Australasian Transport Research Forum 2022 Proceedings
28-30 September, Adelaide, Australia
Publication website: http://www.atrf.info

Investigation into public transport fare noninteractions using large-scale automatically collected data

Tianwei Yin[1], Neema Nassir[1], Joseph Leong[1], Egemen Tanin[2], Majid Sarvi[1]

[1]Department of Infrastructure Engineering, The University of Melbourne

[2] School of Computing and Information Systems, The University of Melbourne

Email for correspondence: tiyin@student.unimelb.edu.au

Abstract

Fare card data provides an unprecedented opportunity to monitor day-to-day variability of travel demand and its responses to service disruptions and special events. However, when passengers take public transport without interacting with the fare system, demand is usually underestimated, which may cause problems for performance measurement and revenue collection. This research aims to investigate the fare noninteractions phenomena of the tram network in Melbourne, Australia. According to a prior evaluation, only 37% of boarding passengers validate tickets. This study utilizes large-scale automatically collated data to measure fare noninteractions, including data collected by Automatic Passenger Counting (APC) and Automated Fare Collection (AFC) systems. Compared to previous studies with small samples of on-board surveys, it contributes to the state of the art as these high coverage data enable the study of the impact of different types of explanatory variables, including time periods, routes, stop location, travel demand variability, presence of an inspector on-board, etc. Moreover, a free service zone is located in Melbourne central business district where passengers are not required to validate tickets. We specifically investigate passengers' behavior at the boundary of a free service zone. Results show that fare noninteractions are lower for stops close to train stations, education facilities, stops that have been frequently inspected, and during the peak hours, but are higher for stops with large boarding flows, crowded services, evening periods and weekends. Importantly, conditioning on other variables, fare noninteractions at the boundary of the free service zone are higher in the morning peak but lower in the afternoon peak. The passenger flow diagram demonstrates the reason behind this may lie in the differences between purposes of trips. This investigation provides a starting point for proposing solutions to deal with the missing AFC data due to fare noninteractions.

1. Introduction

Increasing use of Automated Fare Collection (AFC) media in public transit systems around the world is a source of a massive amount of transit users' travel data. These data typically record passengers' origins, and in some cases destinations, as well as their time of travel. Moreover, these datasets can provide fare card identifiers for trips taken within the same day, and across days, weeks, months, or even years of observations. Recently, such rich datasets with high spatial and temporal resolution, collected over an extended period, have given researchers the opportunity to pursue many possible data-driven methodologies to better understand

passengers' travel behaviors and activity patterns (Munizaga & Palma 2012, Gordon et al. 2013, Nassir et al. 2011, 2015, 2019). However, when passengers take public transport without interacting the fare system, demand is usually underestimated and will cause problems for performance measurement and revenue collection. The noninteractions can be due to multiple reasons, including fare evasion, trips made with other payment methods, transfer trips, trips in the free service zone etc.

This paper presents the results of a quantitative investigation into fare noninteractions in the tram system of Melbourne. We use data collected by AFC and Automated Passenger Counts (APC) systems as a cost-effective method to estimate fare noninteractions and apply an econometric approach to explain the noninteractions taking into account fare evasion and other possible reasons. This study contributes to the state of the art as the observations from APC and AFC data include all travelers along multiple routes, so we have complete information on the actual passenger flows and the number of ticket validations at different locations. It enables the study of the impact of spatial variables, such as land use and network topology. The sample also covers the entire operation of services, with 24 hours and 7 day, managing to identify the temporal variation of multiple explanatory variables. We also analyse the impact of inspection and travel demand on fare noninteractions.

The reminder of this paper is organized as follows. Section 2 specifies the research background and highlights the limitations of the existing approach. Section 3 describes the model framework. Results are presented in Section 4. Conclusions, research limitations and future directions are discussed in Section 5.

2. Research Background

Among the reasons of fare noninteractions, fare evasion is the most critical issue due to its financial impact on the operation. A growing body of literature has examined the factors that explain fare evasion using on-board survey data (Guarda et al. 2016, Cantillo et al. 2022) or face-to-face interviews (Guarda et al. 2016, Delbosc & Currie 2016, 2019) collected on a sample of routes. The survey data used in these studies are usually well-designed and the consistency and randomness are guaranteed to be representative of the network usage. They concluded that fare evasion could be impacted by various factors, including time, location, socio-demographics, service operations, inspection levels etc. However, it is expensive to collect survey data and sample sizes are inevitably small. Particularly to the face-to-face interview, the data is also subject to bias as it relies on the capacity to build a relationship of trust between surveyors and interview passengers (Egu & Bonnel 2020). Fare inspection logs are another data sources to investigate fare evasion. This data is usually continuously gathered by inspectors during their daily shifts. However, the touch-on ratio might be overestimated due to the following reasons (Egu & Bonnel 2020). Firstly, the data recorded by inspectors are closely related to inspection strategies. Sampling bias might occur if inspectors do not systematically follow randomization in selecting areas and trips to inspect. Moreover, for crowded services or longer vehicles, such as tram, on-board passengers cannot be inspected completely, so the fare evasion rates are usually under-reported. When passengers perceive the presence of inspectors, they will also change their behavior. As a result, inspectors in uniform are likely to miss evaders that escape or validate cards when seeing inspectors (Delbosc & Currie 2019). In theory, all those elements should be taken into account to accurately measure fare noninteractions. Unfortunately, this cannot always be done in a satisfactory manner with the survey or inspection data. Hence, a more cost-effective alternative that will continuously collect data to support transit operators is required.

Automated Passenger Counts (APC) systems provide the opportunity to automatically record ridership rates and are becoming more common among transit operators. In contrast to the survey, no human intervention is needed to collect the data, so it can be used to monitor the service utilization of the entire network continuously. Fused with records of ticket validation in AFC data, the number of fare noninteractions can be easily calculated. Although it is hard to distinguish between fare evaders and users with other payment methods, an investigation into those noninteractions will also provide valuable insights for transit operators. For example, considering the touch-on ratio varies significantly with space and time (Sánchez-Martínez 2017), an understanding of what factors influence fare noninteractions will be an important component for decision makers to impute the missing AFC data.

This research is conducted in collaboration with the Victoria Department of Transport, and the service operator, Yarra Trams. The data available for this research comes from Melbourne trams, the largest urban tram system in the world. Concentrated with the inner suburbs, trams are the second most used form of public transport in Melbourne after rail. The multi-modal integrated ticketing system, myki, currently operates across the tram network. Myki fare collection system on Melbourne trams requires passengers to touch on. However, according to a prior investigation, the touch-on rate for the month of June 2012 was only 37%. The low touch-on ratio prohibits the use of myki data for service planning and performance measurement.

The reasons for fare noninteractions for Melbourne tram are varied. Firstly, the fare evasion rates of Melbourne tram are relatively high (Delbosc & Currie 2016). Fare validation requires passengers to touch-on when they board, but they can board the tram from any door without any contact with the driver. Although ticket inspectors are employed to check valid tickets, they only board at a few selected stops and the inspection rates are relatively low (1-2%). Secondly, while passengers are supposed to touch on when boarding, they are compliant if they hold a fare pass or just transferred from other services (train, tram, or bus). AFC systems will not record these types of passengers. Another issue that is specific to Melbourne occurs in the free tram zone located in the Central Business District (CBD) where passengers are actively discouraged to tap on or off tram services. This will make the border of the free tram zone a frontier, as people may risk travelling a few stops without touching on when they board at stops close to the free tram zone. This paper evaluates the factors that impact fare noninteractions at a stop level, using an econometric approach with automatically collected data. The following sections elaborate the data, explanatory variables, and econometric model used in this study.

3. Methodology

3.1. Data description

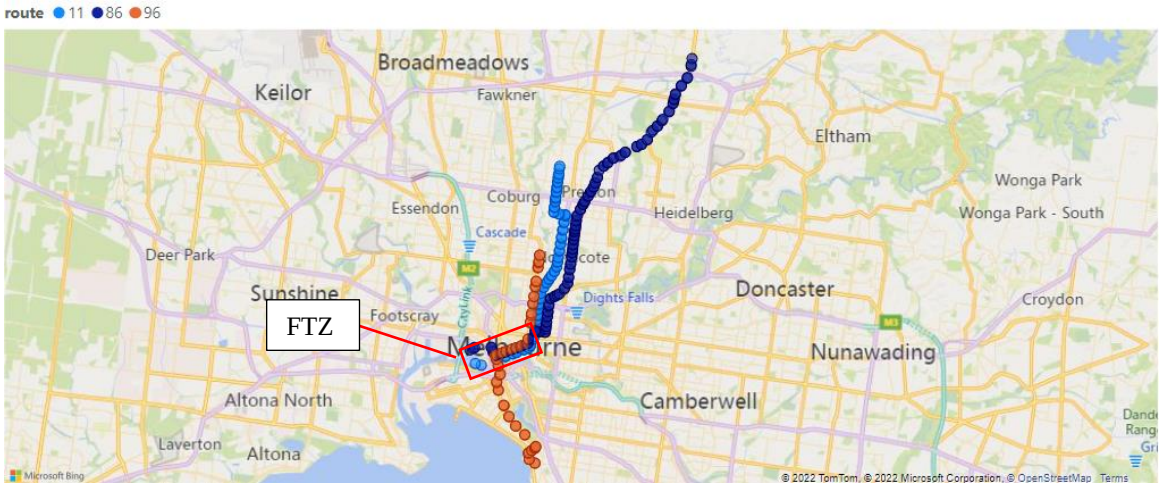
In this study, APC and AFC data collected from three tram routes (route 11, route 86, and route 96) are used, which cover the period from 01/02/2020 to 16/03/2020. Figure 1 shows the geographic distribution of these three tram lines. These three tram routes cross multiple suburbs in Melbourne from north to south, with a total of 99 kilometers in both directions. All three tram routes pass through the free tram zone (FTZ) located in Melbourne CBD.

DILAX counters were purchased for E-Class trams in Melbourne as APC systems since 2015. Each door is equipped with several active infrared sensors and each sensor works with two beams (to distinguish between boarding and alighting). A Passenger Counting Unit will check the status of all the beams and calculate the number of boarding and alighting passengers per door and stores the data. Victoria Department of Transport (DoT) has validated DILAX counters in 2019 and concluded it is accurate enough for most applications but tend to slightly undercount by one or two regardless of the crowding conditions. Since this measurement error

is not correlated with any of the explanatory variables used in this study, we conclude the undercount will not influence the conclusion.

Myki is the AFC system used for the electronic payment of fares in Melbourne. It records the individual fare validation trajectories. It is one-tap-only, so passengers usually tap when they board, but they may also tap later during their ride or when alighting. Passengers are actively discouraged to touch-on/off so as to be eligible for the free trip in the free tram zone.

Figure 1: Geographic distribution of route 11, 86, and 96 in Melbourne, Australia (extracted from GTFS shape file)



Although APC and AFC data is a rich source of information, the quality and accuracy of the data is of relatively high importance because they contain unique information that cannot be substituted by other data sources. Hence, a series of data cleaning and matching steps are employed.

For APC data, while it provides information regarding the passenger flows, the routes on which they operate, stop times and service information, further processing is still required to complete entries in the dataset. Specifically, the General Transit Feed Specification (GTFS) scheduling data provides the ideal source for matching the provided APC data to the relevant schedule, thus providing additional but necessary information about the stops and specific trips/services that pertain to the APC data entries. The matched GTFS-APC data is used to extract the locations and times for all passenger count information.

For AFC data, to ensure that stop locations identified in myki transactions are correct, the Automated Vehicle Location (AVL) dataset is used to infer transaction locations. The AVL dataset is firstly matched with the GTFS dataset utilizing the AVL timepoints (fixed location passing points) and GTFS stop times to determine the best fitting GTFS trip ID. This join has a cardinality of many-to-one, with each trips in AVL data containing several candidate trips in GTFS data. As a result, we use a fuzzy match strategy to find the most appropriate GTFS trip ID for each AVL trip ID. Each candidate is assigned an error using the following formula:

$$\boldsymbol{E_1} = \sqrt{(\boldsymbol{\partial t})^2_{\boldsymbol{start}} + (\boldsymbol{\partial t})^2_{\boldsymbol{end}}} \qquad (1)$$

where $(\partial t)_{start}$ and $(\partial t)_{end}$ denote the error of matching at the start and end time of the trip respectively.

The AVL-GTFS matched dataset is then matched myki data to infer the transaction locations. The condition used for joining is based on the vehicle number present in both datasets and a reasonable window around transaction times. As this join has a cardinality of one-to-

many (each Myki transaction matches with multiple AVL timepoints), the dataset size at this point is increased substantially, depending on the size of the window. For each unique transaction ID, the following error calculation is used:

$$E_2 = t_{timepoint} - t_{transaction} \quad (2)$$

where $t_{timepoint}$ and $t_{transaction}$ denote the AVL timepoint and the transaction time respectively

Using the calculated error, the timepoints that were reached on the subject's vehicle before the transaction was made would produce a negative value and the timepoints reached after the transaction was made would produce a positive value. The dataset is then filtered, with each transaction being assigned their closest previous timepoint, closest next time timepoint, associated trip IDs, sequence of timepoints within the trip, actual start and end times of the trip and the actual arrival time at the timepoint. To determine the specific stop at which a transaction took place, the progress between all transaction's previous and next timepoints is calculated linearly. Once the progress between timepoints has been calculated, the transaction's trip stop sequence value can then be determined by multiplying the progress percentage by the difference in trip sequence of the pair of timepoints.

Finally, the trip ID and stop sequence values are used to join with the GTFS dataset and determine the stop ID, stop coordinates and stop name of each transaction. The matched GTFS-myki data is then joined with the matched GTFS-APC data. Therefore, each record in the joined dataset will contain the total number of boarding passengers from APC data and the total number of transactions from AVL data at a given stop of a particular service trip.

After data processing, we obtain 1113 trips with 44 stops for route 11,1233 trips with 67 stops for route 86, and 1,795 trips with 33 stops for route 96. Since passengers are not required to touch on in the free tram zone, only stops outside the free tram zone were used for this study. We removed records with 0 boarding passengers and finally yielded a total of number of 61,504 observations. This is a far larger amount of data than any other studies with similar objectives.

3.2. Data modelling

3.2.1. Dependent and explanatory variables

The dependent variable to be modelled is the absolute difference between boarding flows recorded by AFC and APC data recorded at a given stop of a particular service trip.

The explanatory variables are distinguished into the following four groups: (1) time periods variables, (2) stop location variables, (3) travel demand variables, and (4) level of inspection variables. Those variables are extracted from multiple exogenous data sources.

Time periods variables APC and AFC systems continually collect data at different time periods. For time of day, we group the time periods into pre-peak (before 6 am), morning peak (6am to 9am), interpeak (9am to 3pm), afternoon peak (3pm to 6pm) and evening (after 6pm) and other time periods. A dummy for the first four periods (***pre-peak, morning, interpeak, and afternoon***) is defined while evening is set to be the reference. For day of week, we define a dummy for trips on weekends (***weekends***), taking the weekday as a reference. We also assume the tap on ratio will be influenced by weather, so a dummy variable ***raining*** that indicates whether it is raining during the trip period is also created. The weather information is obtained from microclimate sensor readings in Melbourne and is matched to trips recorded by automatic sensors based on time periods.

Stop location variables Specific to Melbourne, a Free Tram Zone (FTZ) is located in the Central Business District (CBD), where passengers are actively discouraged to touch-on so as to be eligible for the free trip. Although fare noninteractions inside the FTZ will not cause the revenue loss, this may also lead to problems at stops outside the FTZ. For example, due to a flat fare structure, some passengers may not be willing to pay for a trip with only a few stops and they may risk travelling without paying at the boundary of the FTZ. It is unclear how the fare noninteractions vary due to the free service zone. Hence, two explanatory variables are designed. Stops are firstly categorized based on the topology of the network. Stops before the free tram zone are defined as inbound stops, while stops after the free tram zone are defined as outbound stops. A dummy variable ***inbound stop*** is created, taking the value of one if the boarding stop is an inbound stop, and outbound stops are set to be the reference. To capture the effect of the boundary of the FTZ, we use the dummy variable ***close to the boundary of FTZ***, taking the value of one if the stop is within two stops past the boundary.

Numerous studies have shown that land use may exert varying effects on passengers' behavior (Boarnet & Crane 2001). Two land use types are used to study the effects on fare noninteractions, including train stations and education facilities. While passengers are supposed to touch on when boarding, they are compliant if they are transferred from other services. However, they are not charged if they touch-on. It is unclear whether passengers are willing to touch-on their cards for transfer trips from train stations without further analysis. Hence, a dummy variable ***close to train station*** is created, taking the value of one for stops near a train station to take into account transfer trips from the train. Most of the trips made at education facilities are from students, previous studies revealed students display diverse behaviours with respect to fare evasion (Barabino & Salis 2020). To study this, a dummy variable ***close to education facility*** is created, taking the value of one for stops close to an education facility, including primary schools, secondary schools, and universities. In Melbourne, it was found that the median length of a walking trip to a train station is 721 meters, while the median walk to a tram stop is 360 m (Eady & Burtt 2019). Hence, the stop is defined as ***close to train station*** if it is within 721 meters of a train station and ***close to education facility*** if it is with 360 memters of a education facility.

Travel demand variables The number of fare noninteractions will naturally grow with the passenger boarding flows. The chance of fare noninteractions are likely to increase with more passengers boarding at the given stop. To validate this hypothesis, the number of boarding passengers recorded by APC systems at the given stop (***boarding flows***) is included as an explanatory variable.

Many studies observed higher evasion rates on more crowded vehicles (Mukherjee et al. 2013, Cantillo et al. 2022). The passenger loads can be directly calculated from the boarding and alighting flows recorded by APC systems. Hence, we also included the occupancy of service, based on passenger loads and the seating capacity of the vehicle. We defined the dummy variable ***low occupancy***, taking the value of one if the passenger load at the boarding stop is less than 25^{th} percentile of passenger loads. On the other hand, a dummy variable ***high occupancy*** is also defined, taking the value of one if the passenger load at the boarding stop is more than 75^{th} percentile of passenger loads.

Level of inspection variables Previous studies showed that the fare evasion rate is expected to vary significantly according to the level of ticket inspection (Bucciol et al. 2013, Cantillo et al. 2022, Dauby & Kovacs 2007, Mukherjee et al. 2013). For Melbourne tram, inspectors are assigned to conduct fare inspections at stops. They randomly choose their boarding stops and move on randomized paths to ask passengers to show the valid tickets. The inspection logs record the boarding stop, vehicle number, and the duration for all onboard inspections. They

provide the ideal source for matching the APC data to the relevant duration of a trip that has inspectors on-board. We design two explanatory variables to model the level of inspection.

Firstly, a dummy variable ***if inspected*** is created, which takes the value one if the data is collected when at least one inspector is on-board, and zero otherwise. We also believe the frequency of inspection will change passengers' perceived likelihood of being caught and discourage them from possible fare evasion. Hence, for every stop, we also calculate the total number of times it has been inspected during the study period and use it as an explanatory variable ***frequency of being inspected***.

3.2.2. Negative binomial regression

Since the dependent variable can only be taken as counting numbers, negative binomial regression is used to explain fare noninteractions. Negative binomial regression is a generalized linear model form of regression analysis to model count data (Hilbe 2011). It assumes the dependent variable follows a negative binomial regression and the logarithm of its expected value can be modelled as a linear combination of explanatory variables. Compared to Poisson distribution (Hilbe 2011), negative binomial distribution allows the mean and variance to be different and provides a more accurate model for passenger count data (Guarda et al. 2016). The model is given by the following equations:

$$\boldsymbol{P(Y_i = y_i|\mu_i, \alpha)} = \binom{\boldsymbol{y_i + \frac{1}{\alpha} - 1}}{\boldsymbol{\frac{1}{\alpha}}} \left(\frac{\boldsymbol{1}}{\boldsymbol{1 + \alpha\mu_i}}\right)^{\frac{1}{\alpha}} \left(\frac{\boldsymbol{\alpha\mu_i}}{\boldsymbol{1 + \alpha\mu_i}}\right)^{y_i} \quad (3)$$

$$\boldsymbol{\ln(\mu_i)} = \sum_{k}^{K} \boldsymbol{\beta_k x_{ik}} \quad (4)$$

Equation 3 is the probability density function (PDF) of negative binomial distribution, where μ_i is the mean of the outcome of counts y_i for observation i, and α is the heterogeneity parameter that allows the mean and variance to be different. Equation 4 is the log-link function of the negative binomial regression model, in which x_{ik} represents k^{th} explanatory variable of observation i and β_k is its corresponding parameters.

4. Results

4.1. Myki touch-on rates

Before econometric analysis, we firstly calculated the Myki touch-on rates, which is the total number of transactions from Myki data divided by the total number of boarding passengers from APC data. We removed records with 0 boarding passengers as it generates infinite values. Table 1 shows the average touch-on rates in different time periods of pre- and post-COVID-19 pandemic.

Results suggest touch-on rates are higher between 6am and 6pm comparing to other time periods. Econometric analysis is further required to study the marginal effect of explanatory variables. It is also noticeable that the touch-on ratio is much less during the COVID restriction, indicating many passengers changed travel behaviors and tend to not validate their tickets when board. It is a question of future research to investigate the low touch-on rate during COVID restriction. For this study, we only conduct econometrics analysis for data before March 15th 2020 in the following sections.

Table 1: Average Myki touch-on rates, separated by time periods

	Pre-COVID (before March 15th 2020)	During COVID restriction (after March 15th 2020)
Pre-morning peak (before 6am)	0.29	0.14
Morning peak (6am – 9am)	0.49	0.18
Inter peak (9am – 3pm)	0.44	0.15
Afternoon peak (3pm – 6pm)	0.43	0.14
Evening (after 6pm)	0.31	0.1

4.2. Model analysis with the whole data

According to the dependent and explanatory variables discussed in Section 2.3.1, the negative binomial regression models were applied to explain the number of noninteractions. The correlation between all explanatory variables was analyzed by the variance inflation factor (VIF) (O'brien 2007) and no problems of high correlation between the variables were observed.

Table 2 presents the summary of the resulting model, including the estimated coefficients and the standard error of coefficients. The last column indicates the significance of the explanatory variable, where *, **, *** represents the variable is significant at 95%, 99%, and 99.95% confidence interval.

The most significant variable is the actual number of boarding flows (***boarding flows***), with a positive effect on fare noninteractions. This is obvious as the noninteractions have more opportunity to occur when more passengers board. The number of fare noninteractions naturally grows with the number of passengers boarding. Another explanation is that the difficulty to reach ticket validators with a large boarding group. Variable (***high occupancy***) also has a significant positive impact on fare noninteraction, indicating fare evasion is more likely to occur on crowded vehicles. We also notice when the vehicle occupancy is low (***low occupancy***), fewer noninteractions are observed.

In terms of time periods, the results suggested that fare noninteractions are lower in the morning peak, afternoon peak and interpeak, compared to other time periods. Moreover, more observed during the weekends. The majority of trips in these time periods are commuting trips, which are usually made by regular public transport users. The myki "Pass" product is designed for regular users that allow them to pay up front for unlimited travel during a period with discounts. However, data shows that noninteractions are generally less common for time periods associated with frequent public transport users. It will be important that future research to investigate the proportion of fare noninteractions due to myki "Pass" using survey data. On the other hand, during the evening or weekends, trips are made with varying purposes, such as shopping or recreation. We notice fare noninteractions are higher among these trips.

It is interesting to note that the variable ***if inspected*** only has small impact on the fare noninteractions. However, this is not necessarily true considering the ticket inspection rate. The inspection rate observed in the data is only 1%, so the sample size of inspected observations is inevitably small. A small sample size leads to high variability and biases the hypotheses test of the estimated coefficient. For variable ***if inspected***, the standard error of the coefficient is relatively high, indicating high variability in the sample of observations with inspections. Hence, it is difficult to explain the results of variable ***if inspected***. However, variable ***frequency of being inspected*** turns out to have a negative impact on fare noninteractions, being the second most significant variable of the model. This is due to the fact that the passenger perception of being inspected is influenced by inspection levels. As a result, at stops that have been frequently inspected, the subjective probability that passengers feel they will be checked is high, which

discourages them to evade fares. In line with previous studies (Clarke et al. 2010), our results demonstrate that ticket inspection can be used as a deterrence against fare evasions.

Table 2: Model estimates of fare noninteractions with the whole dateset

AIC: 327746			
variable	coefficient	std. error	significance
time periods			
pre-peak (before 6am)	0.003	0.023	
morning (6am-9am)	-0.033	0.009	***
Interpeak (9am-3pm)	-0.04	0.008	***
afternoon (3pm-6pm)	-0.091	0.008	***
weekend	0.031	0.006	***
Weather			
raining	0.011	0.143	
travel demand			
boarding flows	0.164	0	***
low occupancy	-0.047	0.006	***
high occupancy	0.051	0.006	***
stop location			
close to train station	-0.049	0.006	***
close to education facility	-0.064	0.016	*
close to the boundary of FTZ	0	0.009	
inbound stop	-0.132	0.005	***
inspection level			
if inspected	-0.094	0.034	**
frequency of being inspected	-0.009	0.001	***
intercept	0.272	0.009	***

With regard to locations, our results also indicate that fare noninteractions are lower at inbound stops, which are stops before the free tram zone in the direction of the service. We also notice fare noninteractions are lower at stops close to train stations and education facilities, but the boundary of the free tram zone does not play an important role. However, a previous study (Sánchez-Martínez 2017) shows that fare noninteraction significantly varies with space and time. It is difficult to explain such results with the model estimates of the whole data. Therefore, to understand the interactions between time periods and stop location variables, we used different models for each time periods at each direction. Results are discussed in the next section.

4.3. Model analysis with segmented data

To further understand the spatial and temporal variation of fare noninteractions, we segmented the data based on time and location and applied negative binomial regression to each

Table 3: Model estimates of fare noninteractions with data in the morning peak (Left: inbound direction; Right: outbound direction)

	AIC: 189176			AIC: 56836		
variable	coefficient	std. error	significance	coefficient	std. error	significance
time periods						
weekend	0.029	0.008	***	0.02	0.014	
weather						
raining	0.304	0.019		0.017	0.037	
travel demand						
boarding flows	0.165	0	***	0.166	0.001	***
low occupancy	-0.039	0.008	***	-0.079	0.015	***
high occupancy	0.074	0.008	***	0.049	0.014	***
stop location						
close to train station	0.019	0.018		-0.063	0.015	***
Close to education facility	-0.037	0.023		-0.167	0.035	***
close to boundary of FTZ	0.025	0.012	*	-0.141	0.037	***
inspection level						
if inspected	-0.048	0.045		-0.12	0.081	
frequency of being inspected	-0.018	0.001	***	0.005	0.002	
intercept	0.167	0.008	***	0.128	0.014	

segmentation. We look specifically at the model estimates in the peak hours for both inbound and outbound stops. Data is divided ***into morning inbound, morning outbound, afternoon inbound, and afternoon outbound***. Table 3 - 4 present the model estimates with data for each segment.

Results reveal that there are more significant variables for ***morning inbound*** and ***afternoon outbound*** compared to the other two segments. This can be explained by the purposes of the trips. Most of the trips within these two segments have commuting purposes, so they are likely to be made by regular users. Hence, their behaviors, such as the responses to the service occupancy and land use, are more predictable.

For education facilities, we notice fewer fare noninteractions are observed, especially at outbound stops in the afternoon peak, in which many students travel from schools or universities to their home locations. Most of the trips originating from education facilities are made by students or staff working in those facilities. Previous research (Barabino & Salis 2020) revealed that students with a high school diploma are less likely to be fare evaders, as they may have a higher sense of morality. Since most of the education faculties observed in this study are universities and high schools, our results are in line with previous studies. Moreover, students are eligible for concession fares, which also discourages them to evade fares.

Table 4: Model estimates of fare noninteractions with data in the afternoon peak (Left: inbound direction; Right: outbound direction)

	AIC: 139691			AIC: 98486		
variable	coefficient	std. error	significance	coefficient	std. error	significance
time periods						
weekend	0.054	0.009	*	0.012	0.011	
weather						
raining	-0.016	0.021		0.026	0.027	
travel demand						
boarding flows	0.173	0.001	***	0.159	0.001	***
low occupancy	-0.021	0.009	*	-0.086	0.012	***
high occupancy	0.122	0.009	***	0.023	0.01	*
stop location						
close to train station	-0.013	0.106		-0.042	0.01	***
close to education facility	-0.013	0.019		-0.286	0.042	***
close to boundary of FTZ	-0.016	0.012	***	-0.015	0.019	
inspection level						
if inspected	-0.08	0.048		-0.185	0.068	**
frequency of being inspected	-0.048	0.002	***	0.016	0.002	***
intercept	0.183	0.009	***	0.166	0.01	***

In terms of train stations, our results show that the number of noninteractions is lower at stops close to train stations. It indicates many passengers transferring from train to tram tend to touch-on their cards even though they are not required. We speculate that some passengers may not know they are compliant if they do not touch-on, so they will naturally follow the rules as long as they are not charged. However, this variable is not significant for the inbound stops in the afternoon peak. As shown in Figure 3, this is because fewer transfer trips are made within this segment.

It is also notable that the sign of variable ***close to the boundary of FTZ*** varies with space and time. The sign is positive for inbound stops in the morning peak, but negative for inbound stops in the afternoon peak. The reason behind this phenomenon may lie in the differences between OD flows. We derived the OD flows from AFC data using the developed trip chaining models in the literature (Gordon et al. 2013). Figure 2 and 3 show the OD flow diagram of inbound stops in the morning and afternoon peak respectively. Stops are categorized into inside ***FTZ***, ***close to the boundary of FTZ***, ***close to the train station***, ***close to education facility***, and ***other*** based on their locations. The left-hand side represents the type of boarding stops, right-hand side represents the type of alighting stops, and the width is proportional to the OD flow rate.

As shown in Figure 2, many trips starting at the boundary of the FTZ have destinations inside the FTZ. This number is even underestimated as the OD flow diagram is based on records in the myki data, so fare noninteractions travelling between the boundary and the FTZ

are not included. However, compared to the records in the APC data, there exists a large proportion of trips between those two regions. This implies that fare noninteraction is very significant for those trips. A possible explanation is that passengers may not be willing to pay for trips with only a few stops, so they may risk travelling from the boundary to the free tram zone.

Figure 2: OD flow diagram of inbound stops in the morning peak

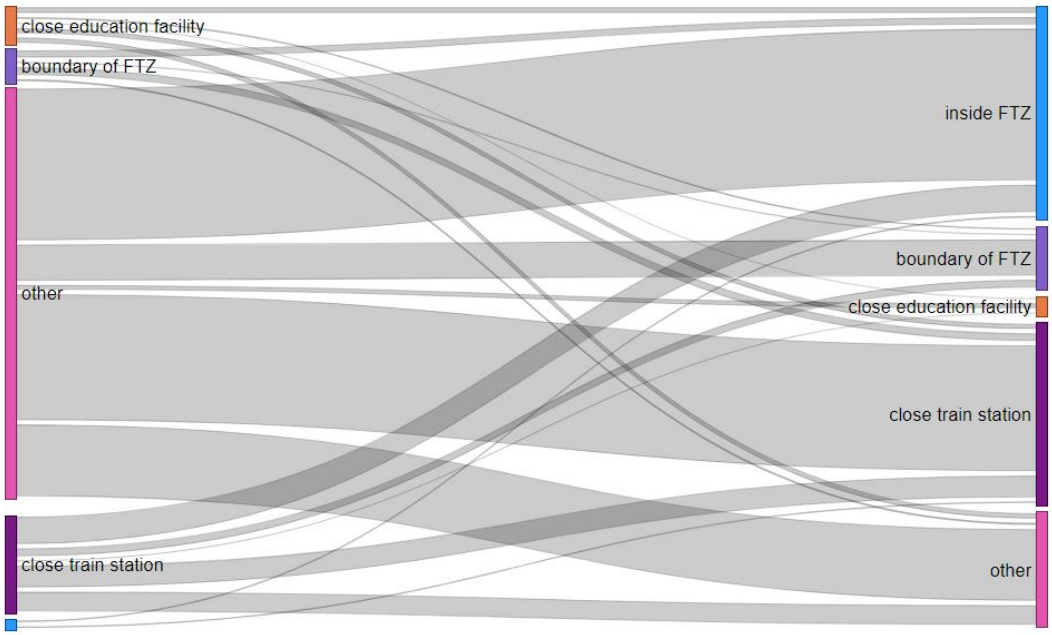
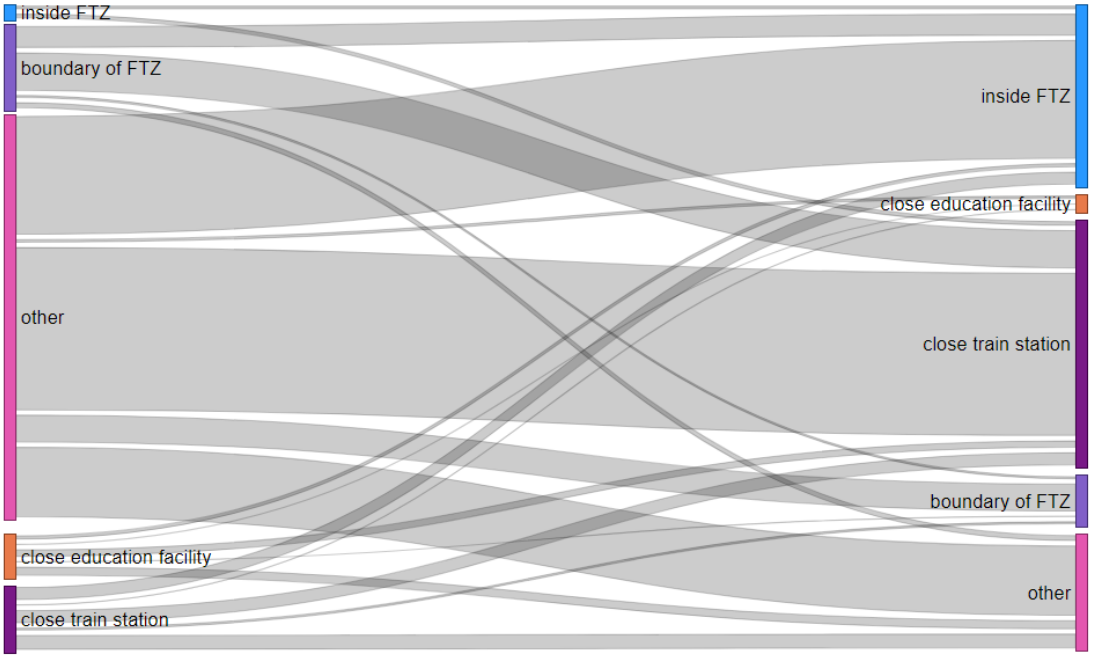


Figure 3: OD flow diagram of inbound stops in the afternoon peak



Unlike the morning peak, OD flow diagram in Figure 3 shows a large proportion of trips generated from the boundary of the FTZ in the afternoon peak have the destination at stops close to train stations. These could be transfer trips connecting passengers' activity locations to train stations prior to the train trip stage. It is practically difficult to evade fares at

train stations, as most of the stations are equipped with turnstiles. Since most of the passengers have to pay for the train trip stage at the train stations and there are no extra charges for transfer trips, there is no point to evade fares at the tram trip stage for transfer passengers.

For outbound stops, we notice that fare noninteractions are less likely to occur at the boundary of the FTZ. Trips generated in this direction are to the outer suburbs instead of the FTZ, so the FTZ will not affect passengers' willingness to touch-on their cards.

5. Conclusion and future research

According to the authors' knowledge, this is the first study that evaluates how different factors affect the fare noninteractions using a large-scale automatically collected data. Compared to previous studies, it presents a cost-efficient methodology for processing multiple sources of data to measure fare noninteractions. We specifically investigate passengers' behavior in validating tickets in the presence of a free service zone and the interactions between time and land use.

The analysis leads to the following conclusions. Firstly, fare noninteractions are lower in the period with more regular public transport users. Although the weekly or monthly pass is designed for those regular commuters, it does not lead to more fare noninteractions during the commuting time. The regression analysis also suggests fare noninteractions made during the commuting time are more explainable, with a smaller interception and more significant variables, compared to other time periods.

Although passengers are not required to touch-on when they are transferring from other services, fewer fare noninteractions are observed at tram stops close to train stations. Moreover, our data indicates that how free service zone affects the fare noninteractions is dependent on the purposes of trips. In the morning peak, fare noninteractions are higher at the boundary of the free service zone, because most of the trips generated there are commuting trips to the work location inside the free tram zone. Passengers may not be willing to pay for trips with only a few stops, so some may risk travelling without payment. In the afternoon, however, as many trips generated at the boundary are transfer trips to the train station, fewer fare noninteractions are observed. Results also show that trips made close to an education facility have fewer fare noninteractions in the afternoon peak, suggesting students with concession are more likely to comply with the rules.

The fare noninteractions are also influenced by the travel demands and the inspection frequency. Fare noninteractions are always higher when they board in a large group or on a crowded vehicle. In terms of inspection level, the inspection frequency significantly affects the touch on rates. Results provide evidence that the subjective probability that a passenger feels he or she will be checked is an important factor that influences passengers' decisions on fare evasions.

There are some limitations to our approach. The methodologies used in this paper are mainly based on statistical analysis and non-intrusive data. Although it correctly identifies the major influential factors of fare noninteractions, it could not capture variables that cannot be observed from the passively collected data. For example, previous studies show that fare evasions can be impacted by gender or income (Guarda et al. 2016). People may also be sensitive to group behavior or social pressures (Gino et al. 2009). Therefore, future research could mix both qualitative and quantitative research methods, relying on both face-to-face interviews and automatically collected data. More variables at stop levels could be obtained from surveys or interviews, such as (1) major purposes of trips, (2) average income of boarding

passengers, (3) the likelihood of people travelling in groups, (4) the proportion of deliberate or unintentional evaders, among others. Moreover, existing data sources cannot measure how many people exit the FTZ without paying. If people coming in from the boundary of the FTZ are not validating, it logically makes sense that when they leave the city they will also exit a few stops outside the FTZ without validating. This method cannot capture the fare evaders who board within the FTZ and exit a few stops outside the FTZ.

Another limitation is that the matching of AFC data with scheduled trips is based on the minimum difference between the vehicle's arrival time and the ticket transaction time. However, sometimes passengers may validate tickets at other stops instead of their boarding stops. For example, some passengers may not be able to find their tickets when they board and will validate tickets a few stops later. This could generate an inaccuracy in the fare noninteractions at some stops. Hence, future research is needed to improve the process of matching APC, AFC, AVL, and GTFS data.

Acknowledgement

This research is financially supported by the Victoria Department of Transport (DoT), Cubic Transportation Systems, and iMOVE Australia. The authors would like to thank DoT, iMOVE, Cubic, and Yarra Trams for providing the AFC, AVL, APC data, as well as their insights and consultations.

References

Barabino, B. & Salis, S. (2020), 'Do students, workers, and unemployed passengers respond differently to the intention to evade fares? an empirical research', Transportation Research Interdisciplinary Perspectives 7, 100215.

Boarnet, M. & Crane, R. (2001), 'The influence of land use on travel behavior: specification and estimation strategies', Transportation Research Part A: Policy and Practice 35(9), 823–845.

Bucciol, A., Landini, F. & Piovesan, M. (2013), 'Unethical behavior in the field: Demographic characteristics and beliefs of the cheater', Journal of Economic Behavior & Organization 93, 248–257.

Cantillo, A., Raveau, S. & Muñoz, J. C. (2022), 'Fare evasion on public transport: Who, when, where and how?', Transportation Research Part A: Policy and Practice 156, 285–295.

Clarke, R. V., Contre, S. & Petrossian, G. (2010), 'Deterrence and fare evasion: Results of a natural experiment', Security Journal 23(1), 5–17.

Dauby, L. & Kovacs, Z. (2007), 'Fare evasion in light rail systems', Transportation Research Circular (E-C112).

Delbosc, A. & Currie, G. (2016), 'Four types of fare evasion: A qualitative study from melbourne, australia', Transportation Research Part F: Traffic Psychology and Behaviour 43, 254–264.

Eady, J and Burtt, D (2019). Walking and transport in Melbourne suburbs. Victoria Walks,

Melbourne, November 2019.

Delbosc, A. & Currie, G. (2019), 'Why do people fare evade? a global shift in fare evasion research', Transport Reviews 39(3), 376–391.

Egu, O. & Bonnel, P. (2020), 'Can we estimate accurately fare evasion without a survey? results from a data comparison approach in lyon using fare collection data, fare inspection data and counting data', Public Transport 12(1), 1–26.

Gino, F., Ayal, S. & Ariely, D. (2009), 'Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel', Psychological science 20(3), 393–398.

Gordon, J. B., Koutsopoulos, H. N., Wilson, N. H. & Attanucci, J. P. (2013), 'Automated inference of linked transit journeys in london using fare-transaction and vehicle location data', Transportation research record 2343(1), 17–24.

Guarda, P., Galilea, P., Paget-Seekins, L. & de Dios Ortúzar, J. (2016), 'What is behind fare evasion in urban bus systems? an econometric approach', Transportation Research Part A: Policy and Practice 84, 55–71.

Hilbe, J. M. (2011), Negative binomial regression, Cambridge University Press.

Mukherjee, C., White, H. & Wuyts, M. (2013), Econometrics and data analysis for developing countries, Routledge.

Munizaga, M. A. & Palma, C. (2012), 'Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile', Transportation Research Part C: Emerging Technologies 24, 9–18.

Nassir, N., Hickman, M. & Ma, Z.-L. (2015), 'Activity detection and transfer identification for public transit fare card data', Transportation 42(4), 683–705.

Nassir, N., Hickman, M. & Ma, Z.-L. (2019), 'A strategy-based recursive path choice model for public transit smart card data', Transportation Research Part B: Methodological 126, 528– 548.

Nassir, N., Khani, A., Lee, S. G., Noh, H. & Hickman, M. (2011), 'Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system', Transportation research record 2263(1), 140–150.

O'brien, R. M. (2007), 'A caution regarding rules of thumb for variance inflation factors', Quality & quantity 41(5), 673–690.

Sánchez-Martínez, G. E. (2017), 'Estimating fare noninteraction and evasion with disaggregate fare transaction data', Transportation Research Record 2652(1), 98–105.