TRB Annual Meeting

A NEW PUBLIC TRANSPORT ACCESSIBILITY ESTIMATION METHOD USING COMMUNITY PUBLIC TRANSPORT CENSUS AND SMART CARD DATA

--Manuscript Draft--

Full Title:	A NEW PUBLIC TRANSPORT ACCESSIBILITY ESTIMATION METHOD USING COMMUNITY PUBLIC TRANSPORT CENSUS AND SMART CARD DATA
Abstract:	Bus stops and train stations play important roles in connecting passengers to the public transport infrastructure. Most studies in the literature consider the accessibility of the bus stops and the train stations as performance indicators of the supply side and utilize the physical environment and the existing infrastructure to estimate the accessibility indexes. This paper treats accessibility as a performance index of both the supply and the demand side and proposes a new searching algorithm to calculate the catchment areas that meet the current realistic public transport demand. Firstly, we formulate the tap-on counts as a function of the demand in each community and the radius of the coverage area. Secondly, we then estimate the bus stop coverage area which identifies the realistic attraction areas where the bus users come from. Thirdly, the bus stop results show that 95.7% of all bus stops in the study area have a coverage area radius of fewer than the standard 400 meters catchment radius, meaning that the current design of the bus network accessibility is currently overestimated and too sparse in comparison to what the travellers currently need. We also prove that the coverage area using our method is more accurate by 42.01% in comparison the traditional 400-meters coverage area. Lastly, for bus stops with a catchment area greater than 400 meters, we further classify them based on the multi-modal interchange behaviors at the bus stop. Similarly, we show that 33.3% of all the train station accessibility is currently overestimated.
Manuscript Classifications:	Public Transportation; Bus Transit Systems AP050; Public Transportation; Planning and Development AP025; Accessibility; GIS; Transformative Trends in Transit Data AP090; Smart card data
Manuscript Number:	TRBAM-23-02714
Article Type:	Presentation
Order of Authors:	Tuo Mao
	Adriana-Simona Mihaita
	Dong Zhao
	Yuming Ou
	Seunghyeon Lee
	Fang Chen
Additional Information:	
Question	**Response**
The total word count limit is 7500 words including tables. Each table equals 250 words and must be included in your count. Papers exceeding the word limit may be rejected. My word count is:	7157
Is your submission in response to a Call for Papers? (This is not required and will not affect your likelihood of publication.)	No

A NEW PUBLIC TRANSPORT ACCESSIBILITY ESTIMATION METHOD USING COMMUNITY PUBLIC TRANSPORT CENSUS AND SMART CARD DATA

Tuo Mao, Ph.D., Corresponding Author
tuo.mao@uts.edu.au

Adriana-Simona Mihaita, Ph.D.
adriana-simona.mihaita@uts.edu.au

Dong Zhao, Ph.D Candidate.
dong.zhao@student.uts.edu.au

Yuming Ou, Ph.D.
yuming.ou@uts.edu.au

Suenghyeon Lee, Ph.D.
seunghyeon.lee@uos.ac.kr

Fang Chen, Ph.D.
fang.chen@uts.edu.au

Word Count: 6407 words + 3 table(s) × 250 = 7157 words

Submission Date: August 1, 2022

ABSTRACT

Bus stops and train stations play important roles in connecting passengers to the public transport infrastructure. Most studies in the literature so far consider the accessibility of the bus stops and the train stations as performance indicators of the supply side and utilize the physical environment and the existing infrastructure to estimate the accessibility indexes. This paper treats accessibility as a performance index of both the supply and the demand side and proposes a new searching algorithm to calculate the ideal catchment areas that meet the current and more realistic public transport demand. Firstly, by utilizing the smart card tap-on counts and the community survey data, we formulate the tap-on counts as a function of the demand in each community and the radius of the coverage area. Secondly, by using this new approach, we then estimate the actual bus stop coverage area which allows us to identify the realistic attraction areas where indeed the bus users come from. Thirdly, the bus stop results show that 95.7% of all bus stops in the study area have a coverage area radius of fewer than the standard 400 meters catchment radius fixed by default by traffic agencies, meaning that the current design of the bus network accessibility is currently overestimated and too sparse in comparison to what the travellers currently need. We also prove that the coverage area using our method is more accurate by 42.01% in comparison the traditional 400-meters coverage area. Lastly, for bus stops with a catchment area greater than 400 meters, we further classify them based on the multi-modal interchange that users need to adopt at the bus stop. Similarly, we show that 33.3% of all the train station accessibility is currently overestimated.

Keywords: Public transport accessibility, Coverage area, Catchment area, public transport planning.

INTRODUCTION

Buses and trains are major components of the public transport system in most cities, so the accessibility of the bus stops and train stations is an important factor which can affect the traveller's mode choice (*1*). Accessibility of a transit stop can be quantified as the time or the distance of walking, riding a bike or driving a car to the nearest transit stop (*2*). Traditionally, the accessibility standard of most transit systems is considered to be 400 meters which is equivalent to 5 minutes of walking (*3, 4*).

As an important indicator of the transit stop accessibility, the coverage area is firstly estimated using a circular buffer with a radius of the access threshold around the transit stop (*5*). Later in 2008, authors in (*6*) evaluated the stop location considering the population living within the coverage area. The traditional way of treating the coverage area as a circular buffer tends to overestimate the stop coverage because the travellers will walk on the pedestrian network instead of walking directly to the bus stop; therefore other authors in (*7, 8*) estimated the actual coverage area by using the actual pedestrian network, which assumes that passengers will travel directly to the nearest pedestrian link and then use the pedestrian network to access the bus stop. As the Geography Information System (GIS) became more mature, research based on different land use patterns and population density along the bus line coverage area was developed around the Gold Coast in Australia (*9*) as well as in several areas of China (*10*). The bus stop accessibility was associated with the bus stop redundancy in (*11*) and with the overlapping service in (*12*). In 2000, an isochrone was used to represent the coverage area of the bus network (*13*). But more case studies were conducted in recent years by using GIS to estimate the bus stop accessibility (*14, 15*) and make further comparison with regards to old versus new catchment areas algorithms. Currie (*16*) compared the public transport demand and the public transport provision in Hobart, Australia to identify the gap, but there is no discussion about the accessibility because the major contribution is to identify the communities with poor and nonexistent public transport services. While pure geographic tools can be used as baseline for modelling the public transport accessibility, we highly believe that complementary algorithms which consider multiple factors should be developed in parallel with the purpose of having a more granular and detailed insight into area catchment planning and a true accessibility metrics estimation.

Challenges and contributions

Existing research about accessibility focuses more on the physical aspect of the transit stop and defines the coverage area as an accessible area inside of the physical network. Moreover, the models are not dynamic, as do not evolve with the population and demand increase in the city; this means they are often made at the beginning of planning a new bus or train service route and not updated once they are being put into practice. Although the accessibility is proposed as a performance indicator of a public transport service, the evaluation of how much the current public transport design meets the realistic accessibility criteria needs to be tailored against the current public transport demand; and this tailoring represents a unique challenge especially when it needs to adapt dynamically and uniquely to each public transit area. Moreover, the travel demand is increasing on a yearly basis, and is highly correlated to the residency and urban living conditions, so public transport services need to keep with with the ever-increasing (or decreasing) needs of urban areas in a dynamic approach.

Our paper proposes a different approach of evaluating the realistic accessibility areas of the public transport users in a demand and supply model, that could lead to a service redesign and optimization. Firstly, the proposed methodology connects the tap-on counts to the current community demand (users). Both the tap-on counts and the community demand are mapped together via a dynamic geographical search and mapping for an area catchment estimation. Secondly, we propose an incremental searching algorithm to match: a) the catchment areas of the public transport stops, b) the demand layer (where the community users live), and c) the supply layer (transit stops associated with their tap-on counts). The output of our proposed algorithm is an estimated bus/train coverage area with realistic passenger attraction indexes. This algorithm reveals whether the current public transport design is well adapted to meet the most recent community demand or not. For example, in the city of Sydney the accessibility threshold is set at 400-meters but this has been established several years back in the late 2000 when the population count and the urban density was much smaller (3.8M people) than in recent years (5.05M inhabitants in 2021).

Lastly, our paper proposes a new approach of modelling the travel demand across multiple public transport modes in the environment of smart cities, which make the connections between the traffic data and other data sources (in this case, the population census data). As a starting point, this paper demonstrate a searching algorithm between different types of geographic data which can be later extended into an automatic feature selection using machine learning in a second stage of development.

METHODOLOGY

Study Area

We select the M2 motorway corridor in Sydney as our study area (see Fig. 1). This corridor connects the Sydney CBD and the North-West part of the city and plays an important role in both the daily commuting and the cargo transportation. The M2 Motorway corridor is selected due to its wide coverage of multi-modal transportation systems including private vehicles, buses, and train. The selected area contains part of the T9 northern train line, part of the Metro line, and part of the T1 Western train line. Besides, the bus lines along the M2 Motorway carry a large amount of travellers and commuters towards the business area around the Maquarie centre and further more towards the city centre.

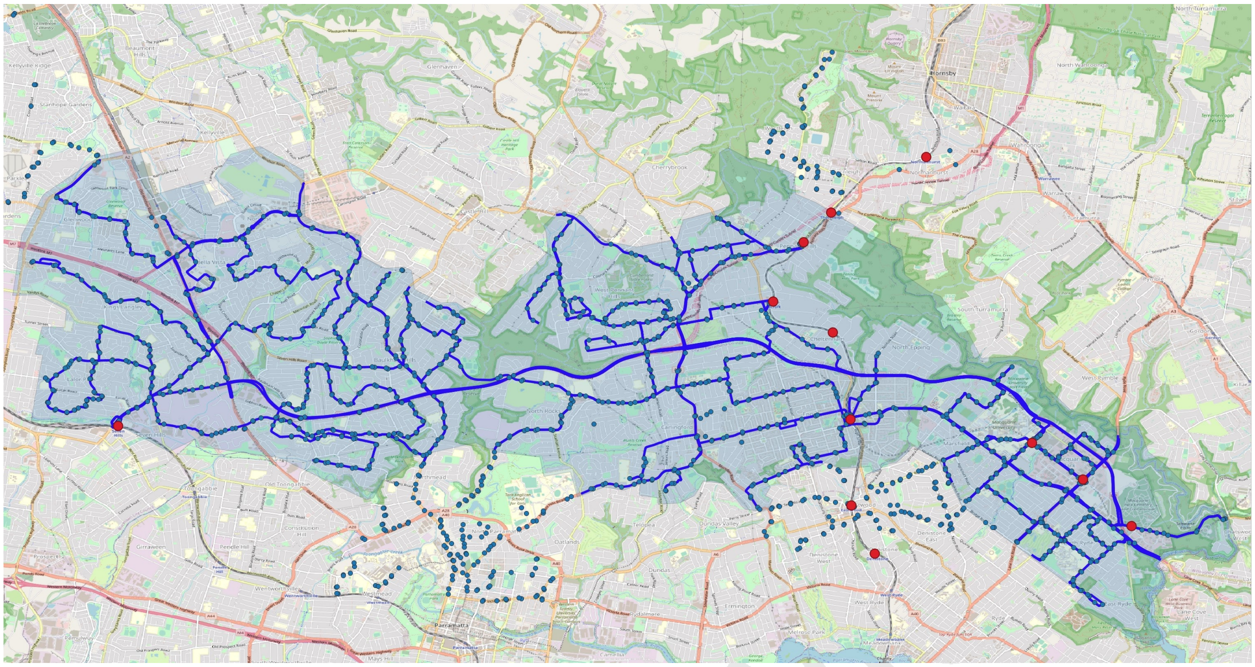


FIGURE 1: Investigated area of the M2 Motorway corridor in Sydney North-West.

In Fig. 1, the blue lines are the bus lines running through the study area. The highlighted blue-shadowed area is the study area spreading from the West to the North part of the city. The blue dots are the bus stops. The big red points are the train stations. All mathematical notations we use are provided in Table 1 below.

Problem Formulation

Bus network: Normally, the bus stop coverage area is defined as the maximal area inside which all passengers will and are capable of using a bus stop. Traditionally, the bus stop coverage area was represented as a circular area that surrounds the bus stop within a fixed radius. In this study, we use a similar definition of the bus stop coverage area to remain consistent with the main notations in the literature. For a bus stop P_i, its coverage area is defined as a circle area S_i with a radius denoted R_i. The outline of the area is C_i. Therefore, the coverage area can be quantified as the area of the circle which is expressed in Eq. (1); in this equation we also define the function to calculate the area of a circle/polygon as $S()$.

$$S_i = S(C_i) = \pi R_i^2 \tag{1}$$

We further define the average daily tap-on counts of the bus stop P_i as X_i. We define as well a boundary as a set M which contains K communities lieing in the study area, and each community k can be represented as a polygon m_k ($\forall k \in \{1,2,..K\}$). The community boundary set M is further defined in Eq. (2):

$$M = \{m_1, m_2, ..., m_k, ..., m_K\} \tag{2}$$

For each community k, we calculate the area of the community ($S(m_k)$), the daily average number of bus users (B_k) and the daily average number of train users (T_k) in the community.

Intuitively, the total number of bus stop P_i users (Y_i) are users within a circle area C_i. We further define the function $f()$ to retrieve the bus users within a circle/polygon which leads us to

TABLE 1: Notations in use for accessibility modelling.

Variable	**Definition**
i	Index of a bus stop
P_i	The bus stop i
S_i	The circular coverage area of a bus stop i
C_i	The outline of the circular coverage area of bus stop i
R_i	The radius of the circular coverage area of bus stop i
X_i	The daily average tap-on counts at a bus stop i
Y_i	The total number of users of a bus stop i
$f()$	The function extracting the number of bus users within a circle/polygon area
i'	Index of a train station
$P_{i'}$	The train station i'
$S_{i'}$	The circular coverage area of a train station i'
$C_{i'}$	The outline of the circular coverage area of a train station i'
$R_{i'}$	The radius of the circular coverage area of a train station i'
$X_{i'}$	The daily average tap-on counts of a train station i'
$f‘()$	The function estimating the number of train users within a circle/polygon area
$S()$	The function which calculates the area of a circle/polygon (e.g. $S(C_i)$ and $S(m_k)$)
M	The set of all communities boundaries
K	The total number of communities boundaries
k	Index of a community
m_k	The polygon of a community k boundary
B_k	The number of bus users in a community k
T_k	The number of train users in a community k
$\cap$	The function calculating the intersection area between two polygons (e.g. $C_i \cap m_k$)

the following equation:

$$Y_i = \sum_{k=1}^{K} f(C_i \cap m_k) \tag{3}$$

Equation (3) shows that the total number of bus users (Y_i) using a bus stop P_i is generated by the bus users within the circle C_i in all the community polygons m_k which intersect with the circle C_i. In Eq. (3), $C_i \cap m_k$ is the intersection polygon of the circle C_i and the community boundary polygon m_k. By calculating $f(C_i \cap m_k)$, we obtain the number of bus users within the intersected polygon. By summing up all the bus users of all the intersected areas, we get the total number of bus users in our proposed bus area. Detailed example of calculating the train users is described in the next section 3.3 and Figure 2.

Train network: Similarly to the Bus Stop Coverage Area, for a train station $P_{i'}$, its coverage area is defined as a circle area $S_{i'}$ whose radius is $R_{i'}$. The outline of the area is $C_{i'}$. Therefore, Eq. (1) is also applicable to this case as well. We also denote the average daily tap-on count of the train station $P_{i'}$ as $X_{i'}$.

Intuitively, the total number of train station $P_{i'}$ users ($Y_{i'}$) represents the users encapsulated within the circle $C_{i'}$. We further define the function f' to retrieve the train users within a circle/polygon and express this by using the following equation:

$$Y_{i'} = \sum_{k=1}^{K} f'(C_{i'} \cap m_k) \tag{4}$$

Detailed example is described in the next section 3.3 and Figure 2.

Assumptions

In order to ease the notations and modelling steps, we make a few assumptions detailed in the following.

A1: Firstly, we treat bus stops in two opposite directions which are located at the same geographical coordinates as being one single bus or train stop. This assumption is majorly because in the current public transport data formatting, the bus stops in two opposite directions always has the same name and the tap-on counts at a station are merged when being reported, regardless of the direction of travel of any public transport mode passing that stop. From the smart card data, we observe that the tap-off counts are also aggregated using the stop names which re-enforces this assumption.

A2: Further, we assume that each bus user taps ON only once per day at their chosen bus stop to start the daily travels. This means that the bus users route only contains the chosen bus stop once per trip. In addition, if the bus user travels the same route back to home (the bus user travels a round-trip), the bus users should tap off once at the same bus stop per day.

A3: In addition, if there is no interchanging from bus to bus or from train to bus, the tap-on counts of an investigated bus stop should be equal to the total number of bus users within its coverage area. This means that for those bus stops that only serve local demands, the tap-on counts should be conservative with the local bus users. This assumption can be eliminated if we have access to more detailed smart card data which contains the transfer information of each user.

Under assumptions A1, A2, and A3, we can further infer that the tap-on counts of a bus stop P_i (X_i) equals to the total number of bus users Y_i as shown in the Eq. (5):

$$X_i = Y_i \tag{5}$$

Similarly under the same assumptions, we can infer that the tap-on counts of a train station $P_{i'}$ ($T_{i'}$) equals to the total number of train users $Y_{i'}$ which is shown in Eq. (6):

$$X_{i'} = Y_{i'} \tag{6}$$

A4: Lastly, we assume that the bus users are evenly distributed in the community. This assumption allows us to link the bus users with the areas of the polygons. Therefore, we can unfold the function f and f' under this assumption. We therefore expand the function $f(C_i \cap m_k)$ in Eq. (7):

$$f(C_i \cap m_k) = B_k / S(m_k) \times (\alpha_{ik} \times \pi R_i^2) \tag{7}$$

where α_{ik} is the ratio of the intersected area ($S(C_i \cap m_k)$) over the area of the circle $S(C_i)$ which is further shown in Eq. (8):

$$\alpha_{ik} = S(C_i \cap m_k)/S(C_i) = S(C_i \cap m_k)/(\pi R_i^2) \tag{8}$$

We make the observation that α_{ik} is dependent on the radius R_i.

Similarly, we expand the function $f'(C_{i'} \cap m_k)$ in the following Eq. (9):

$$f'(C_{i'} \cap m_k) = T_k / S(m_k) \times (\alpha_{i'k} \times \pi R_{i'}^2) \tag{9}$$

where

$$\alpha_{i'k} = S(C_{i'} \cap m_k)/S(C_{i'}) = S(C_{i'} \cap m_k)/(\pi R_{i'}^2) \tag{10}$$

We also make the observation that $\alpha_{i'k}$ is dependent on the radius $R_{i'}$. As a summary, by combining all the equations listed above, we can link the bus stop tap-on counts X_i and the bus stop coverage area radius R_i by using Eq. (11):

$$X_i = Y_i = \sum_{k=1}^{K} B_k/S(m_k) \times (\alpha_{ik} \times \pi R_i^2) \tag{11}$$

We can also link the tap-on counts of a train station $X_{i'}$ and the train station coverage area radius $R_{i'}$ by using Eq. (12):

$$X_{i'} = Y_{i'} = \sum_{k=1}^{K} T_k/S(m_k) \times (\alpha_{i'k} \times \pi R_{i'}^2) \tag{12}$$

For easing the understanding, we plot Fig. 2 which demonstrates the Eq. (12) in a map overview, where the red dot represents the train station and its coverage area, intersected by the community local areas (see polygon $S(m_k)$ and the total number of people travelling from that catchment area to the train station (297.24). The number (297.24) is a decimal number because it's averaged by day using multiple-day data. The same methodology is applied for bus stops as well.

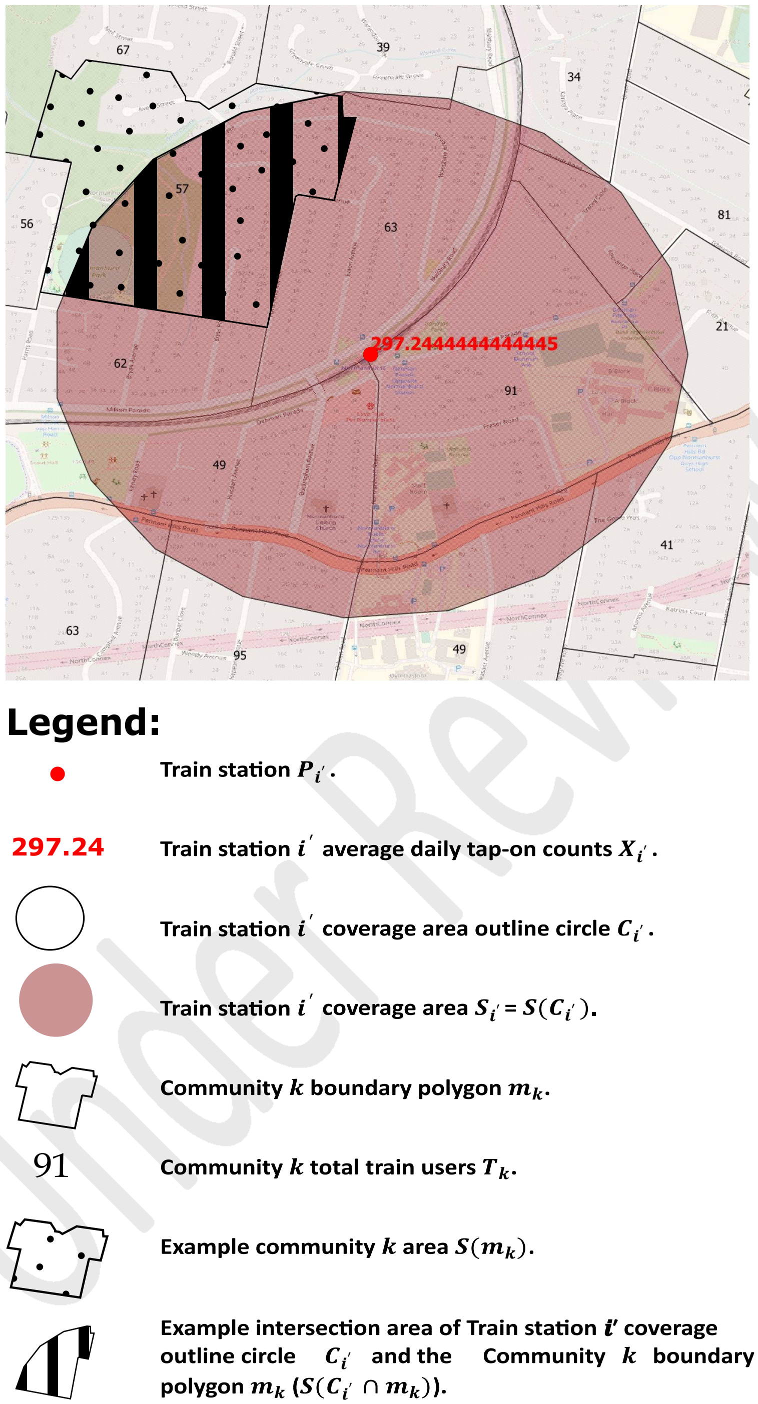


FIGURE 2: An example on how to calculate the tap-on counts and the train station coverage area.

In the next step, we will use the bus stop/train station tap-on counts and the community bus/train users to estimate the bus stop/train station coverage area by applying Eq. (11) and Eq. (12).

Accessibility Search Algorithm

Based on the problem formulation provided in Section 3.2, we propose a searching algorithm to estimate the ideal bus stop/train station coverage area based on ongoing travel demand and

Algorithm 1: Accessibility Search Algorithm

Input: $myStop, Communities, step$
Output: $myStop$
$j = 0$; `//j is a flag.`
while $j = 0$ *or* $mystop.estimated\_bus\_user \leq mystop.daily\_tap\_on$ **do**
  $R = (i+1) * step$; `//Calculate current search radius`
  $X, Y = P_i.x, P_i.y$; `//Get the center point of the bus stop`
  $Get\ C_i\ using\ the\ R\ and\ X, Y$; `//Get the outline circle`
  $intersected\_communities = \{community\ for\ community.intersects(C_i) == True\ in\ Communities\}$; `//Get the intersected communities`
  $Get\ \_the\ \_intersected\_communities[interc\_area]$; `//Get the intersection area`
  $Get\ \_the\ \_intersected\_communities[interc\_bus\_user]$; `//Using the Eq. (7) to calculate the` *interc_bus_user*
  $mystop[estimated\_bus\_user] = sum(intersected\_communities[interc\_bus\_user])$;
  $mystop[Radius] = R$;
  $j = j + 1$;
end
return $myStop$;
; `//The outcome is a` *mystop* `data-frame with the updated` *radius* `and the` *estimated_bus_user* `column.`

population catchment areas. The searching algorithm is provided in Alg. 1 and detailed in the following. In the algorithm, we start with a 'while' loop whose stop condition is that: the estimated bus users in the tested radius reaches (equals or is greater than) the daily tap-on counts. In the 'while' loop, we take mini-steps to increment the tested radius and calculated the number of bus users inside the tested coverage area. The main entries in the algorithm are:

a) *myStop* which is a data-frame containing all the known information of a bus stop. The information contains X(latitude), Y(longitude), *daily_tap_on* (daily average tap-on counts), *estimated_bus_user* (the final estimated number of bus users which is set to be 0 at the first step), and *radius* (which is the final radius of the bus stop coverage area, whose value is set to be 0 initially).

b) *Communities* is a data-frame which contains all the known information of all communities in the investigated area. The columns contain: S (the area of the corresponding community), *total_bus_users* (total number of bus users), *total_train_users* (total number of train users), and *geometry* (the polygon of the community boundary).

c) we need to define the minimum stepping unit of the searching area *step* (for example, 50 meters) before running the algorithm. In the Algorithm, the step is used to increment our searching radius after each 'while' loop until we reach the stop condition. The minimum stepping unit can also be treated as the accuracy of the final output radius. We therefore set the starting searching radius as the minimum stepping unit.

Searching Accuracy and Errors

In Section 3.4, we defined the minimum stepping unit of the searching algorithm *step* which can also be treated as the control parameter of the accuracy. Therefore, to achieve a higher accuracy, we need to select a smaller *step*.

The difference between the estimated bus user (*estimated_bus_user*) and the daily average tap-on count (*daily_tap_on*) is defined as the raw error (sometimes called a residual in various works).

In our study, we conducted a sensitivity test and experimented with 50 meters, 10 meters, and 5 meters as the minimum stepping unit of the searching algorithm (*step*). In addition, we calculated the Mean Absolute Percentage Error (MAPE)(*17*), the Root Mean Squared Error (RMSE)(*18*), and the Max Error as our performance metrics of the errors - see definitions below:

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}([daily\_tap\_on] - [estimated\_bus\_user])^2} \tag{13}$$

where n represents the public transport stop and N is the total number of stops in the network.

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{[daily\_tap\_on] - [estimated\_bus\_user]}{[estimated\_bus\_user]}\right| \tag{14}$$

where t is the prediction id and T represents the total number of predictions.

The max error computes the maximum residual error, which captures the worst case error between the predicted value ($[estimated\_bus\_user]$) and the true value ($[daily\_tap\_on]$). The max error is calculated using the following equation:

$$MaxError([daily\_tap\_on],[estimated\_bus\_user]) = max(|[daily\_tap\_on] - [estimated\_bus\_user]|) \tag{15}$$

Table 2 shows the results of the minimum stepping unit and the errors. By reducing the *step* from 50 meters to 5 meters, all of the observed errors are reduced accordingly (for example MAPE is reduced from 77.27% when choosing step=50m as compared to 6.64% when choosing step=5m; similarly for RMSE and the Max Error parameters, which indicate that step=5m is the concluded best value of the *step* variable in our searching algorithms. Therefore all further results are obtained with an increment of 5 meters radius, giving us maximum granularity in our searching approach.

TABLE 2: Sensitivity testing for finding the minimum stepping unit and the corresponding Error results.

step (meter)	MAPE(%)	RMSE	Max Error
50	77.27	12.28	125.39
10	13.46	5.02	20.11
5	6.64	3.51	11.42

Data

In this study, one of the data sets we use is the census data from the Australian Bureau of Statistics (ABS) - 2016 edition as the last being available for Australia. We use New South Wales (NSW) as our investigated state and Statistic Areas 1 (SA1) as our zoning category. The reason why we use SA1 as our community boundary is that SA1 is the smallest and the finest grained area boundary. From this census, we use the following sources: the "Place of Enumeration Profile" table to search for the "Method of Travel to Work" data sheet. We use the "Bus only", "Train only", "Trains and another mode", and "Bus and another mode" columns to calculate the total number of bus users and the train users inside an SA1 area.

In complementary with this data set, we also utilize the smart-card data to capture the daily average tap-on counts for each bus stop and for each train station. The smart card data contains the tap-on and the tap-off actions for a whole year. We also aggregated the data to estimate the daily average tap-on counts.

RESULTS

We run the searching algorithm for every bus stop and every train station in our study area. Overall, we estimated the area to encompass 1432 bus stops and 12 train stations.

Bus Stop Results

Bus Stop Overall Results

We have investigated all the 1432 bus stops and obtained the optimal radius of the bus stop coverage area. Fig. 3 demonstrates the histogram of the estimated bus stop coverage area radius. We observe

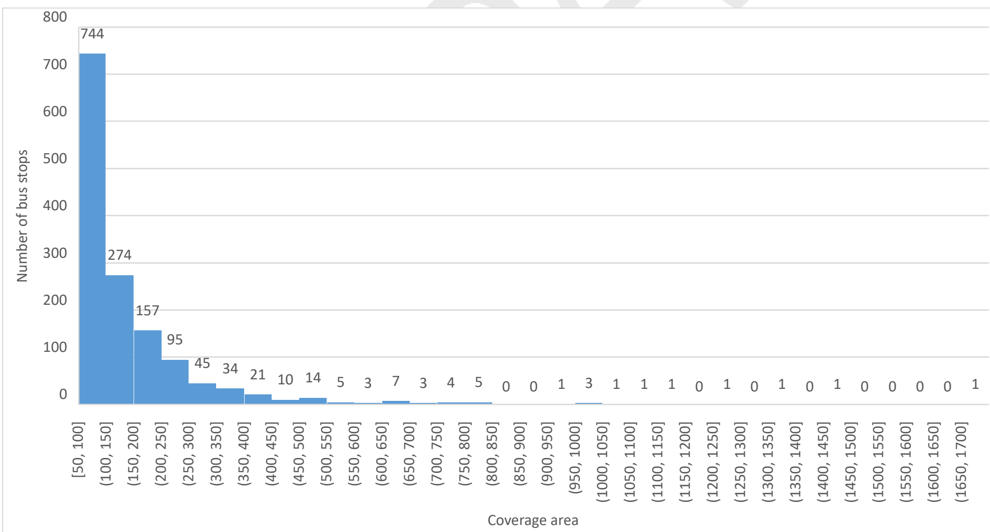
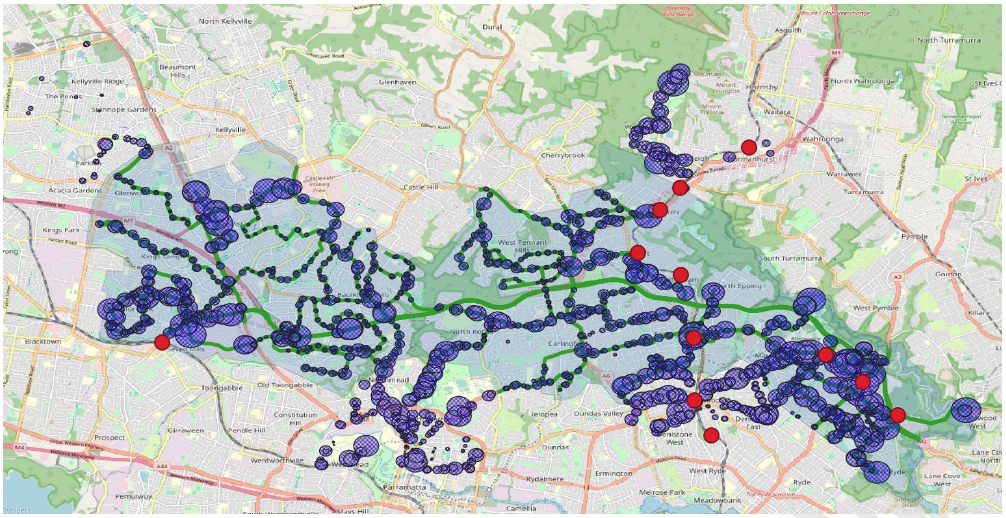
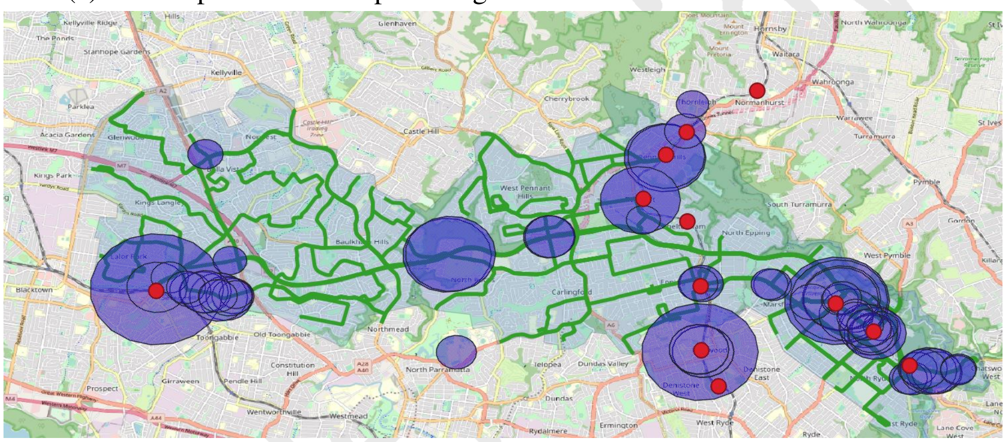


FIGURE 3: Bus stop coverage area radius histogram.

that in total there are 1370 bus stops (which represent 95.7% of all bus stops) whose optimal coverage area radius is less than 400 meters and there are 62 bus stops (which represent 4.3% of all bus stops) whose coverage area radius is more than 400 meters. Interestingly, 744 bus stops have an optimal radius area of less than 100 meters which comes as a significant revelation when compared to the standard adopted area of 400 meters (which seems to be significantly over-estimated given on the current travel needs). Furthermore, Fig. 4a shows the map of the bus stop coverage areas which are less than 400 meters and Fig. 4b shows the map of the bus stop coverage areas which are greater than 400 meters (the latter are identified as major interchange bus stops, or more specifically periphery stops attracting population from suburbs with no train stations for example).



(a) Geo map of the bus stop coverage areas which are less than 400 meters.



(b) Geo map of the bus stop coverage areas which are greater than 400 meters.

FIGURE 4: Geo map of the bus stop coverage areas.

Bus Stop Classification

To extend the current analysis, we further divided the bus stops into four classes:

(1) Small coverage area stops: they have a coverage area radius of less than 400 meters. These bus stops always lie in the areas serviced only by bus routes and present no or low train users within their coverage areas.

(2) Big coverage area stops - bus to bus interchange stops: they are used/shared by multiple bus routes. These bus stops always lie at the crossing intersections of two/multiple bus routes. Theoretically speaking, the radius of the stop coverage areas in this class are over-estimated in our study since our methodology is only applicable for those bus stops with a low proportion of the interchanging behaviour (see the Assumption 3). The same trends show in other classes which indicate a large proportion of interchange travellers in the bus stops.

(3) Big coverage area stops - train to bus interchange stops: they are used/shared by one/multiple bus routes and one/multiple train lines. These bus stops always lie near the train stations since a large number of travellers get off at the train station and transfer to a bus route in order to reach their final destinations in outer suburbs.

(4) Big coverage area stops - low bus user stops: whose surrounding area contains a small number of bus users. These bus stops lie in areas with either a low population density or with a low bus users density. The low bus users density area may happen where travellers choose to walk to the train station and take the train to their destinations.

Train Station Results

Overall, we investigated 12 train stations captured in our M2 area and Table 3 shows the detailed results where the Daily Tap-on counts are provided by the smart tap-on data while the estimated train users are the output of the searching algorithms we have proposed, together with the optimal radius that we have identified for each train station. Fig. 5 shows a map view of the train station coverage areas. We observe that our search algorithm manages to match closely the real data, making it reliable for analysing the ideal catchment areas for citizens.

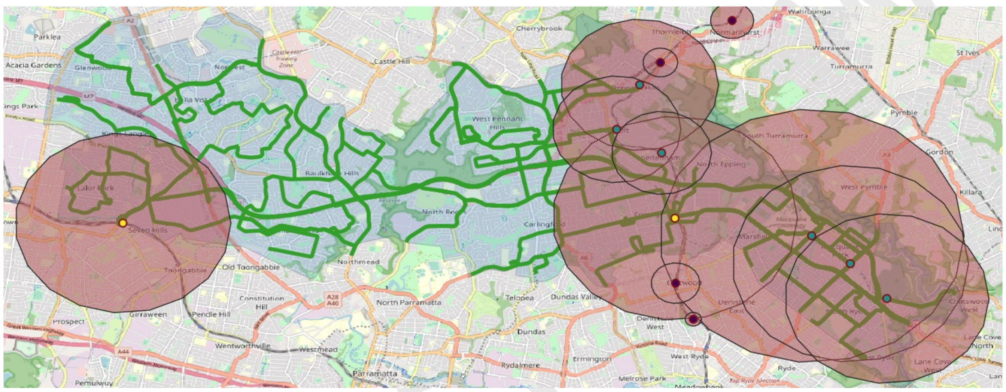


FIGURE 5: Train stop coverage area along the M2 case study network.

TABLE 3: The train stations coverage area radius results.

Stop Name	Daily tap-on counts	Estimated train users	Radius (meter)
Denistone Station	34.69	34.97	200
Thornleigh Station	243.08	244.84	425
Normanhurst Station	297.24	301.16	530
Eastwood Station	808.23	809.10	595
Cheltenham Station	970.42	972.37	1220
Beecroft Station	1639.41	1645.69	1570
Pennant Hills Station	2630.24	2630.66	1930
North Ryde Station	1804.34	1812.74	2495
Seven Hills Station	5055.70	5069.70	2640
Macquarie Park Station	3378.12	3389.21	2890
Epping Station	9212.68	9216.19	2995
Macquarie University Station	8900.21	8925.42	3725

Train Station Classification

We next classify the train stations into 3 classes:

(1) Small train stations that only serve the local train users: they are small train stations which are used by only one or a low number of train routes. Denistone Station, Thornleigh Station, Normanhurst Station, and Eastwood Station are classified in this class. They only serve the T9 Northen line and a small number of local train users (less than 900) within a radius of fewer than 600 meters.

(2) Big train stations that serve bus-to-train transfer users: they are big train stations which are used by only one train route. In this case, a traveller will take a bus to the train station and then take the train to their destinations. The Beecroft, Cheltenham, North Ryde, Macquarie Park, and Macquarie Park Station are classified in this class. The Beecroft Station and The Cheltenham Station serve the T9 Northern line exiting the city to the North, only whereas the North Ryde, Macquarie Park and the Macquarie Park Stations serve the Metro line only to the CBD.

(3) Big train stations that serve train-to-train transfer users: they are big stations which are used by multiple train routes. Therefore travellers will transfer from one train line to another train line in order to reach their destination. The Epping and Seven Hills Stations are classified in this class. The Epping Station serves the T9 Northern Line and the Metro line. The Seven Hills Station servers the T1 Western line, the T1 Richmond line, and the T5 Richmond line.

Overall, the proposed algorithm in this paper allows a good classification of different trains or bus stops based on their functionality and inter-connectivity; this aspect opens the possibility of a more advance coupling between our optimal catchment area finding together with machine learning modelling for an automatic station classification based on multiple features, extracted from either historical and available data sets or from topological aspects of the public transport network, as well as the main purpose for utilisation of each station. This represents a future direction that can open a better and automatic public transport planning as detailed in the last section of our paper.

Comparison to the standard radius method

We compare the coverage area between our method and the existing standard method. We use the commonly defined coverage area radius (which is fixed to 400 meters) as the comparison. We plot the coverage area of both methods in Fig. 6; in this figure, the red areas are the common 400 meters radius catchments and the blue areas are our proposed demand-driven radius. We observe that 1300 bus stops have significantly smaller coverage areas than the 400-meter common coverage area while only 62 stops have bigger coverage areas than the 400 meter common coverage area. This indicates that multiple areas in the city will require a redesign with multiple new public transport stops that would meet the current travel demand. In addition, we calculated the union area of all the investigated bus stops for both methods. The standard 400 meter area coverage totals to an overestimated 116.12 square kilometers, but our proposed coverage areas estimated more accurately the catchment coverage at 67.344 square kilometers; this means that the standard estimation of the coverage area is traditionally overestimated by 42.01% in comparison to our proposed method which is more accurate to capture the needs of more exact bus users in the study area.

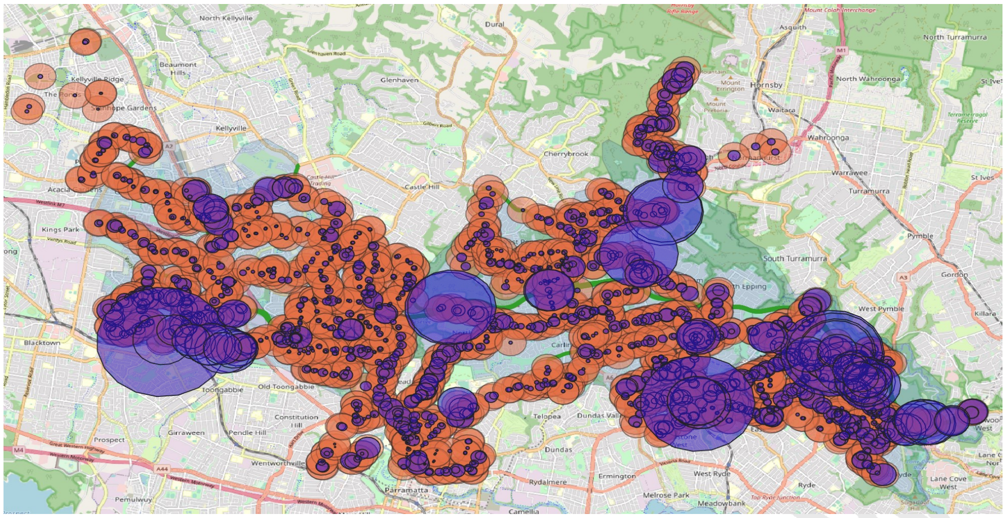


FIGURE 6: Comparison between the 400 meters standard radius (red) and our proposed radius (blue).

CONCLUDING REMARKS

To summarise the work contribution in this paper, we conclude that by discovering the relationship between the tap-on counts and the community demand, we were able to define the tap-on counts as a function of the transit stop coverage area radius and the community demand surrounding the transit stop. Based on this theory, we proposed a searching algorithm to estimate the transit stop coverage area radius. The minimum stepping unit of the searching algorithm is found to be 5 meters and we achieved a MAPE of 6.64%, an RMSE of 3.51, and a Max Error of 11.42 when comparing the estimated tap-on counts with the real smart card tap-on counts.

The bus stop results show that 95.7% of bus stops cover a circular buffer with a radius less than 400 meters which is the walking distance threshold defined in the previous state of art research. We classify the 95.7% of bus stops as the “Small coverage area stops”. For those bus stops whose coverage area radius is more than 400 meters (4.3 % of all bus stops), we further source the component of the transit stop users and classify them into “Bus to bus interchange stops”, "Train to bus interchange stops", and "Low bus user stops". Overall, the coverage area of the proposed area is 67.344 square kilometers comparing to 116.112 square kilometers for the 400 meter radius coverage area. We can observe that people tend to walk a smaller distance to access a public transport service than was originally expected when the traffic authorities designed the bus lines. This showcases a need for a better public transport planning with less generous assumptions on walking time and area coverage for attracting public transit users.

Similarly, train stop results show that four out of twelve train stations (33.33%) have a daily tap-on count lower than 900 and a coverage area radius lower than 600 meters. We classify these four stations as the “Small train stations that only serve the local train users”. The rest eight out of the 12 stations (66.67%) are classified into the “Big train stations that serve bus-to-train transfer users” and “Big train stations that serve train-to-train transfer users”.

In conclusion, in the investigated M2 area from Sydney, most people tend to walk a short distance to access the public transport service such as trains and buses. The walking distance to a train service is relatively longer than the bus service. The bus service coverage area is much less than expected. The train service coverage area per station is much bigger than the service coverage area per bus stop using the proposed method.

Limitations: The main limitation of our study is the data availability for larger area analysis in both time and space. Also, we only have two public transport modes which we would like to expand for larger areas. Several assumptions can be loosen in a future work such as the one when we assume that there is no interchanging from all of the investigated bus stops. In this case, if we have more detailed tap-on tap-off data, we can eliminate this assumption. Another limitation is that when calculating the coverage areas, there are some overlapping area between two successive bus stops. We have not consider the double counting of the passengers in the overlapping areas, which can be true in circumstances when people choose between multiple transport modes which can be equally distanced from their home based on their daily commuting trips and activities.

Future work: As an extension of current work, we will further refine the searching algorithm to consider the overlapping areas in the public transport coverage area identification.

Future work also lies in including more geographic data sets and construct a cross-data-set correlation identification model. This model may utilize the machine learning and AI techniques to automatically search for any correlation between the features.

REFERENCES

1. Holtzclaw, J., *Using residential patterns and transit to decrease auto dependence and costs*, Vol. 11. Natural Resources Defense Council San Francisco, 1994.
2. Murray, A. T. and X. Wu, Accessibility tradeoffs in public transit planning. *Journal of Geographical Systems*, Vol. 5, No. 1, 2003, pp. 93–107.
3. Edwards, J. D., *Transportation planning handbook*. Prentice Hall, 1992.
4. Ammons, D. and M. P. Benchmarks, *Assessing local Performance and establishing community standards*, 2014.
5. Murray, A. T., R. Davis, R. J. Stimson, and L. Ferreira, Public transportation access. *Transportation Research Part D: Transport and Environment*, Vol. 3, No. 5, 1998, pp. 319–328.
6. Foda, M. A. and A. O. Osman, A GIS approach to study the bus transit network accessibility: case study: the city of Alexandria. *Journal of Arab Academy for Science, Technology & Maritime Transport*, Vol. 34, No. 65, 2008.
7. Foda, M. A. and A. O. Osman, Using GIS for measuring transit stop accessibility considering actual pedestrian road network. *Journal of Public Transportation*, Vol. 13, No. 4, 2010, p. 2.
8. Salvo, G. and S. Sabatini, A GIS approach to evaluate bus stop accessibility. In *Proceedings*, 2005.
9. Yigitcanlar, T., N. Sipe, R. Evans, and M. Pitot, A GIS-based land use and public transport accessibility indexing model. *Australian planner*, Vol. 44, No. 3, 2007, pp. 30–37.
10. Liu, S. and X. Zhu, An integrated GIS approach to accessibility analysis. *Transactions in GIS*, Vol. 8, No. 1, 2004, pp. 45–62.
11. Delmelle, E. M., S. Li, and A. T. Murray, Identifying bus stop redundancy: A gis-based spatial optimization approach. *Computers, Environment and Urban Systems*, Vol. 36, No. 5, 2012, pp. 445–455.
12. Kimpel, T. J., K. J. Dueker, and A. M. El-Geneidy, Using GIS to measure the effect of overlapping service areas on passenger boardings at bus stops, 2007.
13. O'Sullivan, D., A. Morrison, and J. Shearer, Using desktop GIS for the investigation of accessibility by public transport: an isochrone approach. *International Journal of Geographical Information Science*, Vol. 14, No. 1, 2000, pp. 85–104.
14. Mohammed, G., *Evaluation of Existing Bus Stop Accessibility and Determination of Optimal Location For Bus Stops Using GIS Techniques: A Case Study of Yeka Sub City, Addis Ababa*. Ph.D. thesis, ASTU, 2020.
15. Tome, A., B. Santos, and C. Carvalheira, GIS-Based Transport Accessibility Analysis to Community Facilities in Mid-Sized Cities. In *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, 2019, Vol. 471, p. 062034.
16. Currie, G., Gap analysis of public transport needs: measuring spatial distribution of public transport needs and identifying gaps in the quality of public transport provision. *Transportation Research Record*, Vol. 1895, No. 1, 2004, pp. 137–146.
17. Armstrong, J. S. and F. Collopy, Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting*, Vol. 8, No. 1, 1992, pp. 69–80.
18. Chai, T. and R. R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development*, Vol. 7, No. 3, 2014, pp. 1247–1250.