Hussain, Etikaf, Bhaskar, Ashish, & Chung, Edward
(2023)
Zone prioritisation for transit improvement using potential demand estimated from smartcard data.
Transportmetrica A: Transport Science, *19*(2), Article number: 2028930.

https://doi.org/10.1080/23249935.2022.2028930

Zone Prioritisation for Transit Improvement Using Potential Demand Estimated from Smartcard Data

Etikaf Hussain[a], Ashish Bhaskar[a*], Edward Chung[b]

[a] *School of Civil and Environmental Engineering, Queensland University of Technology, Brisbane, Australia*

[b] *Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong, China*

[a]*corresponding author: Associate prof. Ashish Bhaskar, School of Civil and Environmental Engineering, Queensland University of Technology, Brisbane, Australia

Email: ashish.bhaskar@qut.edu.au

[a]Email: etikaf.hussain@hdr.qut.edu.au

[b]Email: edward.cs.chung@polyu.edu.hk

Keywords:

smartcard data, transit improvement, gradient boost, potential demand, public transit demand

Zone Prioritisation for Transit Improvement Using Potential Demand Estimated from Smartcard Data

It is of utmost importance to understand the networkwide transit service needs for future planning and effective funding allocations. For this purpose, this study proposes a methodology that uses a zone's transit potential demand as an indicator to prioritise them for public transport-related improvements. This study utilises observed demand (referred to as served demand) from smartcard data to estimate the potential demand. The smartcard data is used to estimate the observed demand of a zone, based upon which high and low trip zones are segregated. An ensemble tree-based Gradient Boosting model is trained and validated using observed trips by employing demographics, socio-economic, and geographic variables. From the analysis, zones with high and low potential demand are identified. Based on the estimated potential demand per unit area, all the zones are clustered into four groups identifying the areas with the lowest, low, medium, and high transit improvement requirements.

Keywords: smartcard data, transit improvement, gradient boost, potential demand, public transit demand

1. Background and literature review

In an urban area, potential or induced travel demand for public transport from a zone can be used as an indicator of the need for transit service improvement. Zones with high potential travel demand are most likely to have a high rate of return on investment in public transit projects as compared to those with low potential travel demand. From a public transport planning point of view, it is of utmost importance to prioritise the zones to receive funds for transit-related improvement (both infrastructure and supply wise).

Overall, travel demand estimation is a well-studied topic and is vital for planners and transport engineers in an urban area. It serves as the primary input for transport and mobility-related infrastructure planning. Demand estimation is essential for the control, operation, and management of urban travel facilities.

There are several models exist in the literature that are explored for demand estimation, ranging from linear models to multivariate, to log-models, etc., (Gaudry and Wills 1978). Traditionally, four-step demand modelling and activity-based modelling are employed to estimate the travel demand. The first step in the 4-step trip demand modelling consists of trip generation, which is usually done based on regression models. More lately, the trip generation is calculated based on activity models, assuming that activities trigger trips. These models, along with the statistical methods with the survey data (Household Travel Survey (HTS)) (Stopher and Greaves 2007), are employed to estimate trip attraction/production from one zone to all other zones (Toole et al. 2015). The HTS is very carefully designed to be able to provide related information. Due to the high amount of time requirement and cost of such surveys, the sampling methods are critical, so that the HTS results need to be generalised over the entire population. Also, because of its cost, the temporal resolution of HTS varies from 3 years to 10 years for different cities. Demand for the base year is calculated using the HTS data projected over the population. For the rest of the years, predicted travel demands are used to plan and construct new transit infrastructure-related decisions. With the advent of new big datasets and their availability, transport planners can access the low cost per sample and finer temporal resolution data (longitudinal data). Examples of such datasets are; loop detector data (Chen et al. 2001), mobile phone data (White and Wells 2002), smartcard data (Barry et al. 2002), Bluetooth data (Bhaskar and Chung 2013; Nusser and Pelz 2000), to name a few. Depending on the type of data, it may have a very high penetration rate. Though, the mentioned datasets are primarily designed for other purposes. For instance, smartcard data's primary objective is fare collection. With these datasets, it is required to develop algorithms and validate them to be used with current models.

The smartcard data records the boarding and alighting information of passengers. If a system is an entry-only or exit-only system, the smartcard data contains only boarding or alighting information, respectively. However, if a system is an entry-exit, it records both boarding and alighting information, as passengers are required to tap their cards while entering and exiting the system. So far, a considerable research has been done to exploit the uses of smartcard data in transportation, most important are being transit origin-destination (OD) matrix estimation (Sánchez-Martínez 2017; Alsger et al. 2016), travel pattern (Kusakabe and Asakura 2014; Naveh and Kim 2019), activity detection (Han and Sohn 2016; Nassir, Hickman, and Ma 2015), transit performance (Trépanier, Morency, and Agard 2009), delay predictions (Yap, Cats, and van Arem 2020), etc.

The smartcard data can be employed to find the number of trips produced or attracted to a zone(Hussain, Bhaskar, and Chung 2021b). However, this transit travel demand is highly biased towards transit service availability, i.e., places with high frequency and high-quality transit services attract more passengers (Hussain et al. 2021). Thereby, the OD matrix from smartcard data represents the served demand (SD) instead of total demand (TD). A transit potential demand (PD) may exist in a zone (or place), which is not served due to spatial and or temporal non-availability of transit services. Apart from the non-availability of transit services, other factors also account for non-utilisation of transit for travelling, such as higher travel time, quality of service provided, fare affordability, security, type of journey, to name a few (Nurdden, Rahmat, and Ismail 2007; Beirão and Cabral 2007). PD is defined as the induced demand that will be present if an appropriate transit service is provided. The PD determination can lead the planners and policymakers to prioritise the areas for transport-related funding and improvements.

Yao (2007) estimated the relative transit PD by proposing a transit need index by employing multiple regression in the literature. The same variables incorporated in multiple regression are also used to develop self-organising maps to cluster the zones based on the independent variables. Another study used the GIS-based approach to determine transit PD (Aljoufie 2014). The study categorises the zones in low, medium, and high need for public transit based on a composite index developed from regression. The above studies used the demand (demand due to workers only in Yao (2007)) and proposed an index for transit prioritisation; however, this study uses a machine-learning algorithm to enhance the prediction and uses observed total transit demand (or served demand) from smartcard data.

Apart from econometric regression, studies have utilised big data sources to identify the transit potential demand or identify the OD zones for which the public transit improvement may be needed. For instance, Olleczek et al. (2014) used the smartcard and cell phone data to identify the OD-pairs with high car trips and low public transit in Singapore. Regt et al. (2017) made use of smartcard and cell phone data to estimate the potential demand for public transport. The differences between the trips made in smart card data and visitors detected by cell phone data are compared to make inferences for the transit need. While developing a demand-oriented coordination model for rail, Li, Luo, and Cai (2019) found the potential demand for the last rail of the day by integrating AFC (rail and bus) and taxi data. The study found potential demand (the missed demand due to transfer non-coordination) by identifying trips made on buses and taxis after the last train had passed near rail stations. In a study, Cheng et al. (2020) determine the taxi demand from GPS data that can potentially shift to rail. The method applies spatial and temporal closeness of historical taxi trips to be potential rail trips.

The above-cited literature portrays that attempts have been made to understand and determine the transit potential demand. However, in most cases, they only consider the regression methods and the projected data from surveys, such as HTS, which may not accurately represent the network demand. Similarly, other studies utilised smartcard data along with other big datasets, i.e., cell phone data, taxi data, and GPS data. Therefore, the contribution of this study is three-fold; firstly, this study presents a methodology to estimate the relative transit potential demand from smart card and census data. Secondly, machine learning algorithms are explored instead of regression models for improved predictability. Thirdly, this study prioritises zones for transit improvement/future funding by employing relative potential transit demand.

To this end, the paper is arranged as follows: Section 2 describes the proposed modelling methodology; Section 3 includes the detailed results obtained; and, the last section (Section 4) provides the conclusion of the study along with future research directions.

2. Methodology

The proposed methodology to estimate potential transit demand by employing demographic and socio-economic variables and smart card data is divided into four parts and presented in Figure 1. Firstly, the transit trips produced/attracted to a zone are calculated from smart card data, and high and low ridership zones are identified, where high ridership zones will be utilised to train the model. Secondly, the training and testing of the model are carried out. Once the model is tested, it is applied to the remaining zones (zones with lower ridership) for the estimation of potential transit demand in the third step. Lastly, the zones are clustered based on their potential demand for transit improvement. The above-stated steps are described in detail in the following paragraphs.

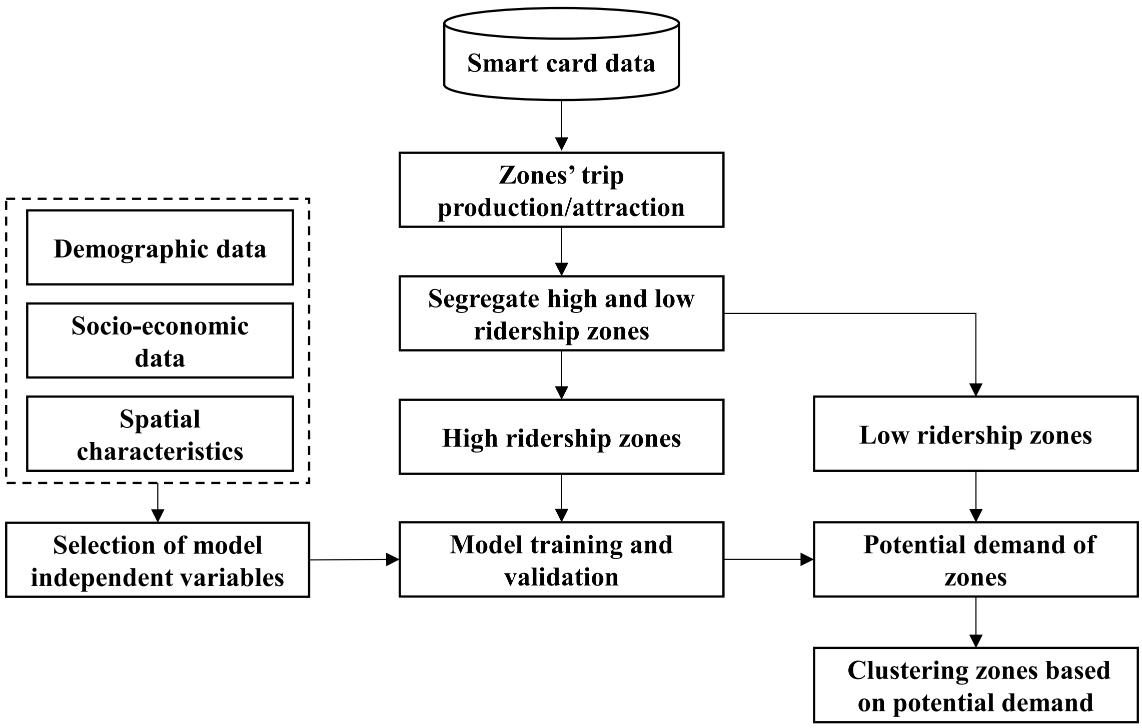


Figure 1 Workflow of the study to determine zones' potential demand and their prioritisation for transit improvement.

For a zone 'z', total transit demand (TD_z) consists of served demand (SD_z) and potential demand (PD_z), written as:

$$TD_z = PD_z + SD_z \quad \forall z \in Z \tag{1}$$

Where subscript 'z' denotes a zone in a set of total zones' Z'. Smartcard data gathered by transit agencies contain either boarding event information in an entry-only system or boarding and alighting event information in an entry-exit system for all passengers (except those using paper tickets). An entry-only system is where passengers are required to tap only once, either while boarding or alighting. In contrast, an entry-exit system is where passenger taps smartcard while boarding and alighting. In any respect, both cases contain trip-leg details made on the transit network. Trip legs can be converted to trips/journeys after performing some calculations. Once trips are extracted from trip-legs from smartcard data, it provides the transit demand which has been served (SD).

Potential demand is defined as the expected number of extra trips that will occur from a zone if appropriate transit service is provided to residents of that zone. The significant factors contributing to impact PD may include spatial and temporal non-availability of transit service. If transit service is available, people may not choose it because of many other factors such as, high walking distance to access and egress public transit facility, longer waiting times at stop, longer travel time (including the number of transfers involved in the journey), affordability, safety, quality of human-public transit interaction infrastructure (e.g., type of bus stop, condition of transit vehicles), etc. However, this study does not aim to determine the absolute potential demand (as per the above definition); instead, this study determines the relative potential demand in low ridership zones in order to have similar demand as that of high ridership zones given the fare, type and quality of transit service in the area remain same.

Based on the above construct, this study considers that there may exist no, or negligible PD in high ridership zones (H). Mathematically,

$$PD_z \approx 0 \qquad \forall z \in H \supseteq Z \tag{2}$$

Hence, for high observed demand zones, Equation (1) can be re-written as:

$$TD_z \approx SD_z \qquad \forall z \in H \supseteq Z \tag{3}$$

The above equation states that TD is approximately equal to SD in zones having high transit usage or in zones having the availability of the high quality of transit service.

Here, the PD estimation is for reference only due to the non-availability of the ground truth. The actual or true PD may vary from PD calculated from Equations (2), (3), and (4). Thereby, this study aims to find the PD for all other zones using the demand from high usage zones, which will be employed to prioritise zones for transit improvement. Mathematically, PD is the difference between total demand and served demand, represented by the following Equation (4).

$$PD_z = TD_z - SD_z \qquad \forall z \in Z \,\&\, z \notin H \tag{4}$$

The above equation can be utilised to estimate the transit potential demand for a zone provided that total demand and served demand for a zone are given. Please note that a suitable transit supply (Hussain, Bhaskar, and Chung 2021a) or accessibility index can also be employed for zone segregation instead of employing the SD. Consequently, the zone with higher supply can be labelled as (H) instead of high ridership and vice versa.

For this purpose, a machine learning model can be trained by employing TD_z for all the zones with higher demand as the dependent variable. Various explanatory variables related to demographics, socio-economic, land use, etc., which contribute to demand production/attraction, can be employed to train the model. Once the model is trained and validated using TD_z for only high utility zones; that when applied to all other zones (low utility zones), will provide the TD_z for those zones. TD_z for those zones are, theoretically, greater than their corresponding SD_z.

2.1 Transit trips production/attraction estimation from smartcard data

The primary objective of accumulating smart card data is the management of fare collection. The structure of collected data by an agency varies across the transit providers. The type of data gathered depends on the type of fare system installed. The data are from either an entry-exit system or an entry-only system. Typical boarding and or alighting information consists of the time and location of the events. The type of datasets and various methodologies to estimate Transit Origin-Destination matrix (tOD) are discussed in detail in Hussain, Bhaskar, and Chung (2021b).

This study does not build on the literature of tOD using smartcard data. It utilises pre-established procedures identified in the literature and emphasises the application of big data generated from smartcard. Estimating tOD from an entry-exit system data is relatively easier and straightforward as it only requires separating trips from journeys. A trip includes boarding and alighting a transit service; however, a journey may consist of many trips to reach the destination. In brief, it is necessary to identify transfer and activity in order to estimate true tOD. Once the tOD is estimated, the row-wise summation gives transit trips produced, and the column-wise summation provides transit trips attracted.

2.2 High demand zones selection

After trips produced/attracted are estimated, the next step comprises the selection of appropriate zones for model training (shown in Figure 1). For this purpose, zones with high and low usage of transit services can be identified by utilising the concept provided in the studies Hussain et al. (2021) and Louail et al. (2015), where zones are clustered into low, medium, and high trip production/attraction zones. More specifically, trips produced/attracted from a zone are divided by its area, giving trips produced/attracted per unit area. The rows are then arranged in ascending order of those values.

According to Hussain et al. (2021), fewer zones contribute to high transit ridership (Figure 2 shows the same phenomenon for the study area). The figure depicts the distribution of trips attracted or produced per unit area from a zone in descending order. Zones above a cut-off value α are considered high trip production/attraction zones. Modelling results can be sensitive to the selection of α. However, the difference in the values of observed zonal demand for low and high demand zones is generally very prominent, as in the case of Brisbane (Figure 2), and analysts can make a reasonable estimate of α by checking the trip production per unit area. It is expected that a small change in α will not drastically change the ranking of a zone, which is the output of this study. For further research, it is recommended to thoroughly test the sensitivity of the modelling with respect to different values of α. Also, a higher value of α will provide a lesser number of high utility zones. Whereas for model training, it is favourable to employ a high number of cases (high utility zones in this case).

After high trip zones are identified, these zones will be further used in model training explained in the next sub-section.

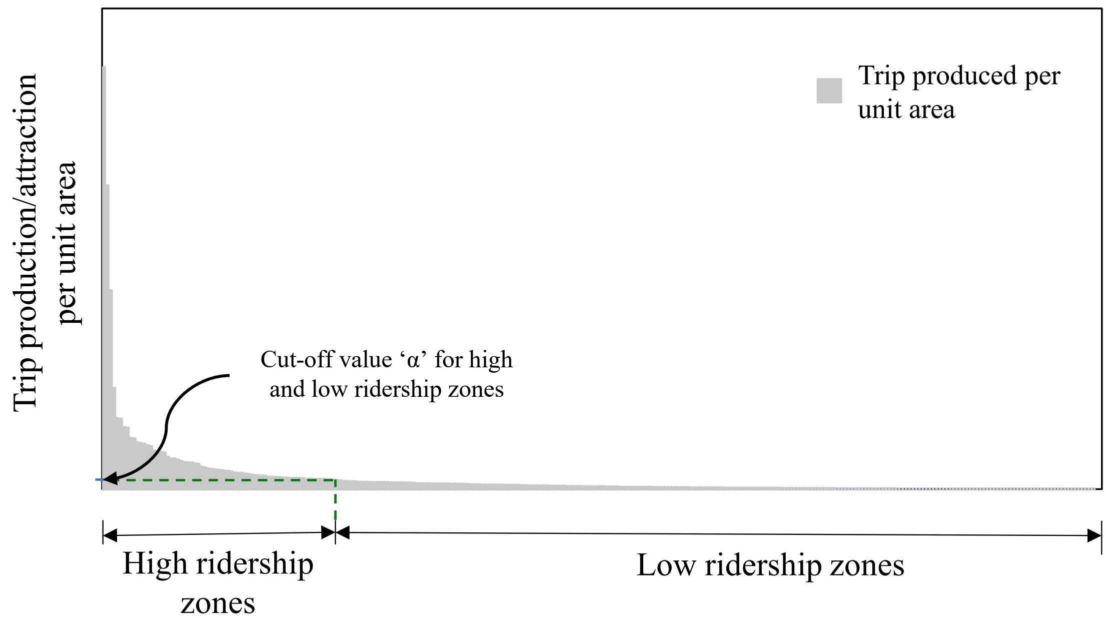


Figure 2 Illustration of the typical distribution of per unit area trip production/attraction in descending order in a transit network.

2.3 Model development

In this study, the model development step is divided into an explanatory variable selection from demographic, socio-economic, and land-use variables and training and validation of the model. The last step is to infer the total demand from the validated model. Below, the sub-section describes these steps in detail.

2.3.1 selection of explanatory variables

For model development, three groups of variables are initially identified from the literature and are selected for analysis. The type of datasets includes demographic variables, socio-economic variables, and spatial characteristics of the zone. The demographic variables mainly comprise the population with different age groups in a zone. The socio-economic variables include population with a different group of income, employed/unemployed persons, the number of households with number of cars, etc. The spatial characteristics of a zone may contain the zone area, the distribution of land use (commercial, residential, etc.), number of park & ride spaces, etc.

To model the total demand, there is a long list of candidate independent variables. However, when calibrating the model, there is a high chance that we have fewer zones with high transit demand. Therefore, it may not be feasible to use all the variables. Hence, applying one of the dimension techniques is required to reduce the candidate independent variables. There exist various dimension reduction techniques such as analysing the correlation amongst the variables, principal component analysis, backward/forward feature elimination/selection, recursive elimination technique, to name a few.

In this study, a feature of the gradient boost model is employed, which provides the variable's importance and rank the variables accordingly. The importance of a variable is shown by the mean increase/decrease in the model's error when that variable is included/excluded (Breiman 2001).

2.3.2 Model selection

Many machine learning algorithms are currently employed for classification, ranking, and regression. Machine learning algorithms are preferred over econometric models due to their enhanced prediction capability.

In order to select a suitable model for potential demand estimation purpose, three machine learning algorithms (artificial neural network, random forest, and gradient boost) and a suitable econometric model (negative binomial model) are chosen and tested for their prediction. An initial investigation suggested that the gradient boost algorithm outperform all other econometric and machine learning algorithms employed.

In this study, a gradient boosting algorithm is selected to model the total demand by utilising the variables found in the last step. Gradient boosting is a supervised ensemble machine learning algorithm that builds shallow independent trees (Friedman 2001). Each tree in gradient boosting learns from the previous tree and improves the results (Boehmke and Greenwell 2020). One of the tree-based method's advantages is that averaging independently grown trees does not allow the model to overfit by employing appropriate hyperparameters.

The primary task of boosting is to combine the model sequentially to the ensemble, which is more effective when the model has high bias and low variance. Given that, it averages the error in regression, which decreases the variance. Thereby, it is efficient for the model to have high variances and low bias. The trade-off between the bias and variance compels the algorithm to start with a weaker model, which is then enhanced in every next model (tree) by transcending previous models' problems (tree).

The gradient founds the improvement. For more details on the model's mathematical formulation and workflow, the readers are suggested to refer to Friedman (2001) and Natekin and Knoll (2013).

2.3.3 Model training and validation

Once the independent variables and models are selected, the next step is to train and validate the model. Gradient boost algorithm includes two types of hyperparameters, i) those related to boosting and ii) those related to tree building. Training these parameters requires fine-tuning before finalising the model.

One of the boosting-related hyperparameters is the number of trees in the model. As discussed above, the averaging method of gradient boost tends not to overfit the model with a slightly higher number of trees; however, a high number of trees can overfit the model. Therefore, it is suggested to find the optimal number of trees. The other boosting related hyperparameter is the learning rate, which dictates how quickly the algorithm learns. Its value ranges from 0 to 1. The smaller the learning rate value, the better the model provided that it tends to overfit the model. Also, with smaller steps, the model may not reach global minima.

Tree related hyperparameters are tree depth and minimum observations in terminal nodes. The former restricts the depth of the tree. The smaller values of tree depth are computationally efficient. On the other hand, a model with higher values can capture more variance. Therefore, it also increases the chance of overfitting. The minimum number of observations in terminal nodes specifies the complexity of trees in the model. It usually does not affect the model since gradient boost utilises the shallow trees. The typical values for this hyperparameter range between 5-15.

After training the model, the next step is to validate the model. In this study, the number of cases (or high trip zones) is limited. Hence the one-left k-fold validation method is found suitable for this purpose. In one-left k-fold validation techniques, one case is excluded from the data, and the model is built before each run. The same model is then used to predict trips for that case. The difference between the observed and predicted value is the error. Two types of errors are calculated, namely mean absolute error (MAE) and root mean square error (RMSE). This process is repeated until all the points are left once from the model.

2.3.4 Total demand and potential demand inference for low trip zones

After training and validation of the model using total demand from high trip zones, it can be employed to give the total demand for low trip zones, from which the potential demand can be computed as per Equation (4). In Equation (4), the demand values (total demand) estimated from the modelling should provide higher values than served demand (from smartcard) for all the zones of analysis. This will enable us to justify the assumption made for Equations (2) and (3).

2.4 Application of the developed methodology

This section provides details about the application of the developed framework for the estimation of potential transit demand.

2.4.1 Study area

The proposed framework to estimate the PD for transit service of a zone used as an indicator to prioritise zones transit improvement is applied to Australia's South-East Queensland region. For the analysis, 298 statistical analysis 2 (SA2) zones with a median area of 7.86 sq. km are included. Figure 3 shows the study area SA2 zone boundary delineation superimposed with the transit routes. The study area has the Pacific Ocean to the East. Also, the seaport and airport lie in Brisbane East. Brisbane central business district (CBD) lies in the Brisbane Inner-city. TransLink operates the public transit services in this region. The public transit modes in the study area consist of buses, trains, ferries, and trams.

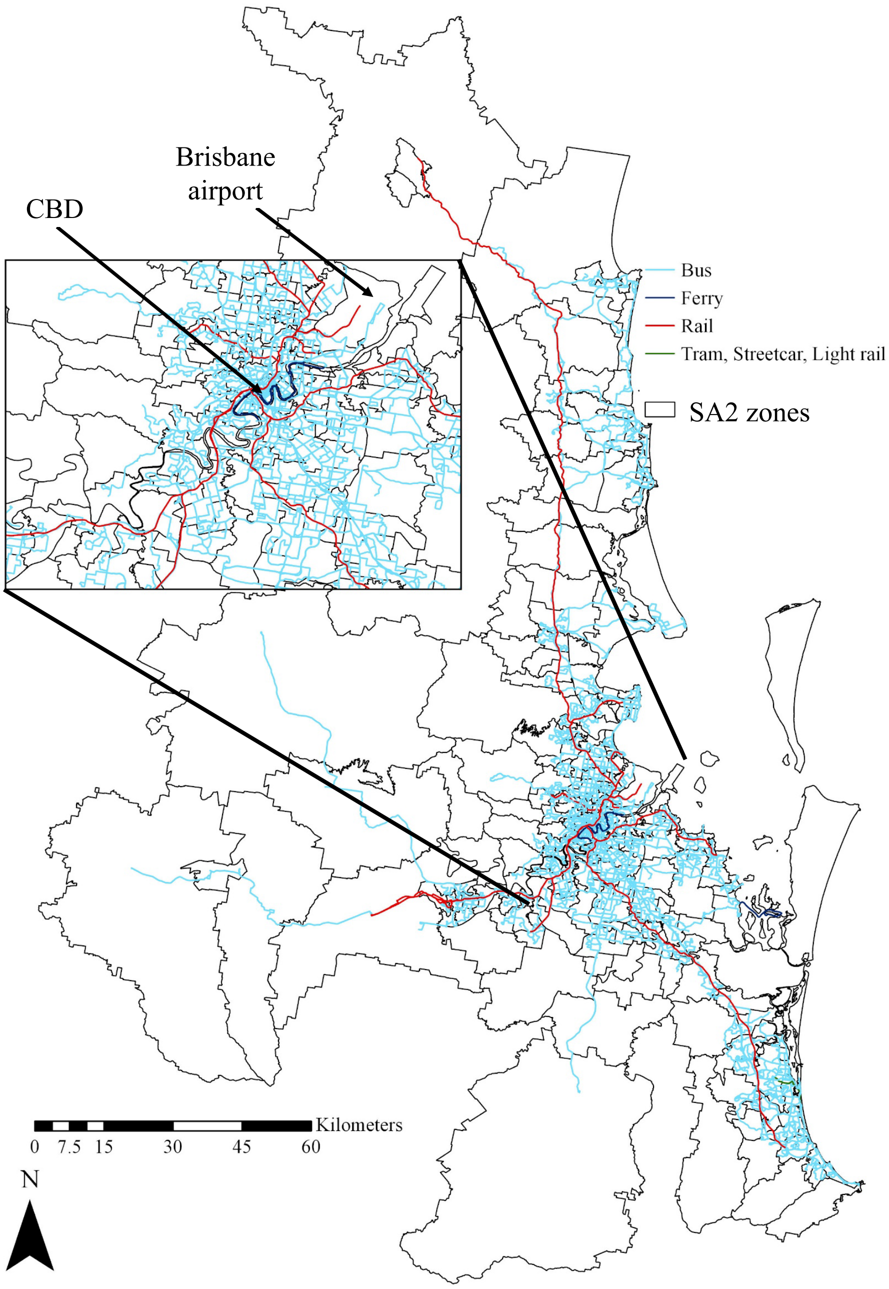


Figure 3 Study area map, South-East Queensland region’s SA2 zone boundaries.

2.4.2 Served demand (SD_z) estimation

The first step to obtain the PD_z is to calculate the SD_z from smart card data. For the analysis, the smartcard data of 7th March 2017, a typical day, are acquired and used for transit OD estimation. The smartcard data contains 31 fields, out of which six fields are retained, which are vital for tOD estimation (Table 1). The rest of the fields are dropped from the data for simplicity and computational efficiency. In smartcard data, EIS (executive information system) is the 20 digits encrypted smartcard number; operations date is the date on which the trip is made; boarding time and alighting time, correspondingly records the start and end time and date of a trip; and, boarding stop and alighting stop show the stop number from where a trip starts and ends, respectively.

EIS	**Operations date**	**Boarding time**	**Alighting time**	**Boarding stop**	**Alighting stop**
20-digit number	2017-03-07	11:33:47	11:50:32	2606	19052
20-digit number	2017-03-07	15:23:49	15:39:28	19064	2630
20-digit number	2017-03-07	07:52:18	08:07:31	C128	C5
20-digit number	2017-03-07	18:10:19	18:22:07	C5	C128
…	…	…	…	…	…
…	…	…	…	…	…
20-digit number	2017-03-07	07:52:23	08:31:27	6399	63

Before application, the smartcard is cleaned for anomalies, such as missing boarding and or alighting time. It is also possible to have an incorrect or missing boarding/alighting stop. All such transactions are deleted, and the rest of the data are applied to calculate SD_z. Before estimating SD_z, the trip (individual trip-leg) represented by each row/transaction from smartcard data is converted to journeys (one trip or combination of trips based on transfer inference).

Various researchers have developed many criteria to differentiate between transfer and activity. For this purpose, as per literature, a threshold for transfer time, which is the time between successive alighting and boarding, is used to separate trip-legs from journeys. This study utilises the criteria of 30 minutes between alighting and next boarding for transfer inference (Hussain, Bhaskar, and Chung 2021b). It suggests, if a person spends less than 30 minutes between alighting a transit service and boarding the next transit service, it depicted transfer. In this case, the two trips would be merged to represent one journey. Furthermore, the physical location of a stop is considered while assigning a zone to a stop, i.e., if a stop geographically lies in a zone, all the transactions made on that stop are considered from the same zone.

After the tOD matrix is estimated from smartcard data, the trips made from a zone to all other zones are summed to get SD_z of that zone.

2.4.3 High demand zone identification

Following the tOD estimation, the trips produced/attracted are divided by that zone's area to get the trips produced/attracted per unit area. Then all the rows are sorted in ascending order based on the resulting values.

From the sorted tOD, high demand zones are identified based on the methodology described in section 2.2. Here, the cut-off value between high and low demand zones (α) is taken as 360 trips per unit area per day. From the sorted tOD matrix, the zones having trips produced per unit area of more than 360 are identified and are labelled as high trip production zones. The model will be trained for these zones (as per Figure 2). Here, we get 67 high demand zones responsible for 63% of the total trips, and the rest of the zones having low demand, for which the served demand will be estimated. It is to be noted that Brisbane CBD is not considered for analysis, as it has a very high number of trips (three times more than the second-highest trip production zone) as compared to other zones. Hence, it is believed as an outlier.

2.4.4 Explanatory variables data

A list of explanatory variables based on demographic, socio-economic, and spatial characteristics of zones are identified. Table 2 presents all the independent variables initially selected for model training.

Table 2 List of independent variables to be considered for model training.

Group	No.	Variables	Min.	Max.	Mean
Demographic variables	1	Persons aged 5-14 years	4	5600	1352.4
	2	Persons aged 15-19 years	0	2286	667.1
	3	Persons aged 65 years and above	0	7349	1550.9
	4	Total population	25	31214	10521.3
	5	Total students	8	10630	3248.1
	6	Persons need assistance	0	1672	468.2
	7	Population density	0.7	6134.9	1500.7
Socio-economic variables	8	Unemployed looking for work	0	1368	390.8
	9	Unemployed labour force	6	9250	2654.6
	10	Percent employment	16.4	73.7	58.1
	11	Jobless persons or with negative income	0	2843	759.5
	12	Persons having low income	3	11301	3669.1
	13	Persons having medium income	3	8279	2717.5
	14	Persons having high income	3	2044	645.5
	15	Persons with unpaid work	19	16645	6074.0
	16	Number of homes with 0 motor vehicles	0	1336	227
	17	Number of homes with 1 motor vehicles	4	4896	1287.6
	18	Number of homes with 2 or more motor vehicles	3	6687	2082.1
Zone characteristics	19	Commercial service area (sq. km)	0	3.02	0.30
	20	Urban residential area (sq. km)	0	16.66	3.20
	21	Rural residential without agriculture (sq. km)	0	105.50	4.90
	22	Rural residential with agriculture (sq. km)	0	86.55	0.80
	23	Zone total area (sq. km)	1.21	2544.66	51.1

Group	No.	Variables	Min.	Max.	Mean
	24	Park n Ride spaces (numbers)	0	1780	102.3

Out of 24 variables, seven belong to demographic, eleven to socio-economic, and six variables to zonal characteristics. The variables are carefully chosen based on the literature. The proposed estimate of the transit PD should not be dependent on the supply. Therefore, we do not consider transit supply or accessibility index as independent variable for the modelling. Parameters given in Table 2 can be used to explain the variation in transit demand in the zones. As mentioned in the earlier section, the number of zones in high demand is limited (only 67 zones). Using many explanatory variables cannot be used for model training as it may not provide an accurate model.

Among many data reduction techniques available, this study utilises the gradient boosting method built-in feature in R package 'gbm' (Greenwell et al. 2020). Figure 4 presents the plot of relative influence on the x-axis and variables on the y-axis. The figure shows that the number of homes with zero motor vehicles is the most influential attribute, followed by jobless persons or with negative income. Out of 24 variables, the first seven variables were selected so that the model does not improve by adding another variable.

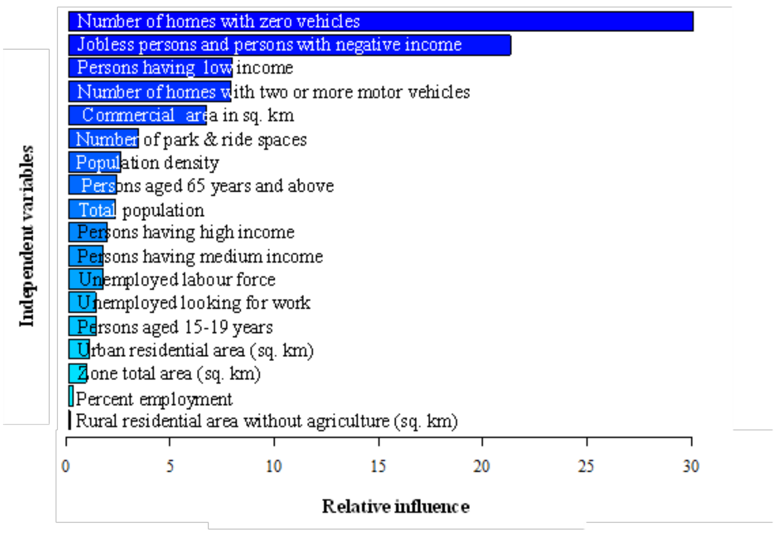


Figure 4 Relative influence for the individual independent variable.

Here, the independent variables chosen for the model training include (in descending order of importance) number of homes with zero vehicles, jobless persons and persons with negative income, persons having low income, number of homes with two or more motor vehicles, commercial area in sq. km, number of park & ride spaces, and population density.

3. Results

This section elucidates the results of the study. The results are related to SD_z, model results and its validation, total demand predicted for zones, and consequently PD. The PD can then be utilised as an indicator for prioritising transit service improvement.

3.1 Served Demand (SD_z)

The SD for the study area is estimated from the smartcard data following the methodology presented in section 2.4.2. The output, i.e., SD_z for the study area, is presented in Figure 5. The green colour shows high trip production zones, while the transition to red colour depicts the other way around. Figure 5 shows most of the zones with high SD are concentrated towards Brisbane City Business District (CBD), with a few exceptions in the Gold-coast region (the South region). Zones with low demand are scattered throughout the area, mostly far from the Brisbane CBD.

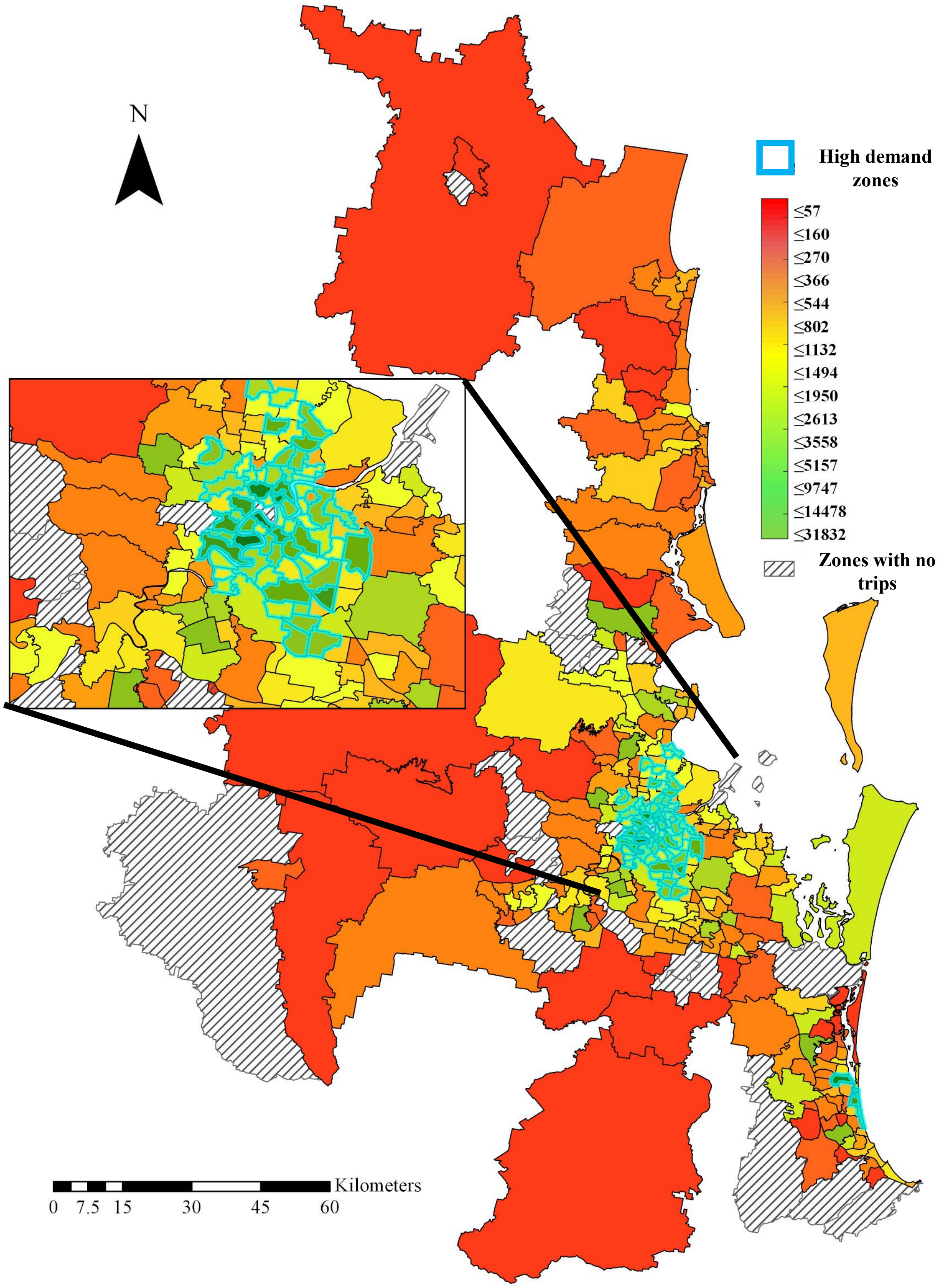


Figure 5 Graphical presentation of SD_z from smartcard data for different zones.

The trips from smartcard data are scaled up by employing a growth factor (Regt et al. 2017). TransLink reports smartcard penetration of around 90%. Therefore, the SD_z calculated from smartcard data is multiplied by a factor of (100/90) to cater for the lost trips (paper-ticket based trips or trips with missing boarding and alighting stops/time).

A minimum of 2 trips is observed from Ipswich – North and Gympie region, while a maximum observed trips from City Centre are 105168 in one day. The study area has 723 median trips. Due to high disparity in the trips originated from CBD and other zones, it is not considered further in the modelling.

3.2 High demand zones

As per the methodology proposed in section 2.2, high demand zones are identified from SD_z. The high demand zones are highlighted in Figure 5. It shows that the high demand zones are mostly near CBD. There are 67 zones responsible for 63% of all the trips produced. As described earlier, CBD has the most trips and is not considered in further modelling. Other 67 zones lie in the high demand zones and are employed in the model training. The remaining 230 zones lie in the low demand zones. Hence, the TD, and consequently, PD will be estimated for all these zones.

3.3 Model results

Three machine learning algorithms and a suitable econometric model are calibrated and validated to predict potential demand, results for which are shown in Figure 6. Two metrics, MAE and RMSE are employed to evaluate the models' results. From Figure 6, analysis indicates that the gradient boost model outperforms other models selected for this purpose. The next closest model is based on artificial neural network, where MAE increases from 1149 trips/zone to 1487 trips/zone, and RMSE increases from 2186 trips/zone to 2462 trips/zone. Therefore, the gradient boosting model is selected for this study, further details of which are given below.

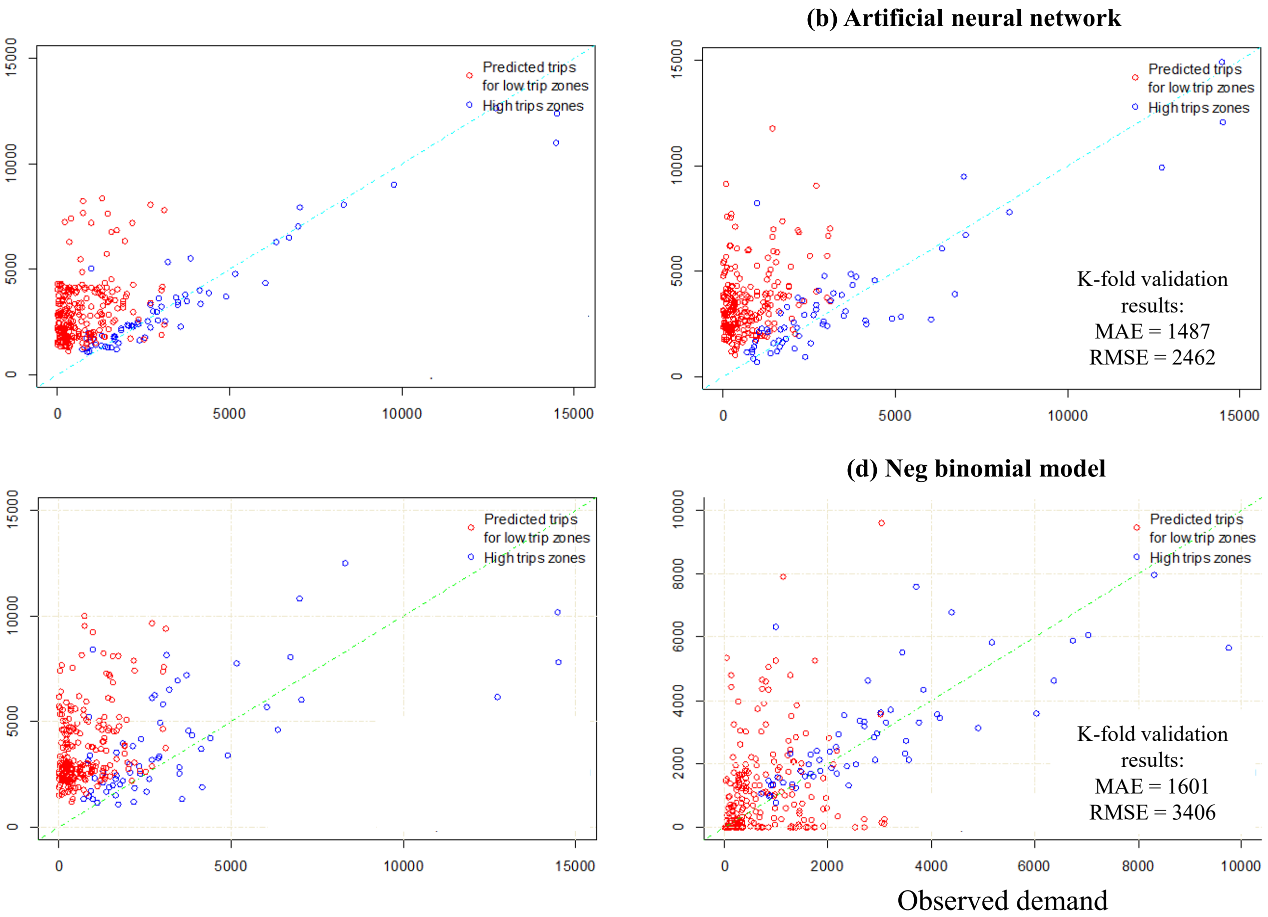
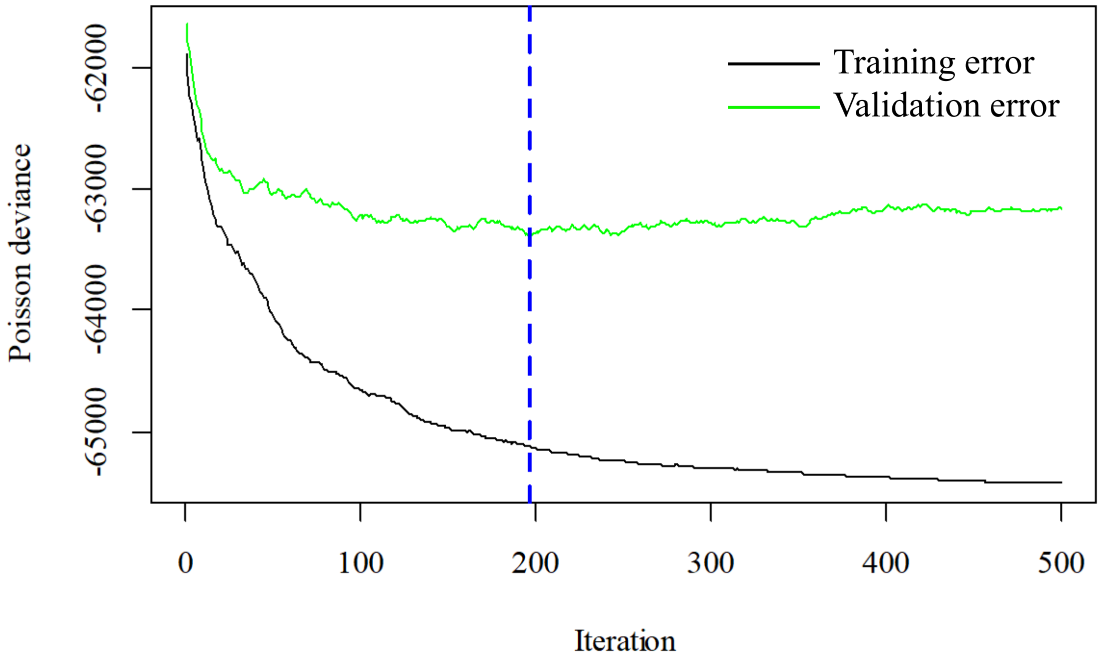


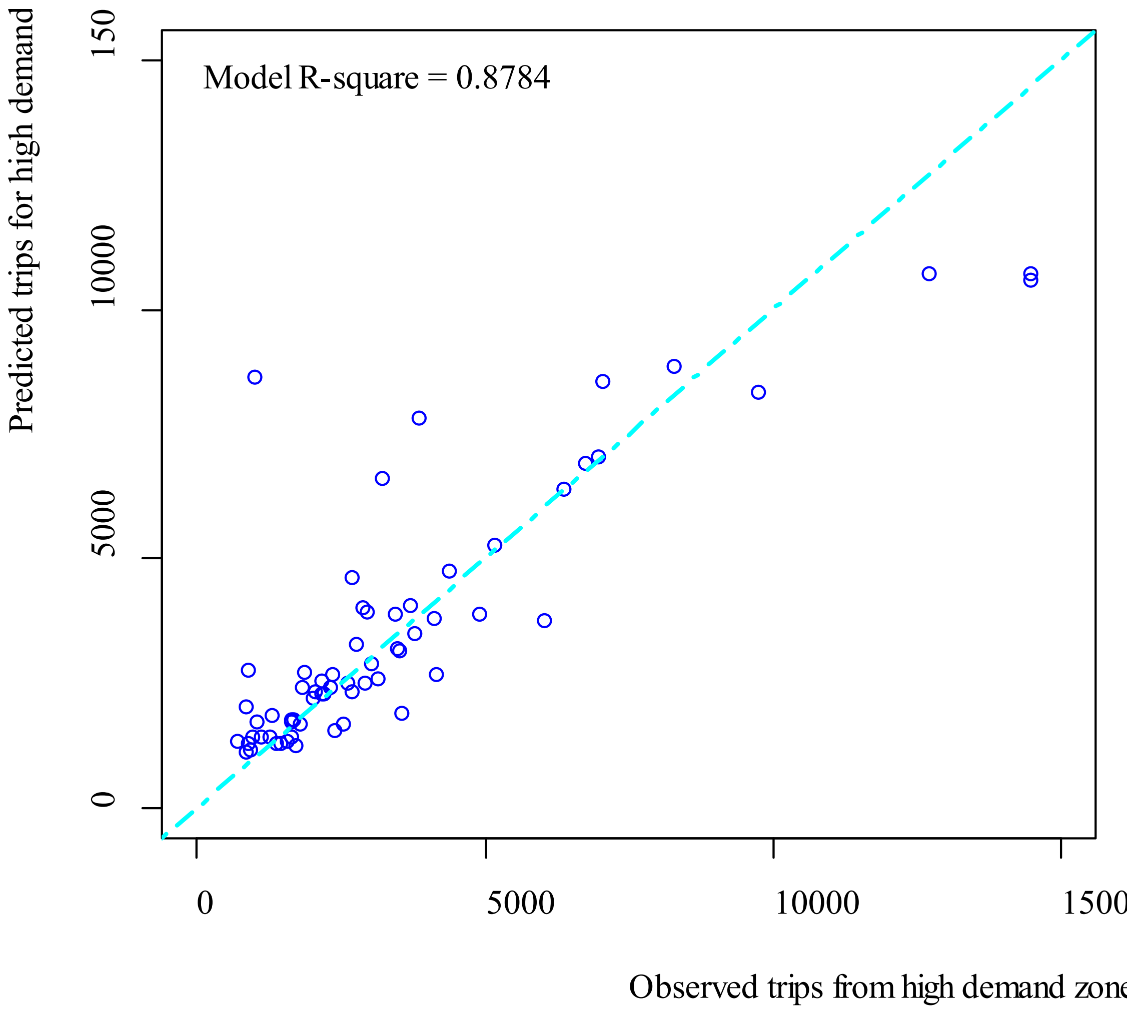
Figure 6 Comparison of various models – (a) gradient boost, (b) artificial neural network, (c) random forest, and (d) negative binomial model, initially employed to predict total demand

Hyperparameter selection is a vital step in machine learning algorithms. Here for the gradient boosting model, the hyperparameters are selected such that the error metric decreases in the model training and validation step. The 'gbm' package of R utilises Poisson deviance as the error function for model evaluation when the model's distribution is set to Poisson. The error metrics employed to evaluate the validation are the MAE and RMSE. All the hyperparameters, i.e., the number of trees, interaction depth, shrinkage (tree depth), and a minimum number of observations in terminal nodes are selected for which the evaluation error metrics are found to be minimum. The distribution of the model is kept as Poisson.

The number of trees is selected based on Figure 7, which shows a plot between the number of iterations and Poisson deviance. It portrays that Poisson deviance decreases with an increase in the number of trees; however, there is a negligible improvement in the validation error after initial improvement. The suggested optimised value is 197 trees for the model. Next, the values of other hyperparameters are selected by considering the error function. The final model has a shrinkage of 0.05, tree depth of five, and the minimum number of observations in terminal nodes are ten.



The selected gradient boosting model has a coefficient of determination of 0.878, which can be considered as an acceptable model and shows an enhanced prediction capability as compared to the earlier proposed similar model by Yao (2007) (R^2=0.678). The relation between the observed trips from smartcard data and predicted trips from the selected gradient boosting model is depicted in Figure 8. The figure shows that most of the points lie on the equity (45-degree) line, advocating a reasonable model fit. Please note that zones having trip values of more than 15000 are not shown in the figure for improved readability. For model training, the disconformity between observed and predicted trips in terms of MAE is 1118, and RMSE is 2152.



Tree-based algorithm inherently uses cross-validation techniques in building models. More specifically, it selects subsets of rows and columns for the training of the model. Nevertheless, one-left k-fold cross-validation is also applied to showcase the model's robustness and presents the model's independence on the input data. In one-left k-fold cross-validation, one row (i.e., one zone) is excluded from the input data, and the model is trained without that zone. The model is then used to predict trips for that left-off zone. In this way, the process is repeated for all zones (67 in this case). The results for the validation process are shown in Figure 9 (in blue). The validation has a MAE of 1149 and RMSE of 2186 trips.

Figure 9 further portrays the predicted demand (TD_z) for low demand zones, shown by red circles. As discussed earlier, all the predicted trips for low zones must lie above the equity line. In Figure 9, most of the points lie above the equity line; however, it also depicted a few red points below the equity line. This shows that there may still be high error exist in the model. Further, each point's perpendicular distance to the equity line is the potential demand of that zone. Hence, the more the perpendicular distance of a point from the equity line, the higher the potential demand. From Figure 9, most of the zones have TD between 2000 to 4000, while a few zones have greater TD (>5000 trips) where their SD is less than 3000 trips.

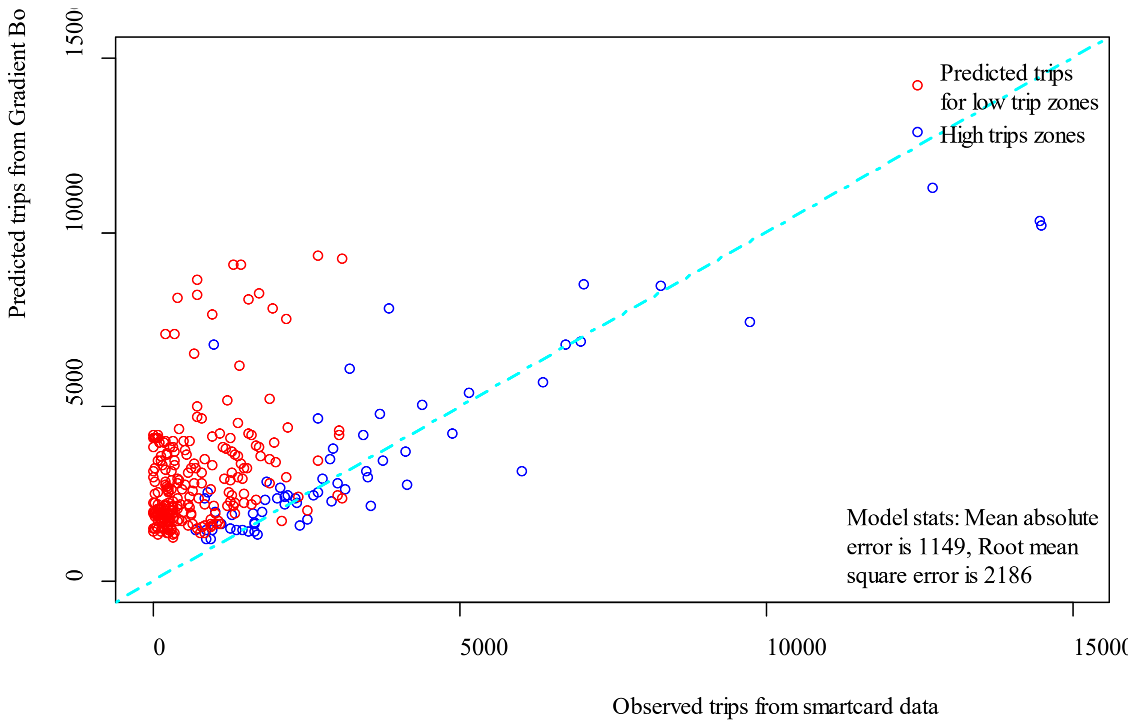


Figure 9 Validated trips (in blue) and predicted trips for low demand zones (in red).

The TD demand for all the low demand zones is depicted in Figure 10. The figure shows overall a high potential demand in the area for public transport. However, the choropleth map can be deceiving because a zone may have a higher PD; however, it would be more prominent in the map if it has more area than the one having less area. Nonetheless, it can be seen from Figure 10 that there sparsely exist zones that can positively contribute to public transport patronage.

It is to be noted that, in this study, the aim is to prioritise the zones for transit improvement by considering transit PD (derived from TD) and not the sole prediction of the PD (or TD) for a zone. As quoted earlier, the selected method (Gradient boosting) may not have given the best model. Nonetheless, the above-predicted TD is utilised for the analysis given below to showcase the proposed method.

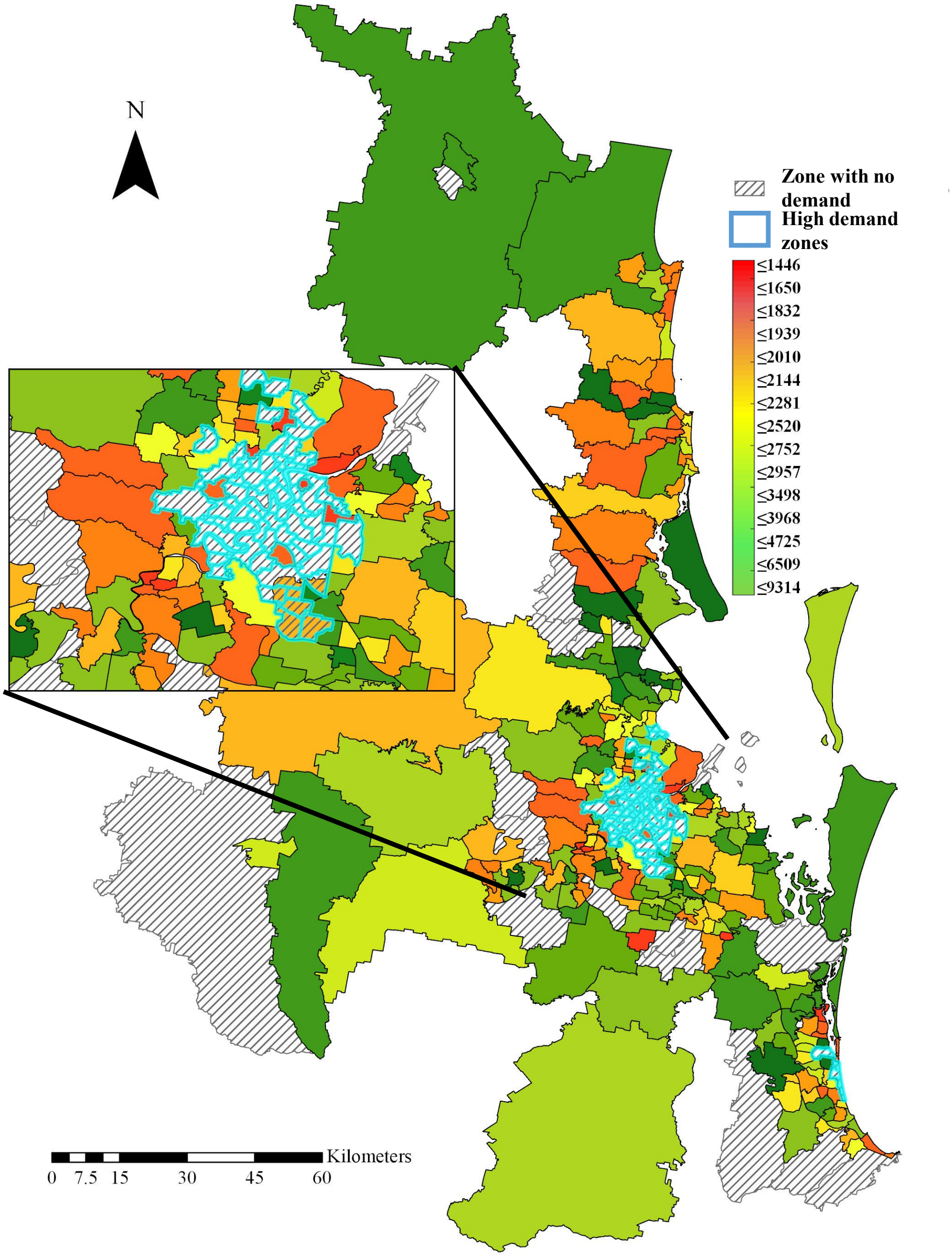


Figure 10 Choropleth map depicting TD estimated for low demand zones.

3.4 Potential demand (PD_z)

From TD_z and SD_z, found earlier, PD_z is calculated by employing Equation (4). The results of PD_z are presented in a choropleth map shown in Figure 11. Each zone has a colour corresponding to the PD value, where the red colour presents low PD, and the transition towards green depicts increasing PD.

Figure 11 portrays that most of the zones have high PD. As noted earlier, zones far away from Brisbane CBD are relatively bigger; consequently, the green colour is prominent in Figure 11.

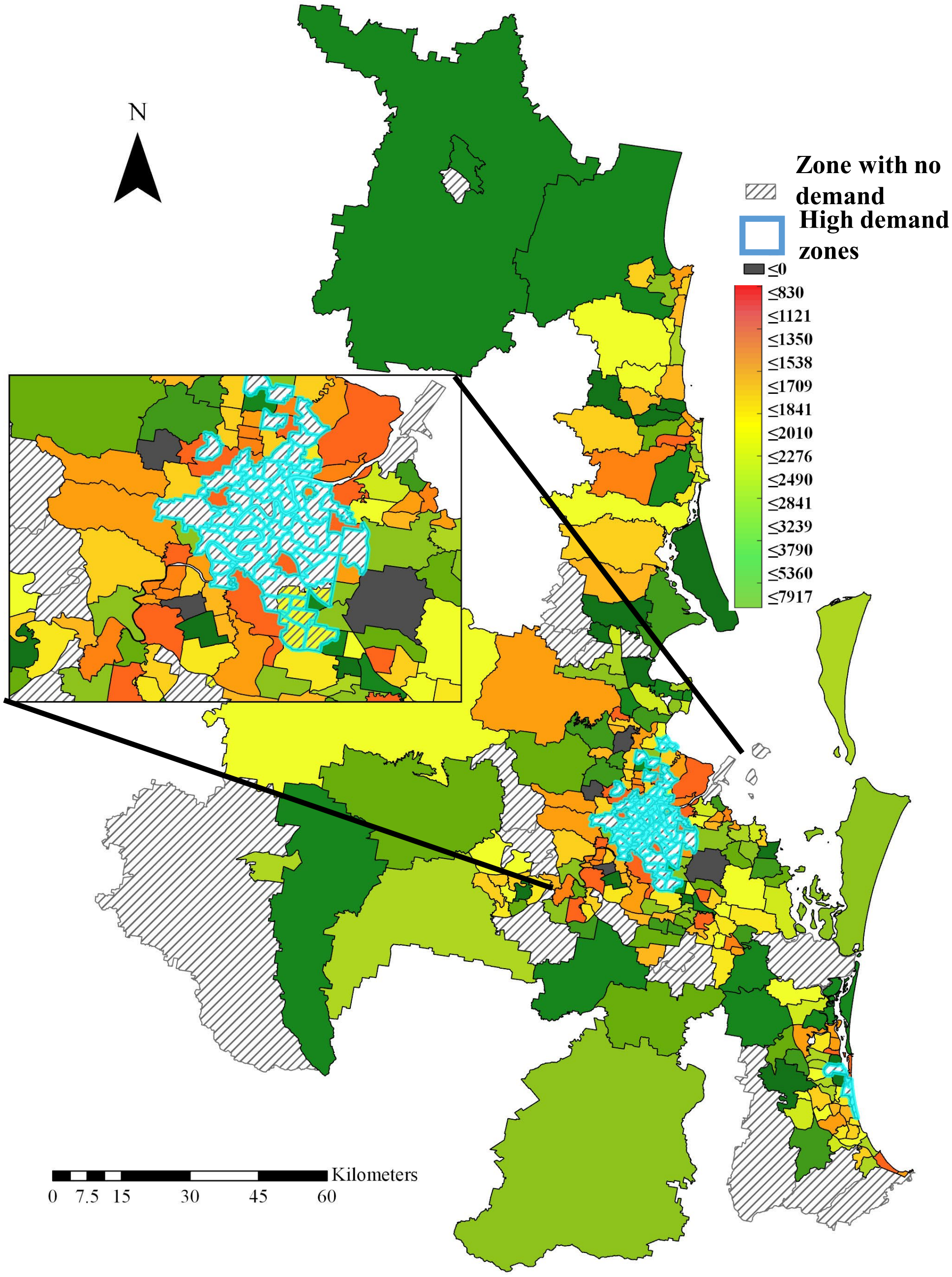


Figure 11 PDz illustration of low demand zones in the study area.

The zone area plays a vital part in the transit trip generation. Further, to compare the potential among two zones, it is required to normalise the PD by dividing the PD of a zone with its geographic area giving output in PD per sq. km. The resulting map is presented in Figure 12, where zones are further divided into four groups based on their PD values.

The figure shows that most of the zones in the vicinity of Brisbane CBD, along with a few zones in the Gold Coast region (in the South), have higher PD than those far away from the Brisbane CBD, for instance, zones in the far North-East, far South-East, South, and Western zones have lesser PD per sq. km. More specifically, the Lockyer Valley – East zone was expected to have the least priority primarily because of their low population density and type of land use (low commercial and urban residential area). In addition, Chapel Hill was highly likely to have high potential demand due to high population density, type of land use and other related variables.

In Figure 12, the higher the PD per unit area value of a zone, the higher the priority of that zone for transit improvement and the other way around. The zone classification is done based on PD percentile values. Zones in the first quarter are classified as the lowest priority zones. Likewise, zones in the top quarter are labelled as the highest priority zones. Low priority zones have the least PD per unit area. Correspondingly, zones having a PD of less than 80 per sq. km lie in this group. The second group, labelled as the low priority group, is based on PD range between 80 and 180 per sq. km. The third group, medium priority zones, has PD between 180 and 360 per sq. km, while other zones with PD greater than 360 per sq. km are termed the high priority zones.

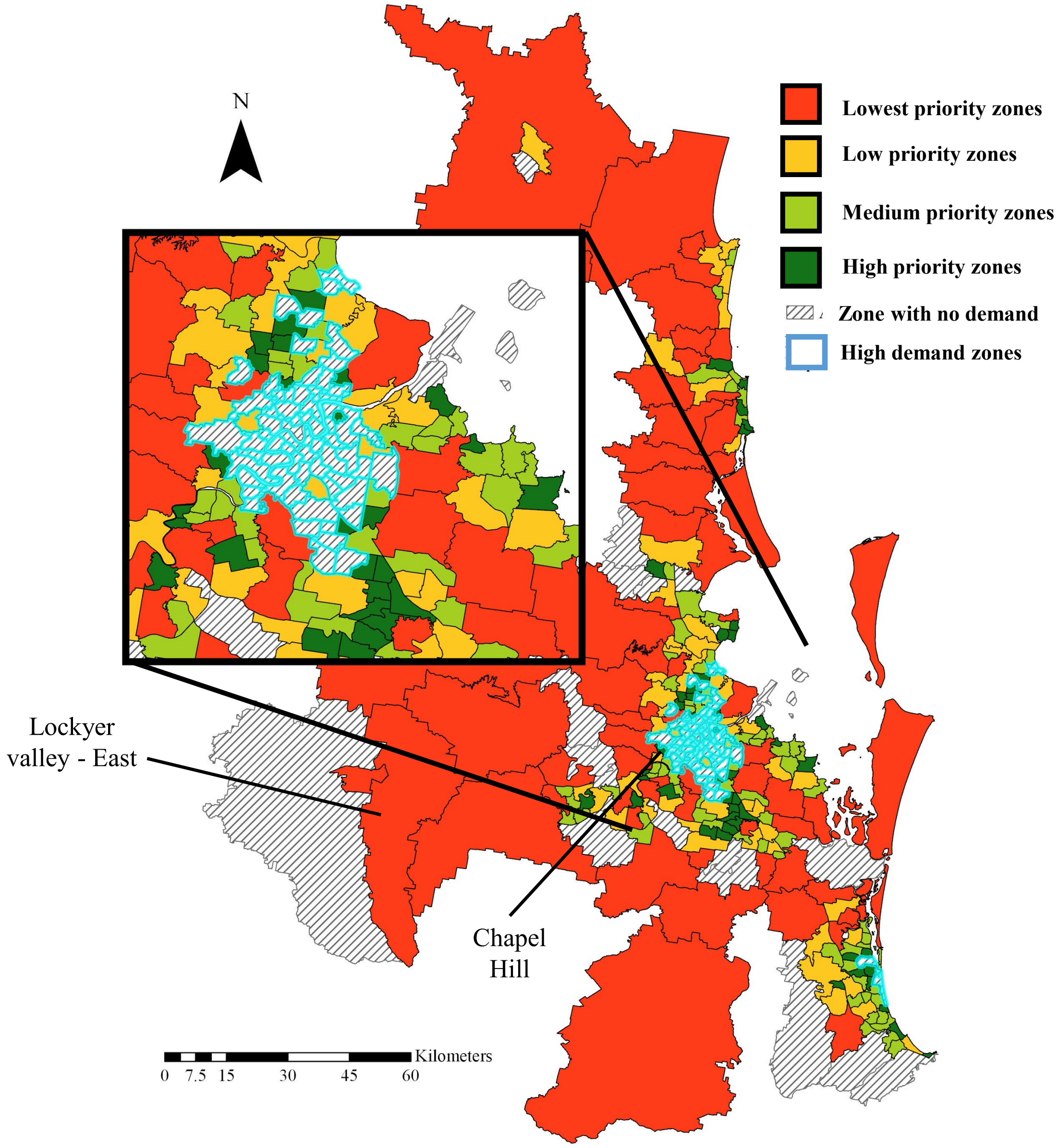


Figure 12 Choropleth maps depicting the priority of zones for public transit improvement based on PD per unit area.

Further to elaborate on the above figure, a histogram is plotted (Figure 13), offering an overview of all the network zones. The figure portrays that most of the zones have minimal tendency to produce high transit even if a good quality of service is provided. Approximately 50% of the zones have less than 180 trips day PD. Out of 232 zones, only eleven zones have unit area PD more than 630 trips per day, which can be treated as high priority zones for public transit improvement (shown mainly by dark green colour in Figure 12).

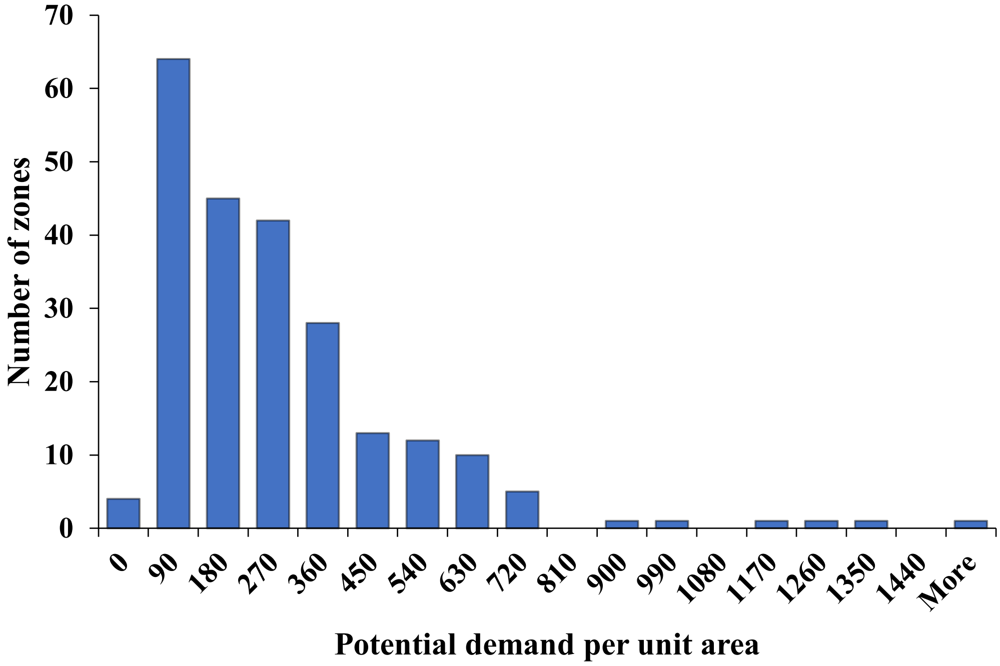


Figure 13 Histogram depicting the distribution of zones in the network based upon the PD from the unit area of each zone in the study area.

4. Conclusion

Transit demand estimation is a widely studied topic in transportation, which is vital for planners and decision-makers. This paper proposes a novel concept to rank the zones for transit improvements based on systematic mining of the transit and census data for a large urban network. The automated fare collection data serve as input to estimate the trip produced from a zone, termed as the served demand. Theoretically, served demand, when combined with potential demand, gives the total transit demand for public transport in a zone. The study assumes that high transit ridership zones have negligible potential demand. Consequently, total transit demand approximately equals served demand in these zones.

The devised methodology is applied to the South-East Queensland region using smartcard data from TransLink for demand estimation. A threshold value on the number of trips per unit area originated from a zone is applied to bifurcate between high and low demand zones. Out of 298 zones, 68 zones are found to be high demand zones, and the remaining zones are low demand zones for which the potential demand will be calculated and prioritise transit improvement.

An ensemble tree-based gradient boosting super machine learning model is trained on high demand zones. The model includes important variables of the zones' demographic, economic, and geographic attributes. The final model has an r-square of 0.878. the MAE of the model is found to be 1118 trips/zone using the one-left k-fold validation technique. The same model is then applied to low demand zones.

Given the modelling is on the same geographical region, in this case on the Brisbane network, it is expected that errors in the transferability of the modelling from high ridership zones to the low ridership zones should not be significant. Moreover, this study ranks the zones for which it is fair to assume that the relative errors should not have much impact. In the absence of the ground truth, it is hard to quantify the errors, and the proposed methodology provides a good tool to practitioners to rank the zones with high potential demand. The total transit demand and corresponding potential demand for the low demand zones are predicted from the developed model, and the results are presented by employing a choropleth map. It is normalised by the zone's total area, giving potential demand per unit area.

Furthermore, all low demand zones are divided into four quantiles based on normalised potential demand and plotted, depicting lowest, low, medium, and high transit improvement areas. The results demonstrate that the proposed methodology can be effectively used to group the zones for transit priority. In addition, the results show that approximately 50% of zones do not intend to produce high trips, even if very high quality and quantity of transit service are provided. On the other hand, eleven zones lie in the high priority for transit improvement.

Future research may include the applications of developed methodology on other regions with census availability at a finer level so that the high demand zones' number is higher, enhancing the confidence in the model results. Besides, this study only considers the origin location (i.e., trip production of a zone) and the destination location (trip attraction) is ignored. Future research may include considering both the origin and destination of transit trips for potential demand estimation. The proposed methodology is generic. The models need to be recalibrated and revalidated if applied to other regions. This may change the explanatory variables due to the region's diversity of transit network, cultural, demographic, and socio-economic attributes. Besides, the results from potential demand estimation will provide a better picture of the network to the transit planners when combined with the transit supply. Thus, these results will make more impact when combined with transit supply.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Aljoufie, M. 2014. "Spatial analysis of the potential demand for public transport in the city of Jeddah, Saudi Arabia." *Urban Transport XX* 1:113-23.

Alsger, Azalden, Behrang Assemi, Mahmoud Mesbah, and Luis Ferreira. 2016. "Validating and improving public transport origin–destination estimation algorithm using smart card fare data." *Transportation Research Part C: Emerging Technologies* 68:490-506.

Barry, James, Robert Newhouser, Adam Rahbee, and Shermeen Sayeda. 2002. "Origin and Destination Estimation in New York City with Automated Fare System Data." *Transportation Research Record: Journal of the Transportation Research Board* 1817:183-7.

Boehmke, Bradley, and Brandon Greenwell. 2020. *The R Series, Hands-On Machine Learning with R*: CRC press.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1):5-32. doi: 10.1023/A:1010933404324.

Chen, Chao, Karl Petty, Alexander Skabardonis, Pravin Varaiya, and Zhanfeng Jia. 2001. "Freeway Performance Measurement System: Mining Loop Detector Data." 1748 (1):96-102.

Friedman, Jerome H. 2001. "Greedy function approximation: a gradient boosting machine." *Annals of statistics*:1189-232.

Gaudry, Marc J. I., and Michael J. Wills. 1978. "Estimating the functional form of travel demand models." *Transportation Research* 12 (4):257-89.

Greenwell, Brandon, Bradley Boehmke, Jay Cunningham, GBM Developers, and Maintainer Brandon Greenwell. 2020. "Package 'gbm'." In.: R.

Han, Gain, and Keemin Sohn. 2016. "Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model." *Transportation Research Part B: Methodological* 83:121-35.

Hussain, Etikaf, Krishna Nikhil Sumantha Behara, Ashish Bhaskar, and Edward Chung. 2021. "Framework to quantify the differences in multi-modal travel demand: Case study on Brisbane network." *IEEE Transactions on Intelligent Transportation Systems*.

Hussain, Etikaf, Ashish Bhaskar, and Edward Chung. 2021a. "A novel origin destination based transit supply index: Exploiting the opportunities with big transit data." *Journal of Transport Geography* 93:103040.

———. 2021b. "Transit OD matrix estimation using smartcard data: Recent developments and future research challenges." *Transportation Research Part C: Emerging Technologies* 125:103044.

Kusakabe, Takahiko, and Yasuo Asakura. 2014. "Behavioural data mining of transit smart card data: A data fusion approach." *Transportation Research Part C: Emerging Technologies* 46:179-91.

Natekin, Alexey, and Alois Knoll. 2013. "Gradient boosting machines, a tutorial." *Frontiers in Neurorobotics* 7 (21). doi: 10.3389/fnbot.2013.00021.

Naveh, Kianossh Soltani, and Jiwon Kim. 2019. "Urban Trajectory Analytics: Day-of-Week Movement Pattern Mining Using Tensor Factorization." *IEEE Transactions on Intelligent Transportation Systems* 20(7):2540-9.

Nurdden, Abdullah, RAOK Rahmat, and Amiruddin Ismail. 2007. "Effect of transportation policies on modal shift from private car to public transport in Malaysia." *Journal of applied Sciences* 7 (7):1013-8.

Nusser, R., and R. M. Pelz. 2000. Bluetooth-based wireless connectivity in an automotive environment. Paper presented at the 52nd Vehicular Technology Conference Fall 2000. IEEE VTS Fall VTC2000., 24-28 Sept.

Olleczek, Thomas, Liang Yu, Joseph Kang Lee, Oliver Senn, Carlo Ratti, and Patrick Jaillet. 2014. Proceedings of the 2014 International Conference on Big Data Science and Computing, Beijing, China.

Regt, Karin de, Oded Cats, Niels Van Oort, and Hans van Lint. 2017. "Investigating Potential Transit Ridership by Fusing Smartcard and Global System for Mobile Communications Data." *Transport Research Record: Journal of the Transportation Research Board* 2652 (1):50-8.

Trépanier, Martin, Catherine Morency, and Bruno Agard. 2009. "Calculation of transit performance measures using smartcard data." *Journal of Public Transportation* 12 (1):5.

White, J., and I. Wells. 2002. “Extracting origin destination information from mobile phone data.” *IET Conference Proceedings*:30-4.

Yao, Xiaobai. 2007. “Where are public transit needed–Examining potential demand for public transit for commuting trips.” *Computers, Environment and Urban Systems* 31 (5):535-50.

Yap, Menno, Oded Cats, and Bart van Arem. 2020. “Crowding valuation in urban tram and bus transportation based on smart card data.” *Transportmetrica A: Transport Science* 16 (1):23-42.