded passengers, special events, track closures, schedule changes, delays, or other anomalies (Li & Hensher, 2013; Tirachini et al., 2013). With the advent of new sensing technologies, such as passenger counters on train doors, ample opportunities are provided for accurately modelling train passenger counts (Grgurević et al., 2022; Yu et al., 2022).

The objective of this study is to investigate the feasibility of a hybrid methodological framework for train passenger count estimation in the South East Queensland passenger rail network. This framework integrates heuristic passenger count estimation using ticketing data with machine-learning models to estimate passenger counts, as observed from in-built door counter sensors in train carriages. To this end, a comprehensive dataset from New Generation Rollingstock (NGR) trains over a single day is utilized, providing detailed information about passengers' boardings and alightings in each carriage. The focus is restricted to this one-day dataset to assess the framework's feasibility before scaling to larger datasets. Several machine learning algorithms commonly used in the literature, including AdaBoost, CatBoost, Random Forest, and XGBoost, are implemented, and their performance for train-level and carriage-level passenger count estimations is evaluated.

2 Methodology

2.1 Datasets

The study utilizes three types of datasets: Automatic Vehicle Location (AVL) for detailed train movements, Automatic Fare Counting (AFC) for passenger origin-destination data via station taps, and Automatic Passenger Counting (APC) from onboard sensor sensors in trains on crowding levels and boarding/alighting.

2.1.1 The NGR trains dataset

The NGR trains in South East Queensland are multi-carriage units (see Figure 1 for a schematic map of NGR carriages), designed for passenger comfort and accessibility. These

trains offer valuable opportunities for research into passengers' travel behaviour and rail transit, as each carriage is equipped with four infrared door sensors that detect passenger movements during boarding and alighting. The dataset includes AVL data, detailing train trajectories, arrivals, departures, and dwell times, as well as APC data, which captures carriage-level boardings and alightings at each station. This data is recorded for each minute of waiting time at stations, providing real-time insights into train operations and crowding dynamics. We utilise data from 24 March 2023 on major routes across South East Queensland, excluding the Ferny Grove, Rosewood, and Sunshine Coast lines north of Nambour, as the 'ground truth' for carriage-level train passenger counts.

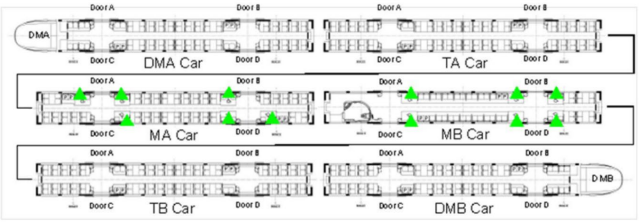

Figure 1: Schematic map of NGR trains showing carriage positioning: DMA at the head, followed by TA, MA, MB, TB, and DMB at the tail. Note: The green triangles shown in the schematic are irrelevant as per the provider's instructions and should be disregarded.

2.1.2 Station-level automatic fare counting (AFC) from passenger ticketing data

Next, we use passenger ticketing data, which records passengers' tap-in and tap-out events at stations alongside NGR observations on the same day, as input for the model. This dataset includes key details such as operation dates, boarding and alighting times and stations, ticket status, passenger count, fare, journey and trip IDs, ticket types, and passenger categories. Ticketing data provides crucial insights into public transport dynamics, including station-level crowding, origin-destination trips, and passenger flows across all lines. However, directly linking passengers to specific trains is challenging due to multiple trains on the same route. Despite this limitation, the data is essential for understanding passenger behaviour and improving transit systems.

2.2 Hybrid methodological approach combining PTM into machine learning for refining passenger count estimates

This study explores a hybrid framework that combines heuristic passenger count estimations from the Public Transport Model (PTM) using historical ticketing data with machine learning algorithms to predict train passenger counts in South East Queensland's rail network, as recorded by door-mounted passenger counters from NGR data. The following sections provide further details on PTM and the machine learning models.

2.2.1 Public transport model (PTM) for passenger-to-train assignment using ticketing data

The Public Transport Model (PTM), developed by Queensland's Department of Transport, is a planning tool used to test network changes and evaluate fare structures through schedule-based transit assignments. These assignments provide time-dependent estimates of total train boardings, alightings at each station, and overall passenger counts based on timetables and demand. However, they do not provide detailed information about carriage-level passenger distributions. In our hybrid approach, we use PTM's transit assignment as an initial estimate of total passenger counts, which is then refined with machine-learning techniques to estimate carriage-level passenger distributions.

2.2.2 Machine-learning models and procedure, and criteria for development and assessment

The proposed methodology evaluates several machine learning algorithms for regression tasks, including AdaBoost, CatBoost, Random Forest, and XGBoost, comparing their performance to one another and using Linear Regression as a benchmark. For further details on the algorithms, refer to Stanulov and Yassine (2024). The goal is to identify the optimal algorithm for accurately estimating onboard passenger counts at carriage-level using data inputs alongside PTM estimates.

To assess the models, we split the data into training and testing sets with an 80:20 ratio, tune hyperparameters for each algorithm using techniques like grid search and train the models on the training set. Performance is evaluated on the testing set using metrics such as R^2 scores, Root Mean Square Error (RMSE), and error distributions to assess the models' robustness and accuracy in estimating passenger counts.

To implement and assess the proposed methodology, we first extract and process both NGR and ticketing data, integrating it with the Public Transport Model (PTM). A cross-comparison between observed NGR data and PTM's scheduled data reveals discrepancies, prompting corrections in planned trajectories, passenger counts, and station stops. After cleaning and preprocessing, and consolidating duplicate entries, 8,516 recordings corresponding to 431 unique trips remain. Further alignment of trajectories and the removal of inconsistent data reduces the dataset to 374 unique trips, retaining 87.9% of the original data.

3 Main findings

The input features for the train-level model include variables such as arrival_time (train arrival time at a station), train_path (the unique path between consecutive junction stations), Direction (forward or reverse travel), Station (current stop), stop_sequence (stop's position in the journey), and PTM estimates for pax_onboard_PTM (total onboard passengers), transit_boardings, and transit_alightings at each station. The target output variable, pax_onboard_adj, represents the total number of passengers onboard, aggregated across all carriages, as recorded by NGR sensors. Figure 2 displays the correlation heatmap showing the relationships between the different features and target variables.

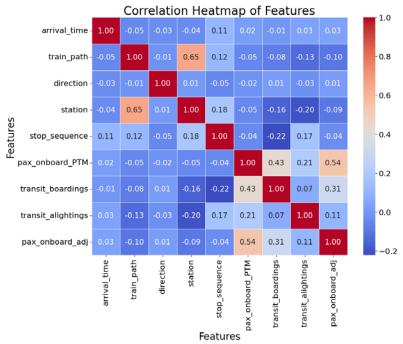

Figure 2: Correlation heatmap across between different the input variables (features) and target total train passenger counts (i.e., pax_oboard_adj)

3.1 Train-level evaluation of the hybrid framework: Assessing approach feasibility and machine learning model selection

We first test our proposed hybrid framework on train-level passenger counts across train trajectories to evaluate its accuracy and feasibility. The goal is to compare the performance of

various machine learning algorithms and select the most suitable model for extending the methodology to carriage-level estimations. To achieve this, we develop several machine learning models- Linear Regression as a baseline, along with AdaBoost, CatBoost, Random Forest, and XGBoost- by conducting hyperparameter tuning for each algorithm. Optimal model parameters are identified, and model performance is evaluated using R^2 scores and RMSE values, as shown in Figure 3.

The results demonstrate the suitability of the proposed hybrid approach, with XGBoost and Random Forest significantly outperforming other models. Random Forest slightly outperforms XGBoost in the training set, indicated by lower RMSE values. As a result, both models are selected for further evaluation in carriage-level predictions.

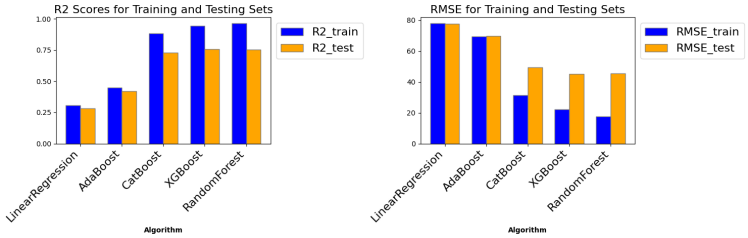

Figure 3: Comparative performance of various machine learning algorithms assessed for train-level passenger count estimations using R^2 and $RMSE$ measures over training and testing sets.

3.2 Carriage-level evaluation of the hybrid framework: Assessing approach feasibility and machine learning model selection

For the carriage-level model, we use the same input variables as in the train-level model. However, the target output variables focus on the number of passengers in individual carriages. As recorded in the NGR data, these outputs include pax_dma_onboard_adj (DMA carriage), pax_ta_onboard_adj (TA carriage), pax_ma_onboard_adj (MA carriage), pax_mb_onboard_adj (MB carriage), pax_tb_onboard_adj (TB carriage), and pax_dmb_onboard_adj (DMB carriage), all tracked by NGR sensors at each station.

Building on the findings from the train-level model, we focus on the Random Forest algorithm, with Table 1 presenting the R^2 scores and RMSE values for both the training and test sets. Overall, Random Forest provides strong carriage-level passenger count predictions, achieving higher accuracy in the training set (R^2 values around 0.95) compared to the test set (R^2 values between 0.59 and 0.71). Figure 4 shows the error distributions for Random Forest predictions across two representative carriages for both training and test sets.

Although prediction accuracy is lower in the test set, it remains acceptable, and the normal error distribution suggests this may not be primarily due to overfitting. The reduced accuracy in the test set is likely due to the limited dataset, which contains data from only one day, and the random selection of test sets. Both factors are expected to improve with a larger dataset.

We conclude by illustrating the model's performance in predicting carriage-level passenger counts across train trajectories, focusing on Train D as a representative example, as shown in Figure 5. The results show that the model performs well in estimating overall passenger counts at the carriage level along the trajectory, despite some local discrepancies at certain stations (e.g., at Kallangur station for carriages MA and MB), likely due to those stations belonging to the test set. Overall, similar performance was observed across most trains in the

dataset, supporting the model's suitability for predicting carriage-level counts and the broader application of this hybrid approach in railway network operations.

Table 1: Quantitative performance results for carriage-level estimates obtained using the random forest algorithm.

Evaluation scenarios		Data			
		Training set		Testing set	
		r^2	$RMSE$	r^2	$RMSE$
Carriage-level	dma	0.9449	4.45	0.6376	11.89
	ta	0.9555	4.01	0.6864	10.58
	ma	0.9571	4.00	0.7121	10.35
	mb	0.9592	4.14	0.7029	10.96
	tb	0.9559	3.87	0.6713	10.62
	dmb	0.9439	3.94	0.5934	10.50
Train-level	---	0.9449	15.62	0.6376	11.89

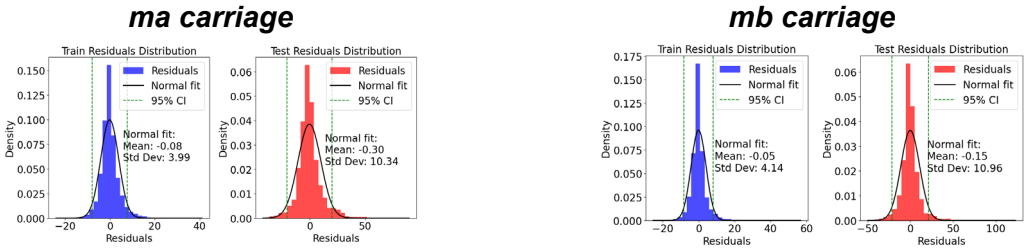

Figure 4: the error distributions for the selected machine learning algorithm (Random Forest) predictions across two representative carriages, for both the training and testing sets.

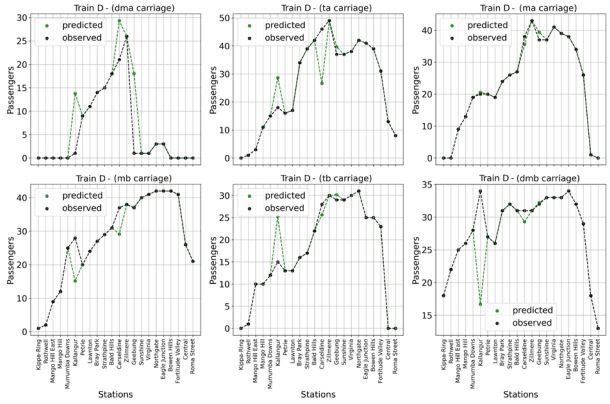

Figure 5: An example of carriage-level passenger count estimations across a train trajectory from the first to the last stop, comparing the proposed model to carriage-level NGR observations.

4 Discussion

This study focused on investigating the feasibility of a hybrid methodological framework, integrating heuristic approaches with machine-learning models, to estimate passenger counts in trains using ticketing data as input, while considering observations from door counts as ground truth using one-day data from Queensland urban rail networks. Investigations were conducted on two modelling approaches: train-level and carriage-level models. A summary of the key findings include:

(1) Heuristic estimations serve as valuable indicators of passenger counts, demonstrated by their alignment with NGR observations regarding the overall trends.
(2) The proposed hybrid methodological framework, combining heuristic approaches with machine-learning models, shows promise in accurately estimating passenger counts in trains.
(3) Among the machine-learning models considered Random Forest and XGBoost outperformed the rest for both train-level and carriage-level estimates.
(4) Carriage-level estimations on the calibration data, lead to accurate estimates of train-level passenger counts as well as total passenger counts.
(5) The model exhibited robust performance for unknown data in terms of prediction consistency and Gaussian error distributions. However, its prediction accuracy was not as strong as in the calibration scenario, likely due to the limited amount of data available from just one day.

Based on the findings in this work, future directions, which is currently under investigations, include:

(1) Conduct longitudinal studies using extended periods of data to improve model robustness and accuracy.
(2) Explore the inclusion of additional variables like day type and seasonality to gain insights into passenger behaviour and enhance prediction performance.

Acknowledgement

This research was part of an ongoing TAP project, or Transport Academic Partnership collaboration between the Department of Transport and Main Roads (TMR) and the University of Queensland. The authors would like to thank the Translink, and Policy, Planning & Investment Divisions of the Department of Transport and Main Roads for the productive collaboration, valuable feedback, access to datasets and contributions to this research. The authors would like to acknowledge the Transport Analysis Unit who are responsible for the development, maintenance, and application of the Public Transport Model.

5 Selected References

Grgurević, I., Juršić, K., & Rajič, V. (2022). Review of automatic passenger counting systems in public urban transport. 5th EAI International Conference on Management of Manufacturing Systems,

Li, Z., & Hensher, D. A. (2013). Crowding in Public Transport: A Review of Objective and Subjective Measures. *Journal of Public Transportation*, *16*(2), 107-134. https://doi.org/https://doi.org/10.5038/2375-0901.16.2.6

Stanulov, A., & Yassine, S. (2024). A comparative analysis of machine learning algorithms for the purpose of predicting Norwegian air passenger traffic. *International Journal of Mathematics, Statistics, and Computer Science*, *2*, 28-43.

Tirachini, A., Hensher, D. A., & Rose, J. M. (2013). Crowding in public transport systems: effects on users, operation and implications for the estimation of demand. *Transportation Research Part A: Policy and Practice*, *53*, 36-52.

Yu, H., Lu, J., Liu, A., Wang, B., Li, R., & Zhang, G. (2022). Real-time prediction system of train carriage load based on multi-stream fuzzy learning. *IEEE Transactions on Intelligent Transportation Systems*, *23*(9), 15155-15165.

Zhang, Q., Liu, X., Spurgeon, S., & Yu, D. (2021). A two-layer modelling framework for predicting passenger flow on trains: A case study of London underground trains. *Transportation Research Part A: Policy and Practice*, *151*, 119-139.