CASPT 2018

Route Choice Stickiness of Public Transport Passengers: Measuring Habitual Bus Ridership Behaviour using Smart Card Data

Jiwon Kim · Jonathan Corcoran · Marty Papamanolis

Abstract Drawing on smart card data this paper introduces a new metric, termed the 'Stickiness Index' to quantify individual travel behaviour decisions across a public transit system. We demonstrate the utility of the Index to distinguish locales where travellers always select the same route (high stickiness) versus those exhibiting a more diverse pattern of route selection (low stickiness). Mapping the Index reveals that stickiness varies geographically across the public transit system as well as over the course of the day. Modelling the features that explain variations in the Index suggest that higher levels of stickiness are associated with high frequency users in situations where there is substantial variability of route travel times across all possible alternatives. We contend the Index has potential to help public transit agencies in decisions concerning system design and scheduling through offering a simplified depiction of travel behaviour.

Keywords: Public transport · Smart card data · Bus route choice · Habitual behaviour · Stickiness Index · Quantile regression

1. Introduction

Collectively, our individual travel behaviour decisions have important consequences for transit systems. Modal selection, timing and route choice each combine to influence travel experience and contribute to congestion and environmental pollution. From a public transit system perspective, people's trip making behaviours have major implications for service

Jiwon Kim
School of Civil Engineering, The University of Queensland
Brisbane, Australia
Email: jiwon.kim@uq.edu.au

Jonathan Corcoran
School of Earth and Environmental Sciences, The University of Queensland
Brisbane, Australia
Email: jj.corcoran@uq.edu.au

Marty Papamanolis
School of Civil Engineering, The University of Queensland
Brisbane, Australia
E-mail: s.papamanolis@uq.edu.au

delivery in their capacity to place loads in particular locations of a system, specific routes and at certain times. In order to design and operate efficient and effective public transit systems that offer a competitive alternate to private transit options, there is a need for us to both capture and develop analytic approaches to reveal transit behaviour decisions and travel experience.

What we know from travel behaviour scholarship is that people's trip making varies by a number of individual factors including gender (Gordon et al., 1989), age (Chudyk et al., 2015), occupation (Rasouli et al., 2015), household characteristics (Dieleman et al., 2002), socio-economics (Kotval-K and Vojnovic, 2015) and is conditioned by land use (Hong et al., 2014) and urban form (Handy, 1996). What is less evident from this literature is the spatio-temporal dimension of travel behaviour and how individual trip making decisions translate to spatial and temporal patterns across a transit system. Traditionally research in this area has been limited by the availability of suitable data sources, a lack of suitable analytic tools and restrained by computation capabilities (Yue et al., 2014). Whereas data and computational constraints have now eased, what has become increasingly apparent in the advent of 'big data' is the growing role played by algorithms to mine data and reveal travel behaviour dynamics placing them in the context of system-wide patterns (Kwan, 2016).

The emergence of smart card systems for automated fare collection has been increasingly recognised by transport researchers as a valuable source of (big) data to understand a growing range of public transit dynamics, including travel demand variability (Morency et al., 2007), origin-destination estimation (Munizaga and Palma, 2012), trip purpose inference (Lee and Hickman, 2013), spatio-temporal patterns (Tao et al., 2014a), service reliability (Ma et al., 2015), modal transfer behaviour (Sun et al., 2015), and passenger experience (Chu and Lomone, 2016), to name only a few. Advantages of these data have been noted in their capacity to contribute to strategic, tactical and operational system management (Pelletier et al., 2011).

Despite a growing body of literature on the use of smart cards in travel behaviour research, a relatively little attention has been paid to investigating habitual aspects of travel behaviours. It has been noted that daily travellers tend to make the same choices over and over again (Gärling and Axhausen, 2003; Schlich and Axhausen, 2003). Such habit effects can lead travellers to

make decisions that are not optimal, from an economical point of view. This poses a great challenge to transport modelling and planning as these behavioural tendencies do not conform to the basic assumption of conventional choice models that passengers make rational decisions to maximize expected utility. As such, there is a need to accurately measure and analyse habitual travel behaviour to take this into account in travel choice models and, thus, to design travel demand management strategies that can effectively influence travel choices. Most existing studies addressing this issue, however, focus on car drivers—habit effects on drivers' mode choice (Innocenti et al., 2013; Thøgersen, 2006) and route choice (Prato et al., 2011; Vacca and Meloni, 2015)—and habitual behaviours of bus passengers remain largely unexplored.

The goal of this study is to fill this gap by investigating habitual behaviours of bus passengers' route choice decisions using smart card data. The main difficulty in studying habitual travel behaviour has traditionally been the lack of suitable data that allow the observation and identification of intrapersonal and interpersonal variability in travel behaviour over a sufficiently long period of time. This issue is addressed in this study through the use of smart card data that cover a period of six months. Drawing on a single large smart card database of bus ridership, we first construct travel trajectories for individual passengers across a metropolis. Travel trajectories comprise a set of spatio-temporal points describing a stop-to-stop sequence of an individual bus passenger travelling through the network. By tracking travel trajectories of each passenger over the six-month period and matching them with other passengers' trajectories, we identify a set of routes that are available for each origin-destination (OD) pair and investigate how those routes are utilized by each passenger. We introduce the concept of *stickiness* to describe an individuals' tendency towards 'sticking' to only one route regardless of the availability of multiple alternative routes. We define a metric called the *Stickiness Index* (SI) to quantify the range of preferences from users that always travel on the same route (high stickiness) to those with a more varied pattern of route selection (low stickiness).

The contributions of this paper to the existing literature are as follows: First, we mine travel patterns based on *trajectories* of individual passengers, rather than boarding and alighting points that many existing studies have focused on. Detecting travel regularities and identifying similar travel behaviours based only on start and end points of journey arguably falls short of being

able to consider a full picture of individual journeys. In this study, whole trajectories of individual passengers are constructed from smart card transaction records and a trajectory clustering method is used to identify travel regularities in travellers' day-to-day route choice patterns. This allows us to capture travellers' perceived route choice sets by taking into account the actual shape of a path in defining similar routes. Second, we take a *big data* approach to understanding habitual route choice behaviour, performing both longitudinal and cross-sectional analyses at a metropolitan-wide scale. Using a six-month smart card dataset that yields 24 million trajectories that were generated by 814 thousand users, we measure the route choice stickiness of individual users for about 5,500 unique OD-pairs and compare each across users and ODs to reveal the key factors that explain route stickiness. Given that most previous studies investigate route choice behaviour from only a few selected OD-pairs with data that covers a few days or weeks, this study extends the scale in terms of the length of the observation period and completeness of geographical coverage and passenger population to offer a more comprehensive examination. Third, we propose a simple metric, the *Stickiness Index*, to measure and quantify individual bus passengers' route stickiness tendency. By extending this metric to the OD-level, gives an indication of the collective behaviour of travellers of a particular OD, this study also reveals individual travel behaviours in the context of system-wide patterns.

The remainder of the paper is organised as follows: The next section provides an overview of the current scholarship on the use of smart card data in travel behaviour studies. Section three proposes a methodology for measuring habitual route choice behaviours of bus passengers by introducing the notion of stickiness and the associated metrics. The case study area (Brisbane, Australia) is introduced in section four along with a discussion of the data and empirical approach. Section five presents the results in terms of both user-level characteristics and system-wide patterns before offering a set of tentative conclusions and avenues for future research in section six.

2. Background literature: *the use of smart card data in travel behaviour studies*

A rich and growing body of literature has emerged in recent years that discusses the use of smart card data in travel behaviour analysis and modelling. In an early piece, Bagchi and White (2005) discuss the potential

of public transport smart card data for travel behaviour analysis in which they summarise their various benefits and limitations over traditional survey data. The benefits include the availability of larger volumes of personal travel data and longer temporal coverages over which user behaviour can be observed, thereby allowing the identification of different user groups based on their travel behaviour and turnover rates. The lack of information on journey purpose, ultimate origin and destination locations of individual users, and actual trip activity chains are identified as limitations of smart card data. Of relevance to the current study are two streams of research: (1) route choice behaviour; and, (2) data mining of travel patterns, each of which is now discussed in more detail.

Route choice behaviour

A number of studies have focused on route choice behaviour of bus passengers. In one study, Jánošíková et al. (2014) employed a multinomial logit model to investigate factors affecting bus passengers' route choice behaviours, based on a choice set inferred from smart card data. Their model considered four attributes including in-vehicle travel time, walking time, number of transfers and headway, where in-vehicle travel time was found to be the most significant characteristic explaining both morning peak and off-peak periods. Schmöcker et al. (2013) used a discrete choice model that incorporated both choice set generation and route choice using a nested model formulation. On the upper level, the choice set (a set of attractive routes) was estimated based on utility maximisation for a user; on the lower level, the user is assumed to take the first arriving bus from their choice set identified on the upper level. The model was tested on three OD-pairs. In another study, Viggiano et al. (2014) investigated bus passengers' route choice strategies in terms of whether a user will board any bus serving their destination ("first bus strategy") or wait for a specific bus to make their trip ("favourite bus strategy'). They conducted an on-line survey and analysed smart card data finding that trip length, use of countdown information, and passengers' willingness to risk waiting for a faster bus that has not yet arrived were correlated with passengers' route choice strategies. Nassir et al. (2015) inferred the set of attractive routes for each regular commuter using smart card data and presented empirical findings for these observed attractive sets. The study selected six OD-pairs and analysed factors affecting route attractiveness for regular passengers using a binary logit model.

In sum, route choice models generally assume that passengers make rational decisions to maximize expected utility. There are, however, other behavioural complexities that are known to exist in passengers' route choice decisions, such as habit effects, bounded rationality, risk attitudes, learning, and adaptation (Liu et al., 2010). A review of scholarship that examine the effects of habits reveals very few studies that have focussed on the habitual aspects of route choice behaviour using smart card data. In one study examining London commuters, Kurauchi et al. (2014) investigated the consistency or variability in bus passengers' routing decisions over a two week period using a Markov chain model. They computed an n-Step Markov model to capture the probability of choosing the same route or a different route given the choices made on n previous days. Their results revealed that most London commuters do not stick to the same line each morning but show a certain degree of variation in bus route choice.

Data mining travel patterns

Data mining techniques applied to smart card data have been shown to have utility in their capacity to identify and analyse user groups based on their spatial and temporal trip patterns. In one study, Nishiuchi et al. (2012) identified proportions of frequent spatial (route choice) patterns and frequent temporal (departure time choice) patterns for each user. By associating spatial and temporal patterns, the study observed varying levels of correlation and regularity in spatial and temporal patterns for different user categories. Ma et al., (2013) applied a density-based spatial clustering of application with noise (DBSCAN) to cluster passengers with similar travel patterns, in terms of number of travel days, number of similar first boarding times, number of route sequences, and number of similar stop ID sequences. Kieu et al. (2015) applied a DBSCAN to mine spatial patterns (regular ODs) and temporal patterns (habitual boarding times) separately. A market segmentation analysis is then performed by categorizing passengers into four groups based on regularity or irregularity in their spatial and temporal trip patterns. Goulet-Langlois and colleagues (2016) used a principal component analysis (PCA) approach to cluster passengers' activity sequence patterns that were inferred from four-week smart card data. Their study revealed 11 clusters showing different characteristics in the longitudinal activity sequences, characterised by different proportions of time spent in each activity status (e.g., at work, home, and in transit).

Spatial and temporal trip patterns have been also investigated at a network level. Tao et al. (2014b) analysed and visualised aggregate flow patterns using flow-comaps created based on travel trajectories constructed from smart card data. In a final study, Zhong et al. (2016) investigated temporal distributions of trip start times and aggregate flow patterns at stations and compared the patterns in London, Singapore and Beijing using one-week of smart card data.

In sum, the scholarship in data mining travel patterns has expanded to encompass approaches ranging from traditional descriptive data analyses to emerging techniques in machine learning and predictive modelling. Previous studies, however, have focused on boarding and alighting points in mining travel patterns, which have limitations in comprehensively identifying travel regularities and similarity between routes. Mining whole journey trajectories could overcome such limitations, which have been adopted by recent studies (e.g., Tao et al., 2014b) including the present study in this paper.

Despite a growing body of literature in smart card-based travel behaviour research, there remains a paucity of studies which investigate habitual aspects of passengers' route choice behaviours. Furthermore, there is a desire for developing analysis methods that utilize trajectories of passengers, which could not only capture individuals' travel patterns more accurately but also allow them to be easily extended to other smart card systems or other data sources that produce trajectory data. In this paper we redress some of these gaps in current scholarship through the examination of both system and individual route choice behaviour using passenger trajectories from smart card data and a new metric at a scale that is likely to yield a more comprehensive picture of metropolitan travel dynamics.

3. Methodology: Measuring Route Stickiness of Bus Passengers

This paper aims to measure and analyse bus users' route choice behaviours with a particular emphasis on capturing the 'stickiness' of users' route choice. The route choice stickiness is defined as the user's tendency of persistently making the same route choices in traveling between a given origin-destination (OD) pair. A user with high stickiness will tend to always use the same route despite the availability of multiple routes to the user. On the other hand, a user with low stickiness will tend to use the available routes more evenly. Since the route choice stickiness is defined for a particular user travelling a particular OD, we can also define the OD-level

stickiness by aggregating the stickiness tendencies for the users in that OD. That is, the stickiness of an OD pair is defined as the collective tendency of the users of that OD towards the persistent selection of the same route choices.

With these definitions, three questions are addressed in this paper:

- What are the characteristics of bus passenger route choice stickiness?
- What factors explain a user's route choice stickiness?
- What are the geographic dynamics of OD-level stickiness?

To investigate these questions, we first define the *Stickiness Index* (SI) metric to quantify user-level and OD-level stickiness. The notations associated with the SI are summarized in Table 1.

Table 1. Notation summary

Symbol	Description
k	Origin-destination (OD) pair index
i	User or passenger index for OD-pair *k*, $i = 1, \cdots, I^k$
j	Bus route index for OD-pair *k*, $j = 1, \cdots, J^k$
I^k	Total number of bus passengers for OD-pair *k*
J^k	Total number of bus routes connecting OD-pair *k*
J_i^k	Total number of bus routes connecting OD-pair *k*, that are used by user *i*
$n_{i,j}^k$	Number of journeys made by user *i* travelling OD-pair *k* using route *j*
N_i^k	Total number of journeys made by user *i* to travel OD-pair *k*, $N_i^k = \sum_{j=1}^{J^k} n_{i,j}^k$
$p_{i,j}^k$	Proportion of journeys made by user *i* using route *j* when traveling OD-pair *k*, $p_{i,j}^k = n_{i,j}^k / N_i^k$
S_i^k	User-level Stickiness Index, i.e., the SI of user *i* in travelling OD-pair *k*
S^k	OD-level Stickiness Index, i.e., the SI of OD-pair *k*, measured as the weighted average of user-level SI for users of OD-pair *k*.

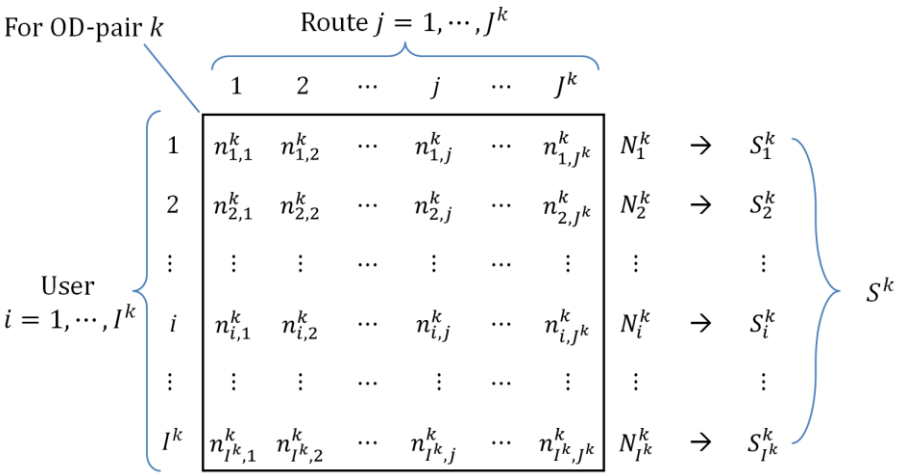

Figure 1. User-Route journey frequency matrix and Stickiness Index (SI) defined at individual user-level (S_i^k) and OD-level (S^k)

User-level Stickiness Index (SI)

The goal of defining the *Stickiness Index* (SI) is to quantitatively measure a public transit user's route stickiness tendency, a tendency towards choosing the same route repeatedly for a given OD-pair. After devising a suitable formula for the SI, we found that *Simpson's Diversity Index* (Simpson, 1949), a metric commonly used in ecology to measure biodiversity, can be effectively adapted to our context and offers a simple and intuitive way of quantifying route choice diversity. This study, thus, adopts and modifies this diversity index to derive an expression for the stickiness index to measure route usage concentration or the level of competition of routes within each user's route choice set.

For given OD-pair *k*, we first construct a *User-Route matrix* as shown in Figure 1, where rows correspond to a set of users for that OD ($i = 1, \cdots, I^k$), columns correspond to a set of available routes ($j = 1, \cdots, J^k$), and the entry in the i^{th} row and j^{th} column ($n_{i,j}^k$) represents the frequency of journeys made by user i using route j to travel OD-pair k. Next, we compute a diversity index, denoted by D_i^k, for each row using the following formula:

$$D_i^k = \sum_{j=1}^{J^k} \left(p_{i,j}^k\right)^2 = \sum_{j=1}^{J^k} \left(\frac{n_{i,j}^k}{N_i^k}\right)^2 \qquad (1)$$

where $n_{i,j}^k$ is the number of journeys made by user *i* using route *j* to travel OD-pair *k*, N_i^k is the total journeys made by user *i* to travel OD-pair *k*, and $p_{i,j}^k$ is the proportion of the journeys made using route *j* (i.e., $n_{i,j}^k/N_i^k$) such that $\sum_{j=1}^{J^k} p_{i,j}^k = 1$. Given $p_{i,1}^k$, $p_{i,2}^k$, $\cdots, p_{i,J^k}^k$ representing the proportional abundances of the J^k routes in the journey sample for user *i* travelling OD-pair *k*, D_i^k is equivalent to a weighted mean of the proportional abundances of the routes with the proportional abundances themselves being used as the weights. D_i^k can be also interpreted as the probability that two journeys taken at random from the sample will belong to the same route.

The diversity index D_i^k in Eq. (1) is known as *Simpson's Diversity Index* in ecology, which measures species diversity within a habitat (Simpson, 1949), or *Herfindahl-Hirschman Index* in economy, which measures market concentration or the level of competition of companies within a market or industry (Herfindahl, 1950; Hirschman, 1945). Simpson's Diversity Index takes into account both *richness* (i.e., the number of species present) and *evenness* (i.e., relative abundances of different species) in measuring species diversity, where diversity is considered to increase as richness increases (the number of different species is larger) and evenness increases (the population is more evenly distributed). In our context, this can be interpreted that D_i^k measures how diverse the chosen routes are within a bus user's journey data sample, where the route choice diversity increases when the *richness* of the user's choice set (the number of routes that are considered as alternatives by the user) increases and when the *evenness* of the route usage (the relative frequency of the used routes) increases. The value of D_i^k in Eq. (1) ranges between $1/J_i^k$ and 1, where $D_i^k = 1/J_i^k$ occurs when all J_i^k routes have been equally used (e.g., $p_{i,1}^k = p_{i,2}^k = \cdots = p_{i,J_i^k}^k = \frac{1}{J_i^k}$) reflecting the maximum diversity and $D_i^k = 1$ occurs when only one route is used and all the other routes have not been used at all (e.g., $p_{i,j}^k = 1$ and $p_{i,1}^k = \cdots = p_{i,j-1}^k = p_{i,j+1}^k = \cdots = p_{i,J_i^k}^k = 0$) reflecting no diversity in the user's route choice behaviour. The fact that a higher value of D_i^k represents lower diversity is somewhat counterintuitive, however, because it is natural to expect that higher diversity index means higher diversity.

Route *stickiness* considered in this study may be viewed as an inverse of the concept of *diversity*, measured by D_i^k, as the case described by the lowest

diversity ($D_i^k = 1$) is what we describe as the highest level of stickiness (e.g., a user 'sticks' to only one route) and the case described by the highest diversity ($D_i^k = 1/J_i^k$) is what we describe as the lowest level of stickiness (e.g., a user has no tendency to stick to one particular route). One may, thus, suggest that we could directly use D_i^k as a stickiness index. There is, however, an important difference between diversity and stickiness. While diversity takes into account both *richness* and *evenness*, stickiness is more strongly related to *evenness* rather than *richness*. Considering that *evenness* measures the diversity of the choices made by a user (e.g., how each of the available routes is used) and *richness* measures the diversity of the choice set itself (e.g., how many routes are available), stickiness in this study aims to characterise the distribution of chosen routes *within* a given choice set and, therefore, the size of the choice set itself is not considered as part of the behaviour that characterises the user's stickiness tendency, but rather external to the user. To remove the effect of choice set *richness* on stickiness index and allow the stickiness index to be compared across different users and ODs with varying choice set sizes, diversity index D_i^k needs to be normalised by the size of choice set. To achieve this, we define the *Stickiness Index* (SI) of user *i* travelling OD-pair *k*, S_i^k using a simple transformation as follows:

$$S_i^k = \frac{J_i^k \times D_i^k - 1}{J_i^k - 1} \qquad (2)$$

where J_i^k denote the number of routes that have been actually used by user *i* when travelling OD-pair *k*. Given all possible routes identified for OD-pair *k*, which is captured by J^k columns in the User-Route matrix in Figure 1, J_i^k represents the size of user *i*'s actual choice set, where the (actual) choice set refers to a subset of the J^k routes that are chosen by user *i* repeatedly (i.e., more than once or a pre-defined small number of times). In other words, J_i^k corresponds to the number of columns with non-zero $n_{i,j}^k$ entries in the i^{th} row of the User-Route matrix ($J_i^k \leq J^k$). The value of S_i^k in Eq. (2) ranges between 0 and 1, where $S_i^k = 0$ represents no stickiness and $S_i^k = 1$ represents the maximum stickiness.

To illustrate how our stickiness index S_i^k differs from the diversity index D_i^k, Figure 2 shows values of S_i^k and D_i^k at various levels of *richness* (i.e., the number of alternative routes J_i^k) under three different *evenness* scenarios:

Scenario 1 (low evenness), Scenario 2 (medium evenness), and Scenario 3 (high evenness), respectively. Scenario 1 considers the case where a user uses one dominant route 90% of the time and uses the other routes equally for the remaining 10% of the time. For instance, if a user makes 100 trips for a given OD and there are three routes available to the user, it is assumed that 90 trips are made on one route and 10 trips are made equally on the other two routes with each accounting for 5 trips. If this user has six alternative routes instead of three, then 90 trips are made on one route and 10 trips are made on the other five routes with 2 trips for each route. Scenario 2 considers the case where 50% of route usage is allocated to one dominant route and the other 50% is evenly distributed across the remaining routes in the choice set. Scenario 3 is the case where all the available routes are used equally. Observations from Figure 2 are summarised as follows:

- For **Scenario 1** (low evenness), diversity index *D* decreases very slowly (i.e., diversity increases very slowly) with a corresponding increase in the number of alternative routes (dark blue circles). Although diversity is seen to increase as richness increases, this is not the case for Scenario 1 because of decreasing evenness (i.e., since one dominant route always makes up 90% of the usage and the remaining routes make up some fraction of 10% of the usage, as the number of routes increases this fraction then becomes smaller and evenness decreases). For the same scenario, stickiness index *S* increases with increasing number of alternative routes (dark blue triangles). This can be interpreted as the user who 'sticks' to one route despite a large number of available routes exhibiting a higher stickiness than a user who 'sticks' to one route given a small number of available routes.
- For **Scenario 2** (medium evenness), overall, the values of *D* and *S* (light blue circles and triangles) are smaller than those in Scenario 1 (i.e., diversity increases and stickiness decreases as evenness increases). As the number of routes increases, *D* decreases (more quickly than in Scenario 1) and *S* increases, showing similar patterns described in Scenario 1.
- For **Scenario 3** (high evenness), since all routes are equally used, evenness is at its highest level regardless of the number of routes. In Figure 2, *D* decreases with increasing number of routes (red circles), reflecting that diversity increases with increasing richness given the same level of evenness. For stickiness, *S* remains at its lowest level (i.e., zero stickiness) regardless of the number of routes, reflecting

the situation when a user shows no tendency towards repeatedly choosing a specific route, given any numbers of routes.

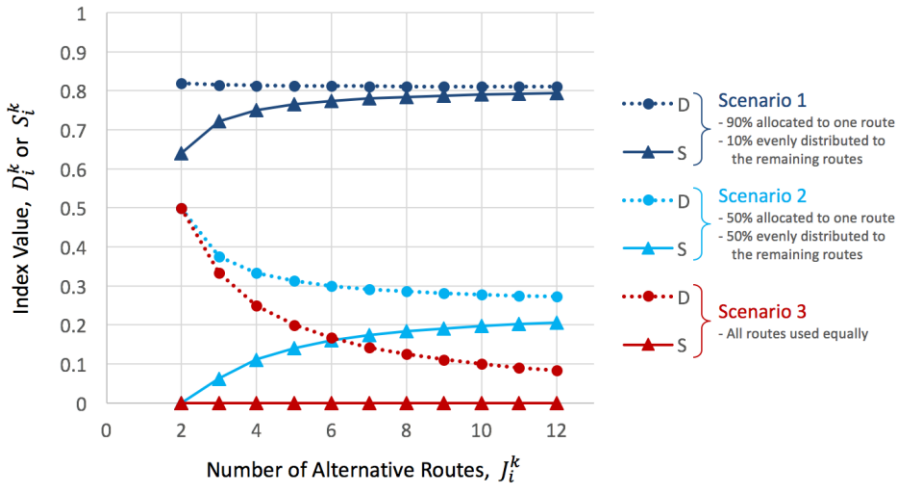

Figure 2. Values of diversity index D_i^k and stickiness index S_i^k at different numbers of alternative routes J_i^k under low, medium, and high evenness scenarios (Scenario 1, 2, and 3, respectively).

OD-level Stickiness Index

Given the user-level SI (S_i^k) obtained for each row of the User-Route matrix in Figure 1, the OD-level SI, denoted by S^k, is defined as a weighted average of S_i^k with users' journey frequencies as the weights as follows:

$$S^k = \sum_{i=1}^{I^k} w_i^k S_i^k \qquad (3)$$

where w_i^k denotes the weight of user i within the user group for OD-pair k, defined as $w_i^k = \frac{N_i^k}{\sum_{i=1}^{I^k} N_i^k}$. The OD-level SI reflects the average tendency of the users of a particular OD and allows us investigate whether there are any geographic patterns associated with bus users' route choice stickiness.

It is worth noting that, although the SI has a clear physical meaning in that its value quantifies a user's route usage frequencies in terms of a level between 0 (all routes are used equally) and 1 (only one route is used), there

is no direct interpretation that could be considered ‘good’ or ‘bad’. Rather the SI should be used as a relative measure describing that a particular user or OD is ‘stickier’ or ‘less sticky’ than another.

4. Study Area, Data and Empirical Approach

4.1 Study Area and Data

Brisbane, the state capital of Queensland, Australia constitutes the study area. Brisbane is the nation’s third most populous city with a population of 2.31 million inhabitants (Australian Bureau of Statistics (ABS), 2016). Common to other metropolitan areas in Australia, Brisbane is a car-oriented with 64.5 per cent of trips being made by private vehicle and 12.8 per cent by public transport (Australian Bureau of Statistics (ABS), 2013). Its public transit system comprises buses, trains, and ferries.

We draw on 6 months of smart card data that have been supplied by TransLink (Brisbane’s sole transit agency). The smart card database contains a total of 69,573,878 transaction records for all trips made across all modes (bus, rail and ferry) between November 2012 and April 2013. The smart card system in Brisbane (called *go* card) is the principal means by which in excess of 80 percent of all urban public transport passengers pay their transit fares. As illustrated in Pelletier et al. (2011), smart card data are collected by touching the smart card against an on-board reader at the time of boarding which are subsequently sent and stored on a central server. It is noteworthy that in Brisbane, public transport passengers are required to touch their *go* cards against the on-board readers at the time of both boarding and alighting. Hence, the Brisbane smart card database drawn on in this study includes information about both boarding and alighting locations and times providing the necessary information to generate journey trajectories. Other important transaction information collected by the *go* card system includes route ID, direction (i.e., inbound or outbound), card ID, journey ID and trip ID. Translink defines a *trip* as the act of travelling from point A to point B with no transfers (a single trip) and a *journey* as the act of travelling from the origin to the final destination, which may involve one or a number of trips (transfers) using different transport modes (TransLink, 2016). Translink uses ‘60 minutes’ as a threshold to determine ‘transfer’ of trips in that two consecutive trips are considered as one journey if a card holder touches on for the second trip within 60 minutes of touching off on the previous trip. If the time between touching-off of the previous trip and touching-on of the

next trip is longer than 60 minutes, these two consecutive trips are considered as two separate journeys. Each journey ID represents a single journey (i.e. trip chains or linked trips) made by the same card holder and trip ID indicates each trip segment within the journey.

4.2 Estimation of Stickiness Index (SI)

To measure the route stickiness of bus passengers, we compute the SI using Eq. (2). The computation process entails the following four analytic steps:

- Generate bus passenger journey trajectories from smart card data.
- Identify OD pairs and available routes for each OD.
- Construct the User-Route matrix for each OD.
- Calculate user-level and OD-level SIs.

Each of these four analytic steps is now discussed in detail.

Generating Journey Trajectories

In order to obtain the full journey trajectory of a given passenger, it is necessary to first identify the list of locations and times of the intermediate stops between the passenger's boarding and alighting points recorded in the smart card database. This can be achieved by simply matching the smart card data with the associated General Transit Feed Specification (GTFS) data. The GTFS, developed by Google, is a common data format for describing a public transit system's scheduled operations and associated geographic information and has been widely adopted by public transit agencies around the world to publish and share their transit data (Google Developers, 2015). The GTFS data consists of a collection of comma-separated values (CSV) files, which provide information on routes, trips, stops, stop times, route shapes, and so on. TransLink adopts the GTFS to describe the bus, rail, and ferry services and this study draws on these data that are made available online (Queensland Government, 2016). After matching and processing the *go* card data, a total of 24,040,538 bus journey trajectories were extracted, which were comprised of 813,593 users, suggesting that on average a user took 5 bus journeys per month (or 1 or 2 journeys per week) during the coverage of our data.

Identifying OD-pairs and Routes

Following the capture of the set of journey trajectories, the next step is to construct the User-Route matrix. This is achieved by via a two-step clustering process, where journey trajectories are first clustered into OD

groups based on their origin and destination points and then, within each OD group, journeys are clustered into route groups based on their whole trajectories. Given the large volume of trajectory data, it is computationally impractical to attempt such a clustering process in a naïve way. To reduce the computation time, we adopt three strategies:

- *The use of a subset of data to identify OD groups*: A set of significant OD-pairs that are repeatedly used and undergo minimal change need to be identified. Since our smart card dataset contains the data for the entire smart card user population, any major OD-pair that is used by at least one person throughout the day will be detected from a single days' worth of data. To address the potential of day-of-week variations in the OD-pair set, data covering a full week are included in the subset. After experiments, we determined that the OD-pair set, once identified from one full-week of data, is consistent from week to week. As such, we employ 9 days of data that cover 5 weekdays from one selected week and 4 weekends (that include 2 Saturdays and 2 Sundays) from two weeks to derive the list of representative OD-pairs for our study region.
- *Clustering origin and destination points*: In order to avoid creating multiple instances of similar OD-pairs, two origin (destination) points are grouped into the same origin (destination) if they are within 500 metres of one another. This strategy is illustrated in Figure 3, in which two neighbouring origin points are grouped into a single origin region. A grid-based spatial index is used to accelerate query processing for neighbouring origin and destination point lookups.
- *The use of simplified trajectories to identify route options for an OD*: Trajectory clustering requires a way to determine the similarity between trajectories and measure similarities based on all points (corresponding to each bus stop comprising a given route) within trajectories is computationally expensive. Instead, we use simplified trajectories to more efficiently verify trajectories' route and shape similarity. Simplified trajectories are created by approximating each original trajectory with 4 representative points using the Douglas-Peucker algorithm (Douglas and Peucker, 1973). Two trajectories are grouped into the same route group if all intermediate point pairs between trajectories are within 500 metres of one another. This process is illustrated in Figure 3, where simplified trajectories 1 and 2 form one route group (as their 2nd and 3rd point pairs are within the

respective 500 metre radius) and simplified trajectory 3 forms another route group, giving two alternative route options for the given OD. Testing established that 500 metres represented the most appropriate radius to perform this operation, offering sufficient simplification to reduce computational demands whilst maintaining the necessary differentiation across trajectories.

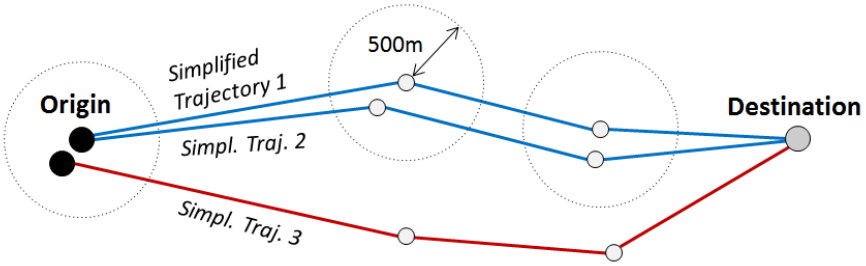

Figure 3. Illustration of the proposed simplified approach to identifying origin-destination groups and the associated route alternatives

Constructing User-Route Matrix

Drawing on the list of representative OD-pairs and the set of route options for each OD-pair generated in the previous steps to next construct the User-Route matrix for each OD as shown in Figure 1. Before constructing this matrix, we impose the following two criteria to exclude invalid ODs and routes.

- *Valid OD Criterion*: An OD-pair must contain a minimum of 10 users having a minimum of 2 journeys each for that OD.
- *Valid Alternative Route Criterion*: A route must be used at least twice by a user in order for that route to be considered as an alternative route for the given user and enter the user's choice set. The threshold of two was used as the minimum required frequency to ensure the route set represents the routes that are used repeatedly, excluding those routes that might have been used only once and therefore potentially signify trips that might be considered 'sporadic'. The routes that were not used by a given user were not considered as the route choice set for that user because we cannot confidently determine whether this user was aware of that route but

did not use it or the user did not know about that route at all. As such, this study defines the alternative route set as the routes that are available and known to as well as are likely to be used again by the given user. With this criterion, the entries with $n_{i,j}^{k} < 2$ in the User-Route matrix will be ignored and replaced with $n_{i,j}^{k} = 0$.

Computing Stickiness Indices

Using the User-Route matrix constructed for each OD-pair, user-level SI (S_i^k) and OD-level SI (S^k) are computed using equations. (2) and (3), respectively.

4.3 Regression Analysis

To examine the factors explaining a bus user's route choice stickiness, we compute a set of regression models. Two complementary types of regression model are employed, Ordinary Least Squares (OLS) and Quantile regression. Quantile regression has been shown to offer a useful complement to OLS, given that it uses select points in the dependents' distribution (such as the median) to produce coefficient estimates that are unaffected by extreme values at the tails. This is especially the case when examining our SI that follows a non-normal distribution. Examining both the mean (OLS) and median values (along with additional points in the dependent's distribution) is important and provides useful information regarding the distribution of the overall SI profile. Table 2 presents the list of variables used in the regression analyses.

Table 2. List of variables used in bus passenger route stickiness modelling

Category	Variable	Description
DEPENDENT VARIABLE		
	S_i^k	*User-level Stickiness Index (SI) defined for user i of OD-pair k*
INDEPENDENT VARIABLES		
User characteristics	*#Journeys per day*	*The average number of journeys made by user i per day across all journey groups*
	OD Usage Fraction	*The fraction that OD-pair k is used by user i compared to all other ODs used by that user*
	#Alternatives	*The number of alternative routes that user i has used to travel OD-pair k*

OD characteristics	*# OD Users*	*The number of users travelling OD-pair k during the six-month observation period*
	Distance between Origin and CBD	*Distance between the origin point and the centre of Brisbane CBD*
	Distance between Destination and CBD	*Distance between the destination point and the centre of Brisbane CBD*
	Is Outbound	*1 if OD-pair k is in the outbound direction; 0 if OD-pair k is in the inbound direction*
Journey characteristics	*Is Weekend*	*1 if journeys are made during weekends 0 if journeys are made during weekdays*
	Is PM	*1 if journeys are started during PM periods 0 if journeys are started during AM periods*
	Travel Time Variation	*The coefficient of variation (CV) of the average route travel times experienced by user i across all his/her alternatives routes*
Interaction	*#Alternatives × Travel Time Variation*	*The product of # Alternatives and Travel Time Variation variables*

The dependent variable is user-level SI (S_i^k) measured for each user *i* and OD *k*. Independent variables are selected to capture *user-specific characteristics* (for user *i*), *OD-specific characteristics* (for OD-pair *k*), as well as *journey-specific characteristics*. To capture journey characteristics, journey trajectories are divided into four journey groups based on the day-of-week (DOW) and time-of-day (TOD) of each journey's start time: {Weekday-AM, Weekday-PM, Weekend-AM, Weekend-PM} and calculate SI for each journey group. As such, for each pair of user and OD (*i* and *k*), we produce four SI measures associated with different DOW-TOD journey groups. The definitions for each independent variable are described next.

User Characteristic Variables

#Journeys per day is defined as the average number of journeys made by user *i* per day. This variable indicates whether the user is a frequent user (e.g., a regular commuter) or an occasional user. The value is the same across all four DOW-TOD journey groups.

OD Usage Fraction is defined as the fraction that OD-pair *k* is used by user *i* compared to all other ODs used by the same user. This variable is an indicator of how important a given OD *k* is to user *i*.

#Alternatives is defined as the number of alternative routes that user *i* has used for travelling a given OD-pair *k*, which corresponds to J_i^k in Equation (2). A route is considered as an alternative route for user *i* (or enters user *i*'s choice set) if it has been used at least twice by the user. This variable is an indicator of the number of options available to user *i*, or in other words the size of user *i*'s choice set.

OD Characteristic Variables

#OD Users is defined as the total number of users travelling OD-pair *k* during the six-month observation period. This is an indicator of the popularity and busyness of a given OD-pair.

Distance between Origin and CBD is defined as the Euclidean distance between a given origin of a trip to the centre of Brisbane's Central Business District (CBD). This variable is employed to characterise the relative proximity of an origin to the city centre.

Distance between Destination and CBD is defined as the Euclidean distance between the destination point and the centre of Brisbane CBD. Similar to *Distance between Origin and CBD*, this variable is employed to characterise the location of a destination in terms of its relative distance from the city centre.

Is Outbound is an indicator variable that takes the value 1 if a given OD-pair is in the outbound direction (specified as when the origin is closer to Brisbane CBD than the destination) and 0 if it is in the inbound direction (in other words, the destination is closer to Brisbane CBD than the origin), indicating whether the associated journeys are travelling towards or away from the city centre.

Journey Characteristic Variables

Is Weekend is a binary indicator variable that takes the value 1 if S_i^k is measured based on the journeys made on Saturday or Sunday (weekend journey group) and 0 if the journey is made on a weekday (weekday journey group).

Is PM a binary indicator variable that takes the value 1 if S_i^k is measured based on journeys made after 12pm (PM journey group) and 0 if it is measured based on the journeys made before 12pm (AM journey group).

Travel Time Variation is defined as the Coefficient of Variation (CV) of average route travel time across different routes used by user *i* travelling OD-pair *k*. We first extract travel time data experienced by user *i* from their journey trajectories associated with OD-pair *k*. Next, we obtain the average travel time for each alternative route and estimate the CV of these average route travel times across the user *i*'s alternative route set. This variable is employed as an indicator of the variability of route travel times across all alternatives.

Interaction Variable

#Alternatives×Travel Time Variation is an interaction term that is employed to capture the interaction effects between *#Alternatives* and *Travel Time Variation*. This is based on the assumption that the effect of the number of alternatives on SI might be varying at different levels of travel time variation across the alternatives and vice versa.

In order to directly compare the relative importance of independent variables, independent variables are standardized by applying the standardization formula $X' = \frac{X-\bar{X}}{s(X)}$, where $\bar{X}$ and $s(X)$ are the mean and standard deviation of variable X.

5. Results and discussion

5.1 Stickiness Index Estimation Results

Following processing the trajectory data, a total of 100,373 distinct OD-user (*k*, *i*) pairs were identified that are then used as input for the Stickiness Index (SI) computation. These distinct OD-user (*k*, *i*) pairs involve 5,460 OD pairs, 685 origins, 631 destinations, and 82,706 users. For each OD and user, the SI was measured for four day-of-week (DOW) and time-of-day (TOD) groups {Weekday-AM, Weekday-PM, Weekend-AM, Weekend-PM}. It is noted that not all OD-user pairs have these four SIs as some users may not travel during particular DOW and TOD periods. Similarly, some ODs may not have a sufficient number of users that are required to be a valid OD for

particular DOW or TOD. Table 3 shows the number of the estimated SIs for each DOW-TOD group, in conjunction with other descriptive statistics such as mean, standard deviation, minimum, and maximum. The sample "Combined" represents the combined sample of SIs from all four DOW-TOD groups. The composition of the combined sample is Weekday-AM (60.9%), Weekday-PM (33.8%), Weekend-AM (2.7%), and Weekend-PM (2.6%), indicating that 94.5% of the SI observations are obtained from weekday journeys.

The estimated SI values range from 0.0 to 0.9208 as shown in Table 3. While the theoretical range of SI is between 0 and 1, the SI estimates in our experiment are always less than 1 as we only consider users who have used at least two alternative routes with each route travelled at least twice, as a way to identify a user's available choice set. The value of 1 can only occur when a user has only used one alternative route given multiple possible alternative routes that are available to the user. The histogram and empirical cumulative distribution function (CDF) of SI are presented in Figure 4. Overall, the distribution of a SI is highly right skewed with a long tail, as can be seen in Figure 4(a) for the combined sample. The peak in SI is observed close to zero, indicating that a large portion of the users show no or a very low level of stickiness, i.e., users tend to use alternative routes with equal likelihood. There are, however, a significant portion of the sample that show a high SI value. For instance, approximately 6 percent of the sample accounted for the SI range greater than or equal to 0.5. If we consider a user with two alternative routes, the SI of 0.5 occurs when the user has used one route 85 percent of the time and the other 15 percent of the time. Similarly, the SI of 0.8 occurs when the user uses one route 95 percent of the time and the other 5 percent of the time.

The SI values in Table 3 show that the mean and the standard deviation of the combined sample are 0.1447 and 0.1731, respectively, indicating a high level of dispersion, which can be identified from the fact that the standard deviation is greater than the mean. Overall, the mean SI is smaller in the Weekend samples (0.0891 and 0.0843) than in the Weekday samples (0.1493 and 0.1456) suggesting that users are less sticky (to a particular route) when it comes to the weekend journeys than for weekday journeys. The difference between weekday and weekend samples can also be identified from the empirical CDFs in Figure 4(b).

Table 3. Stickiness Index (SI) estimation results: summary statistics for different day-of-week and time-of-day groups

Sample	#Obs. (%Sample)	Mean	Std. Dev.	Min.	Max.
Combined	100373 (100%)	0.1447	0.1731	0.0000	0.9208
Weekday-AM	61123 (60.9%)	0.1493	0.1728	0.0000	0.9200
Weekday-PM	33920 (33.8%)	0.1456	0.1788	0.0000	0.9208
Weekend-AM	2759 (2.7%)	0.0891	0.1222	0.0000	0.7257
Weekend-PM	2571 (2.6%)	0.0843	0.1202	0.0000	0.7432

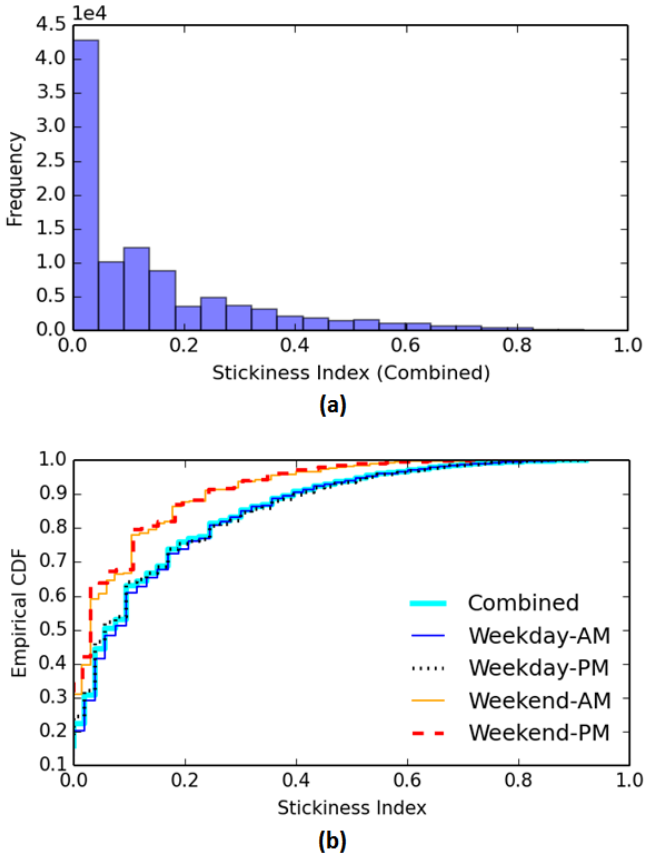

Figure 4. Stickiness Index (SI) estimation results: (a) histogram of SI for the combined sample (b) empirical cumulative distribution functions (CDF) for day-of-week and time-of-day groups

5.2 Regression Results

OLS and quantile regression are computed across on the 100,373 observations and the model results are presented in Table 4. As noted previously, we standardized the independent variables before the regression so that we can compare the relative importance of independent variables based on the magnitude of the estimated coefficients.

OLS Regression

First, we examine the main effects of the variables. The OLS results show that *#Journeys per day* ($\beta = 0.0335$; $p < 0.01$), *OD Usage Fraction* ($\beta = 0.042$; $p < 0.01$), and *Travel Time Variation* ($\beta = 0.0546$; $p < 0.01$) have the greatest effects on the SI (see Figure 5). These variables are each positively associated with the SI, suggesting that a user tends to stick to one route (*Stickiness Index* ↑) when the user is a more frequent bus user (*#Journeys per day* ↑) and when the user travels his/her frequent ODs (*OD Usage Fraction* ↑). Due to the interaction term, the coefficient of *Travel Time Variation* is interpreted as the main effect of *Travel Time Variation* on the stickiness when *#Alternatives* = 0, i.e., the number of alternatives is at its standardized mean value. Given that #Alternatives = 0, the positive value of the *Travel Time Variation* coefficient indicates that a user tends to stick to one route (*Stickiness Index* ↑) as the travel time difference among alternative routes increases (*Travel Time Variation* ↑). The next significant variables based on their coefficient magnitudes are *Distance between Origin and CBD* ($\beta = -0.0144$; $p < 0.01$), *Is Outbound* ($\beta = 0.0123$; $p < 0.01$), and *Is Weekend* ($\beta = -0.0142$; $p < 0.01$). This indicates that a user is stickier to a particular route when the user starts journeys from areas closer to the CBD (*Distance between Origin and CBD* ↓), when the user travels away from the CBD (*Outbound* = 1), and when the user travels during weekdays (*Is Weekend* = 0). With the relatively less effect, *Is PM* ($\beta = -0.0056$; $p < 0.01$) suggests that SI is higher during AM periods than during PM periods (*Is PM* = 0). Variables *#Alternatives* ($\beta = -0.0010$; $p > 0.1$), *#OD Users* ($\beta = 0.0014$; $p < 0.01$), and *Distance between Destination and CBD* ($\beta = -0.0025$; $p < 0.01$) have the coefficient values close to zero, showing the least effects on SI.

The interaction effect captured by *#Alternatives* × *Travel Time Variation* ($\beta = -0.0204$; $p < 0.01$ shows that there exists a significant interaction. For *Travel Time Variation*, this means that the effect of the route travel time difference on SI is different at different values of the number of alternative routes. With the interaction term, the effect of *Travel Time Variation*

becomes $0.0546 - 0.0204 \times \#Alternatives$, which indicates that the route travel time difference becomes a more significant factor when the number of alternative routes is smaller. For instance, a user tends to stick to one route, i.e., a high level of stickiness is observed, when there are only a few alternative routes and the travel time on one route is significantly shorter than the other routes.

Table 4. Ordinary least square regression and quantile regression at the 25th, 50th, 75th and 95th percentile stickiness index values (number of observations = 100,373).

Category	Independent variable	Coefficient β (Standard Error)				
		OLS regression	**Quantile regression at 0.25 quantile**	**Quantile regression at 0.5 quantile**	**Quantile regression at 0.75 quantile**	**Quantile regression at 0.95 quantile**
User characteristics	*#Journeys per day*	0.0335*** (0.001)	0.0080*** (0.000)	0.0256*** (0.000)	0.0549*** (0.001)	0.1010*** (0.001)
	OD Usage Fraction	0.0420*** (0.001)	0.0103*** (0.000)	0.0317*** (0.000)	0.0659*** (0.001)	0.1100*** (0.001)
	#Alternatives	-0.0010 (0.002)	-0.0054*** (0.001)	-0.0084*** (0.001)	0.0079*** (0.002)	0.0207*** (0.003)
OD characteristics	*#OD Users*	0.0014*** (0.001)	0.0007*** (0.000)	0.0011*** (0.000)	0.0019*** (0.001)	-0.0025** (0.001)
	Distance between Origin and CBD	-0.0144*** (0.001)	-0.0032*** (0.000)	-0.0088*** (0.001)	-0.0207*** (0.001)	-0.0237*** (0.002)
	Distance between Destination and CBD	-0.0025*** (0.001)	-0.0008*** (0.000)	-0.0022*** (0.001)	-0.0008 (0.001)	0.0024 (0.002)
	Is Outbound	0.0123*** (0.001)	0.0039*** (0.000)	0.0094*** (0.001)	0.0135*** (0.001)	0.0108*** (0.002)

Journey characteristics	*Is Weekend*	-0.0142*** (0.001)	-0.0040*** (0.000)	-0.0093*** (0.000)	-0.0148*** (0.001)	-0.0175*** (0.001)
	Is PM	-0.0056*** (0.001)	-0.0033*** (0.000)	-0.0059*** (0.000)	-0.0076*** (0.001)	-0.0081*** (0.001)
	Travel Time Variation	0.0546*** (0.002)	-0.0039*** (0.001)	0.0216*** (0.001)	0.1034*** (0.002)	0.1903*** (0.003)
Interaction	*#Alternatives × Travel Time Variation*	-0.0204*** (0.002)	0.0130*** (0.001)	0.0017 (0.002)	-0.0568*** (0.003)	-0.1132*** (0.003)
	Intercept	0.1447*** (0.000)	0.0275*** (0.000)	0.0954*** (0.000)	0.2271*** (0.001)	0.4432*** (0.001)

*, **, ***: *Coefficients significantly different from zero at 10% (p<0.10), 5% (p<0.05), and 1% (p<0.01) significance level, respectively.*

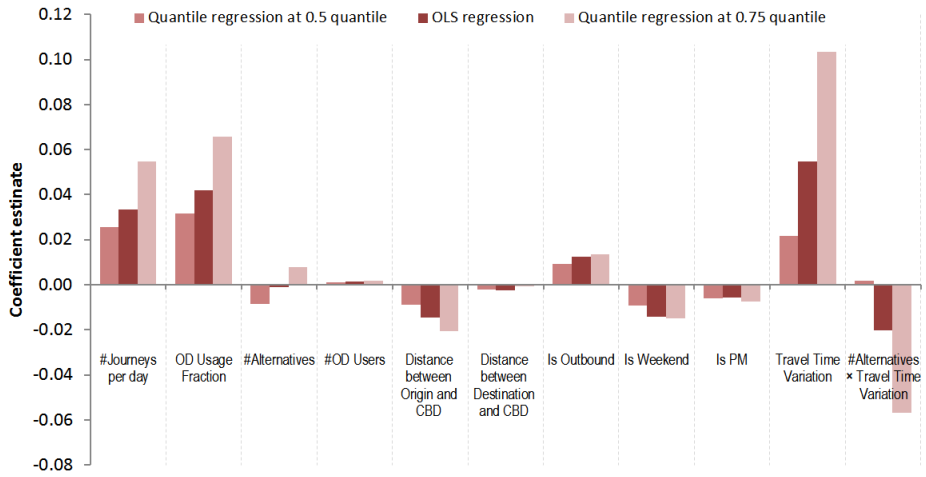

Figure 5. Comparison of estimated coefficients for OLS regression (only show variables with *, **, *** signs in Table 4)

Quantile Regression

Table 4 shows the quantile regression results for the 0.25, 0.5, 0.75, and 0.95 quantiles and Figure 6 presents the quantile regression estimates for each variable. The shaded grey area around the quantile estimate lines in each of the plots depicts the 90 percent pointwise confidence interval. The horizontal dashed line shows the OLS estimate, in conjunction with its 90 percent confidence interval shown by the two dotted lines. Overall, the quantile regression coefficients are significantly different from the OLS coefficients at upper and lower quantiles, indicating that the change in a lower or upper quantile of the SI produced by a one unit change in a given independent variable is different from the change in the mean of the SI produced by a one unit change in the same independent variable.

In all of the cases, the quantile regression estimates vary by quantile, suggesting that the effects of the independent variables on the dependent variable are not constant across different quantiles of the dependent. Figure 6 (a) and (b) show that the coefficients of *#Journeys Per Day* and *OD Usage Fraction* increase with quantile, indicating that increases in these variables lead to greater increases in SI at high values of SI, conditional on the values of the independent variables. For *#Alternatives* in Figure 6 (c), the coefficient initially degreases in the negative direction up to 0.5 quantile and then increases in the positive direction, crossing the zero line at around 0.65 quantile. After around 0.95 quantile, the coefficient again starts decreasing, but we will confine our discussion to the results between 0.05 and 0.95 quantiles. The pattern in this range indicates that *#Alternatives* influences SI negatively at lower values of SI (0.05 to 0.65 quantile), while influencing SI positively at higher values of SI (0.65 to 0.95 quantile). This suggests that having more alternatives makes users with a low SI become less sticky, while making users with a high SI become stickier. Figure 6 (d), shows that the coefficient of *#OD Users* is uniform over the quantile range of 0.05 to 0.8 and, in particular, for estimates between 0.2 and 0.8 quantiles that lie within the confidence intervals for the OLS regression. This suggests that the effects of *#OD Users* on SI is constant over the majority of the SI distribution. For the higher values of SI (> 0.8 quantile), however, the coefficient decreases sharply towards a negative value, indicating that an increase in *#OD Users* leads to greater decreases in the SI in the upper tail of the SI distribution. For Figure 6 (e) *Distance Origin-CBD,* (g) *Is Outbound*, (h) *Is Weekend*, and (i) *Is PM*, the signs of coefficients remain consistent

across all quantiles and their effects are shown to become stronger in the upper quantiles. A stronger effect at higher values of SI indicates that these variables are a stronger predictor of stickiness when a certain level of stickiness tendency exists in the data than when there is no or little stickiness. For *Distance Destination-CBD* in Figure 6 (f), the relationship is negative at lower quantiles with a decreasing slope of the quantile plot. The slope, however, starts increasing after 0.65 quantile, leading to the relationship becoming positive after 0.80 quantile. A potential interpretation of this sign change might be that users with low stickiness are more likely to drop their preference toward a specific route (e.g., use whatever available routes more evenly) when travelling to further outer suburbs, whereas users with high stickiness tend to be even more stickier as their destinations are toward outer areas. The extent to which this interpretation has a clear meaning on stickiness behaviour is quite small, however, as the coefficient is near zero across all quantiles, showing that this variable has a relatively low effect on stickiness compared to other independent variables. Figure 6 (j) and (k) show that the effects of *Travel Time Variation* and *#Alternatives* × *Travel Time Variation* on SI are also close to zero over the lower half of the SI distribution. At the upper half of the distribution, however, the effects of these variables become stronger and the slopes of quantile plots become steeper. Both the separate and combined effects of *#Alternatives* and *Travel Time Variation* become stronger at higher values of SI. For instance, at 0.95 quantile, the effect of *Travel Time Variation* on SI becomes $0.1903 - 0.1132 \times \#Alternatives$ (see Table 4), which has a steeper slope and larger constant term than the expression obtained from the OLS regression (i.e., $0.0546 - 0.0204 \times \#Alternatives$). This suggests that the effect of difference in route travel time across alternative routes becomes more significant for users with higher stickiness. Also, the tendency that route travel time difference plays a more important role when the number of available routes is smaller becomes stronger in the high-stickiness group than in the low-stickiness group.

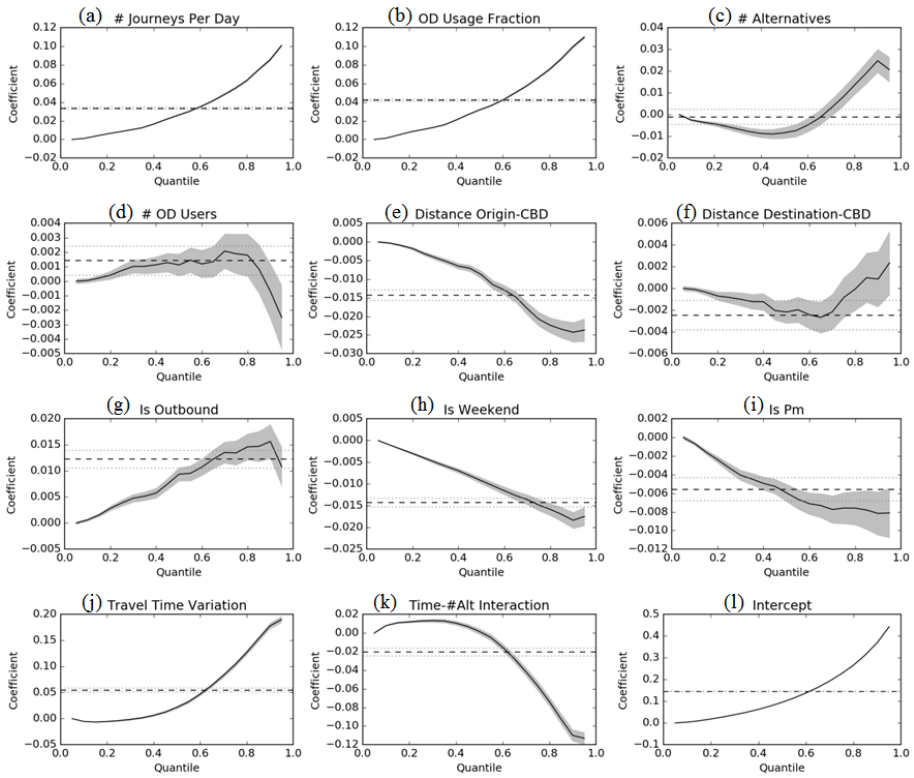

Figure 6. Comparison of estimated coefficients across different regression methods (only show variables with *, **, *** signs in Table 4)

Next, we investigate how the relative importance of independent variables changes for different quantiles of SI. Figure 7 presents pie charts depicting the relative magnitudes of regression coefficients estimated at 0.25, 0.5, 0.75 and 0.95 quantiles. A noticeable change is observed in *Travel Time Variation*, where its proportion increases as the quantile increases (7% → 17% → 30% → 32%), indicating that *Travel Time Variation* becomes more important at higher values of SI. Another observation is that the relative importance of *#Journeys Per Day* and *OD Usage Fraction* remain relatively consistent across quantiles with only a marginal increase at the 0.5 quantile, which are 14% → 20% → 16% → 17% for *#Journeys Per Day* and 18% → 25% → 19% → 18% for *OD Usage Fraction*, respectively.

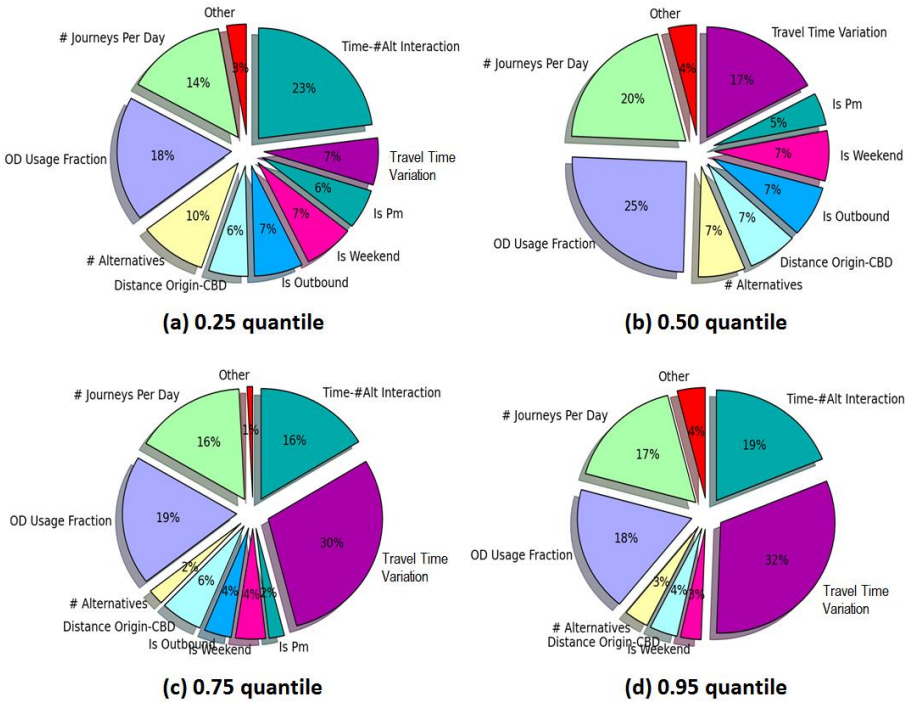

Figure 7. Relative magnitudes of estimated coefficients at different quantiles

5.4 Geographic Patterns in OD-level Stickiness Index

Figure 8 is a matrix of 24 maps that collectively depicts the spatial distribution of SI. This spatial distribution is examined for the top 3 origins and destinations (determined by the total number of users travelling between a given OD pair) distinguishing between morning and afternoon/evening trips in addition to those made during the week and weekend. The location of the origin or destination is shown with the triangle symbol. Variations in SI are captured at a given origin or destination using a variable circular symbol where larger circles are indicative of higher levels of SI and smaller circles indicate low SI. The locations of the circles correspond to the associated destinations (for a given origin) or origins (for a given destination).

The number of locations and associated users for a given origin or destination in addition to a measure of spatial autocorrelation (in addition to its statistical significance), namely Moran's I (Moran, 1950) is employed to

capture the degree to which SI exhibits a spatial distribution that is either clustered, dispersed or random. These 3 metrics are printed in the bottom left corner of each map.

Starting with a visual inspection of the spatial patterns across the matrix, we find distinct variations in SI between weekday and weekend trips as well as for journeys made during morning and afternoon/evening periods. Here we can see that in general people exhibit lower levels of SI during the weekends, an expected finding given that weekend travellers tend to be less temporally constrained allied with weekends tending to involve activities that are less structured (in terms of both timing and location) compared to weekdays. We argue that both of these features combine to explain why we see comparatively low SI for both the top 3 origins and destinations and for morning and afternoon/evening periods. For weekday trips we find some interesting variations between both origins and destinations as well as between morning and afternoon/evening journeys. For origins, the afternoon/evening period appears to be associated with higher levels of SI, whereas for destinations it is the morning period that exhibits higher SI and the relative difference to the afternoon/evening period appears greater. This can be explained by drawing on the regression results associated with *#Journeys per day* and *OD Usage Fraction* which indicate that regular commuters tend to show higher SI. Since the top 3 origins and destinations are located in the CBD, journeys *from* these top 3 origins would include more regular commuting trips in the afternoon/evening period than in the morning (e.g., work-to-home trips) thereby showing relatively high SI in the afternoon, while journeys *to* the top 3 destinations would contain more regular commuting trips in the morning period (e.g., home-to-work trips) leading to higher SI in the morning.

Next, looking at spatial autocorrelation, we find that although there is some variation in I values across the matrix, in all but one instance, each map exhibit significant ($p<0.01$) evidence of positive spatial autocorrelation or in other words, SI is spatial clustered, where high SI locales are co-located next to other high SI locations, and low SI proximate to other low SI locations. There is some evidence to suggest that destinations exhibit higher levels of spatial autocorrelation than origins (given their generally higher I values) and that this is particularly the case for weekdays and the morning period. As noted previously we would expect such spatial clustering of SI given that weekdays in general tend to be associated with higher levels of habituality

wherein there are typically relatively high levels of fixity in terms of the timing and location of activities. As such, it follows that we observe higher levels of SI during this period given the more routinized nature of weekdays over weekends.

Overall, the study findings suggest that SI varies by OD-pair and that there is a certain systematic pattern in its geographical distribution. Observed differences in the SI by OD-pair can stem from both different travel behaviour (e.g., users of some ODs may be stickier than users for other ODs) and the network structure (e.g., transit service on some ODs leads to higher SI than that for other ODs). For instance, the regression results show that user-specific factors such as #Journeys per day and OD Usage Fraction have high impact on SI. As such, when a certain OD has a high SI, it may be because its users are more frequent bus users (*#Journeys per day* ↑) and/or because that OD is a major OD to its users (*OD Usage Fraction* ↑). The network structure or spatial differences in transit service can also influence the differences in the SI. For instance, a user may choose a route with higher frequency more likely and, hence, ODs with large difference in service frequency across their available routes may result in high SI values.

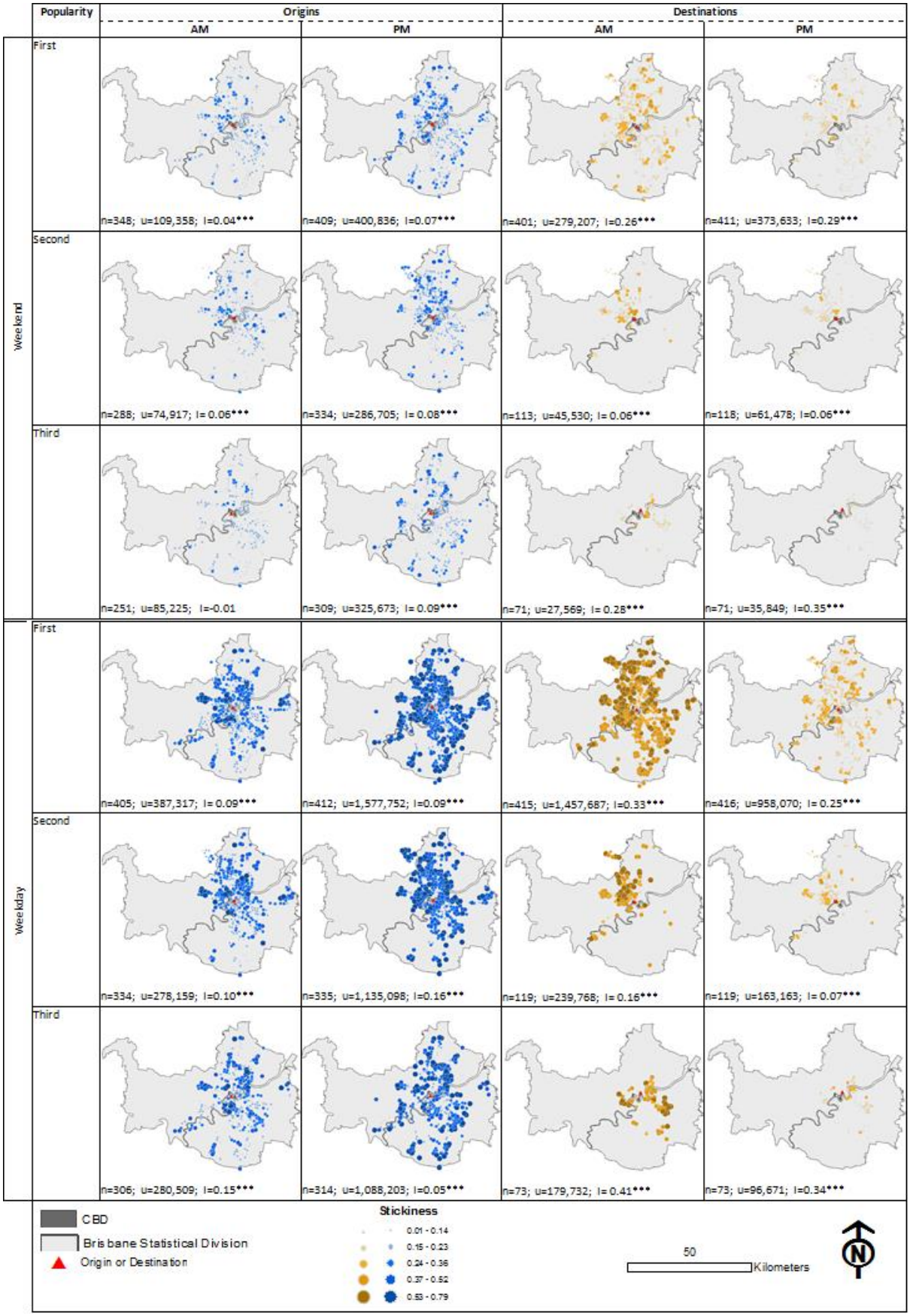
Popularity
Origins
Destinations
AM
PM
AM
PM
Weekend
Weekday
First
Second
Third
n=348; u=109,358; I=0.04***
n=409; u=400,836; I=0.07***
n=401; u=279,207; I=0.26***
n=411; u=373,633; I=0.29***
n=288; u=74,917; I= 0.06***
n=334; u=286,705; I= 0.08***
n=113; u=45,530; I= 0.06***
n=118; u=61,478; I=0.06***
n=251; u=85,225; I=-0.01
n=309; u=325,673; I= 0.09***
n=71; u=27,569; I= 0.28***
n=71; u=35,849; I=0.35***
n=405; u=387,317; I= 0.09***
n=412; u=1,577,752; I=0.09***
n=415; u=1,457,687; I=0.33***
n=416; u=958,070; I= 0.25***
n=334; u=278,159; I=0.10***
n=335; u=1,135,098; I=0.16***
n=119; u=239,768; I= 0.16***
n=119; u=163,163; I= 0.07***
n=306; u=280,509; I=0.15***
n=314; u=1,088,203; I=0.05***
n=73; u=179,732; I= 0.41***
n=73; u=96,671; I=0.34***
CBD
Brisbane Statistical Division
Origin or Destination
Stickiness
0.01 - 0.14
0.15 - 0.23
0.24 - 0.36
0.37 - 0.52
0.53 - 0.79
50
Kilometers
N

Figure 8. The spatial distribution of OD-level stickiness index represented by (a) origin and (b) destination

6. Conclusions

Our study has added to the travel behaviour literature by introducing the notion of 'stickiness' in public transit users' route choice behaviour to investigate the habitual aspects of users' day-to-day trip making decisions. We define the route stickiness of a transit user as the user's tendency towards 'sticking' to only one route regardless of multiple available routes for a given OD-pair. To quantify this tendency and compare it across different users and OD-pairs, we propose the *Stickiness Index* (SI) by employing Simpson's Diversity Index, a commonly used metric that measures biodiversity in ecology, and adapting it to the measurement of route choice (non-)diversity or stickiness. The SI quantifies the range of preferences of users that always select to travel on the same route (high stickiness) to those with a more varied patterns of route selection (low stickiness), permitting the habitual behaviours of individual bus passengers' route choice decisions to be quantified and modelled to reveal its drivers and dynamics.

Drawing on 6 months of public transit smart card data in Brisbane, Australia, this study estimated SI values for a total of 82,706 bus users over 5,460 OD-pairs. To understand what explains SI, we develop OLS and quantile regression models using the SI of a particular user for a particular OD as the dependent variable and a set of user-specific, OD-specific, and journey-specific attributes as independent variables. The regression analysis results suggest that a user's SI increases when the user is a more frequent bus user (*#Journeys per day* ↑) and the user travels his/her frequent ODs (*OD Usage Fraction* ↑). The results also suggest that SI tends to increase when the travel time difference among available routes is high (e.g., the travel time on one route is significantly shorter than the other routes) and the positive effect of the travel time difference on the SI becomes more significant when the number of alternative routes gets smaller. The study further investigated the spatial distribution of SI by visualising geographic patterns of OD-level SIs for selected major origin and destination locations. A visual inspection reveals that in general people exhibit higher levels of SI on inbound trips (to the Brisbane CBD) in the morning period whereas for outbound trips (from the Brisbane CBD) it is the afternoon period that exhibits higher SI. This is consistent with the results from the regression

analysis, which indicates that SI is higher when trips are more constrained in terms of both timing and location (e.g., commuting trips). In addition, a spatial autocorrelation analysis reveals that SI tends to be spatially clustered, where high SI locales are co-located next to other high SI locations, and low SI proximate to other low SI locations. Overall, the study findings show that there exist systematic spatial and temporal patterns in SI, suggesting possibilities of better understanding and even predicting bus passengers' route choice stickiness tendency. Knowing which routes are associated with the highest levels of user-level stickiness is of practical importance to transit agencies as an indicator of their sensitivity to shifts in timetabling. Furthermore, routes identified as low SI could be an indicator of where adjustments to timetabling could be made with minimal impacts to users' routine travel behaviour.

This study has stepped out a new analytic framework with the capacity to be re-deployed in other situational contexts. To this end, this study represents the first step towards a comprehensive empirical understanding of how variations in individual travel behaviour aggregate to contribute to system-wide dynamics. Given access to a both longer coverage of data as well as characteristics of individual users (for example, child, student, adult and senior citizen) our current analytic framework can be extended to consider how persistent our findings remain given variations in both urban design (for example, the development of a new master planned residential housing estate) allied with changes in the design and operation (for example the introduction of a new busway) of the public transport network and the travel behavioural response by different user groups.

Transport researchers, urban planners and transport policy makers must be mindful of the importance and role that travel behaviour plays when planning, designing and re-designing urban spaces and their associated public transport accessibility. In this paper we have offered a way in which we can build a new evidence base with the potential to contribute to transport policy through an understanding of travel behaviour dynamics. The use of such information will become increasingly important as the call to design urban environments in a manner that best facilitates the use of public transport and overcomes the sustainable transport challenges associated with urbanisation becomes increasingly important. Adopting an evidence-based approach will ultimately help us to plan, build and organise both the built

form and public transport provision in our cites in a way that best facilitates a more sustainable future.

Acknowledgements

This research is supported by Queensland Department of Transport and Main Roads (TMR), under the Transport Academic Partnership (TAP) agreement with the University of Queensland. We would like to thank the TransLink division of TMR for access to the data on which the paper is based. The interpretations of the analysis are solely those of the authors and do not necessarily reflect the views and opinions of TransLink or any of its employees. We would also like to extend our gratitude to anonymous reviewers for their constructive comments and suggestions.

References

Australian Bureau of Statistics (ABS) [WWW Document], 2016. . Reg. Popul. Growth Aust. 2014-15. URL http://www.abs.gov.au/ausstats/abs@.nsf/mf/3218.0 (accessed 5.2.16).

Australian Bureau of Statistics (ABS) [WWW Document], 2013. . 2011 Census QuickStats. URL http://www.censusdata.abs.gov.au/census_services/getproduct/census/2011/quickstat/0?opendocumen&navpos%C2%BC220 (accessed 5.2.16).

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transp. Policy, Road User Charging: Theory and PracticesW. Saleh 12, 464–474. doi:10.1016/j.tranpol.2005.06.008

Chu, K.K.A., Lomone, A., 2016. Reproducing Longitudinal In-Vehicle Traveler Experience and the Impact of a Service Reduction with Public Transit Smart Card Data. Transp. Res. Rec. J. Transp. Res. Board 2541, 81–89. doi:10.3141/2541-10

Chudyk, A.M., Winters, M., Moniruzzaman, M., Ashe, M.C., Gould, J.S., McKay, H., 2015. Destinations matter: The association between where older adults live and their travel behavior. J. Transp. Health, Transport, travel and mobility in later life 2, 50–57. doi:10.1016/j.jth.2014.09.008

Dieleman, F.M., Dijst, M., Burghouwt, G., 2002. Urban Form and Travel Behaviour: Micro-level Household Attributes and Residential Context. Urban Stud. 39, 507–527. doi:10.1080/00420980220112801

Douglas, D.H., Peucker, T.K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Cartogr. Int. J. Geogr. Inf. Geovisualization 10, 112–122. doi:10.3138/FM57-6770-U75U-7727

Gärling, T., Axhausen, K.W., 2003. Introduction: Habitual travel choice. Transportation 30, 1–11. doi:10.1023/A:1021230223001

Google Developers [WWW Document], 2015. . Gen. Transit Feed Specif. GTFS. URL https://developers.google.com/transit/gtfs/ (accessed 3.19.16).

Gordon, P., Kumar, A., Richardson, H.W., 1989. Gender Differences in Metropolitan Travel Behaviour. Reg. Stud. 23, 499–510. doi:10.1080/00343408912331345672

Goulet-Langlois, G., Koutsopoulos, H.N., Zhao, J., 2016. Inferring patterns in the multi-week activity sequences of public transport users. Transp. Res. Part C Emerg. Technol. 64, 1–16. doi:10.1016/j.trc.2015.12.012

Handy, S., 1996. Methodologies for exploring the link between urban form and travel behavior. Transp. Res. Part Transp. Environ. 1, 151–165. doi:10.1016/S1361-9209(96)00010-7

Herfindahl, O.C., 1950. Concentration in the steel industry. Columbia University.

Hirschman, A.O., 1945. National Power and the Structure of Foreign Trade. University of California Press.

Hong, J., Shen, Q., Zhang, L., 2014. How do built-environment factors affect travel behavior? A spatial analysis at different geographic scales. Transportation 41, 419–440. doi:10.1007/s11116-013-9462-9

Innocenti, A., Lattarulo, P., Pazienza, M.G., 2013. Car stickiness: Heuristics and biases in travel choice. Transp. Policy 25, 158–168. doi:10.1016/j.tranpol.2012.11.004

Jánošíková, Ľ., Slavík, J., Koháni, M., 2014. Estimation of a route choice model for urban public transport using smart card data. Transp. Plan. Technol. 37, 638–648. doi:10.1080/03081060.2014.935570

Kieu, L.M., Bhaskar, A., Chung, E., 2015. Passenger Segmentation Using Smart Card Data. IEEE Trans. Intell. Transp. Syst. 16, 1537–1548. doi:10.1109/TITS.2014.2368998

Kotval-K, Z., Vojnovic, I., 2015. The socio-economics of travel behavior and environmental burdens: A Detroit, Michigan regional context. Transp. Res. Part Transp. Environ. 41, 477–491. doi:10.1016/j.trd.2015.10.017

Kurauchi, F., Schmöcker, J.-D., Shimamoto, H., Hassan, S.M., 2014. Variability of commuters' bus line choice: an analysis of oyster card data. Public Transp. 6, 21–34. doi:10.1007/s12469-013-0080-x

Kwan, M.-P., 2016. Algorithmic Geographies: Big Data, Algorithmic Uncertainty, and the Production of Geographic Knowledge. Ann. Am. Assoc. Geogr. 106, 274–282. doi:10.1080/00045608.2015.1117937

Lee, S.G., Hickman, M., 2013. Trip purpose inference using automated fare collection data. Public Transp. 6, 1–20. doi:10.1007/s12469-013-0077-5

Liu, Y., Bunker, J., Ferreira, L., 2010. Transit Users' Route-Choice Modelling in Transit Assignment: A Review. Transp. Rev. 30, 753–769. doi:10.1080/01441641003744261

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transp. Res. Part C Emerg. Technol. 36, 1–12. doi:10.1016/j.trc.2013.07.010

Ma, Z.-L., Ferreira, L., Mesbah, M., Hojati, A.T., 2015. Modeling Bus Travel Time Reliability with Supply and Demand Data from Automatic Vehicle Location and Smart Card Systems. Transp. Res. Rec. J. Transp. Res. Board 2533, 17–27. doi:10.3141/2533-03

Moran, P.A.P., 1950. Notes on Continuous Stochastic Phenomena. Biometrika 37, 17–23. doi:10.2307/2332142

Morency, C., Trépanier, M., Agard, B., 2007. Measuring transit use variability with smart-card data. Transp. Policy 14, 193–203. doi:10.1016/j.tranpol.2007.01.001

Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. Transp. Res. Part C Emerg. Technol. 24, 9–18. doi:10.1016/j.trc.2012.01.007

Nassir, N., Hickman, M., Ma, Z.-L., 2015. Behavioural findings from observed transit route choice strategies in the farecard data of Brisbane. Presented at the Australasian Transport Research Forum, Department of Infrastructure and Regional Development.

Nishiuchi, H., King, J., Todoroki, T., 2012. Spatial-Temporal Daily Frequent Trip Pattern of Public Transport Passengers Using Smart Card Data. Int. J. Intell. Transp. Syst. Res. 11, 1–10. doi:10.1007/s13177-012-0051-7

Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transp. Res. Part C Emerg. Technol. 19, 557–568. doi:10.1016/j.trc.2010.12.003

Prato, C.G., Bekhor, S., Pronello, C., 2011. Latent variables and route choice behavior. Transportation 39, 299–319. doi:10.1007/s11116-011-9344-y

Queensland Government [WWW Document], 2016. . Gen. Transit Feed Specif. GTFS—South East Qld. URL https://data.qld.gov.au/dataset/general-transit-feed-specification-gtfs-seq (accessed 3.19.16).

Rasouli, S., Timmermans, H., Waerden, P. van der, 2015. Employment status transitions and shifts in daily activity-travel behavior with special focus on shopping duration. Transportation 42, 919–931. doi:10.1007/s11116-015-9655-5

Schlich, R., Axhausen, K.W., 2003. Habitual travel behaviour: Evidence from a six-week travel diary. Transportation 30, 13–36. doi:10.1023/A:1021230507071

Schmöcker, J.-D., Shimamoto, H., Kurauchi, F., 2013. Generation and calibration of transit hyperpaths. Transp. Res. Part C Emerg. Technol. 36, 406–418. doi:10.1016/j.trc.2013.06.014

Simpson, E.H., 1949. Measurement of diversity. Nature 163, 688. doi:10.1038/163688a0

Sun, L., Jin, J.G., Lee, D.-H., Axhausen, K.W., 2015. Characterizing Multimodal Transfer Time Using Smart Card Data: the Effect of Time, Passenger Age, Crowdedness, and Collective Pressure. Presented at the Transportation Research Board 94th Annual Meeting.

Tao, S., Corcoran, J., Mateo-Babiano, I., Rohde, D., 2014a. Exploring Bus Rapid Transit passenger travel behaviour using big data. Appl. Geogr. 53, 90–104. doi:10.1016/j.apgeog.2014.06.008

Tao, S., Rohde, D., Corcoran, J., 2014b. Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. J. Transp. Geogr. 41, 21–36. doi:10.1016/j.jtrangeo.2014.08.006

Thøgersen, J., 2006. Understanding repetitive travel mode choices in a stable context: A panel study approach. Transp. Res. Part Policy Pract. 40, 621–638. doi:10.1016/j.tra.2005.11.004

TransLink [WWW Document], 2016. . Go Card Journey Trip. URL http://translink.com.au/tickets-and-fares/fares/go-card-journey-and-trip (accessed 5.1.16).

Vacca, A., Meloni, I., 2015. Understanding Route Switch Behavior: An Analysis Using GPS Based Data. Transp. Res. Procedia, SIDT Scientific Seminar 2013 5, 56–65. doi:10.1016/j.trpro.2015.01.018

Viggiano, C., Koutsopoulos, H., Attanucci, J., 2014. User Behavior in Multiroute Bus Corridors. Transp. Res. Rec. J. Transp. Res. Board 2418, 92–99. doi:10.3141/2418-11

Yue, Y., Lan, T., Yeh, A.G.O., Li, Q.-Q., 2014. Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. Travel Behav. Soc. 1, 69–78. doi:10.1016/j.tbs.2013.12.002

Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F., Schmitt, G., 2016. Variability in Regularity: Mining Temporal Mobility Patterns in London, Singapore and Beijing Using Smart-Card Data. PLOS ONE 11, e0149222. doi:10.1371/journal.pone.0149222