Data Mining and Modeling for Smart Transit Management

Siddharth Gupta · Mark Hickman · Karthik Srinivasan

Abstract The incorporation of Advanced Data Collection Systems (ADCS) into transportation systems in recent years has made vast datasets available to analysts. However, even today, in several situations the data collected from public transit systems are not in conformity with the transit specification, commonly available as the General Transit Feed Specification (GTFS). This presents a challenge for the evaluation of service characteristics. This paper primarily uses data from the Smart Card (SC) based Automatic Fare Collection (AFC) system implemented in Brisbane's public transportation network. It demonstrates how these data can be used for trip identification and subsequent schedule matching to talk about service performance in a robust and automated manner. The paper then discusses a novel approach for spatio-temporal analysis of journey data. The approach is based on machine learning and has been demonstrated to capture a variety of phenomena in the transit network. Finally, the service reliability and occupancy metrics derived from schedule matching are used to model passengers' route choice in the network.
The techniques developed for schedule matching and spatio-temporal analysis discussed here, though implemented on SC data in this paper, are extensible to other datasets providing journey information as well.

Keywords: Schedule Matching · Spatio-temporal Analysis · Smart Card · Route Choice · Service Performance

Siddharth Gupta
Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, USA
Email: sid1@mit.edu

Mark Hickman
School of Civil Engineering, The University of Queensland,
St. Lucia, Brisbane, Australia
Email: m.hickman1@uq.edu.au

Karthik Srinivasan
Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, India
E-mail: karthikks@iitm.ac.in

1 Introduction

Advanced Data Collections Systems (ADCS) such as Automated Fare Collection (AFC) systems and Automatic Vehicle Location (AVL) systems among others are becoming fairly ubiquitous in public transit networks. These systems provide vast quantities of data for mining and analysis.

Transit operations commonly follow a specification called the General Transit Feed Specification (GTFS) (Google 2014). In this specification, a unique ID identifies each trip in the network on a day. Similar trips operating on multiple days can have the same trip ID. Data from AVL systems have widely been used for the identification of trips (Bertini and El-Geneidy 2003, Furth et al. 2006). Some AFC data have been used for more aggregate measures of transit performance (Trépanier et al. 2009; Navick and Furth 2002), while only very limited research to date has explored the exclusive use of AFC data for vehicle schedule adherence, travel time studies, and trajectory analysis (Chu and Chapleau 2008; Lee et al. 2012; Sun et al. 2012). Very limited work is available regarding trip identification using farecard data.

If a cogent mapping is established between the set of observed trips and the set of scheduled trips, performance metrics can be determined for each trip ID. Some studies have attempted this matching (Stratham 2002 and Bullock et al. 2005). These are either applicable under special conditions, such as large headways or availability of location and timestamp information at the initial stop, or still worse, are manual. Section 2 attempts to overcome these shortcomings and develops a robust methodology to perform schedule matching. An important derivative from a well-founded mapping is the ability to quantify service reliability and occupancy. These have been demonstrated for sample trips in this paper.

Once schedule matching is performed, detailed information about each trip and journey in the transit system becomes available. Information regarding these is available in high resolution of space and time, much higher than traditional modes of collecting information such as through surveys. There has been much discussion (Kobayashi and Miller 2014, Dodge et al. 2008, Dykes and Mountain 2003) on exploiting both spatial and temporal characteristics from such data sets in recent works. There is, however, further scope for improvement in such analysis (Geurs and Wee 2004, Neutens et al. 2011, Páez et al. 2012, Kwan 1998, Wu and Miller 2001, Miller 1991).

Kieu et al. (2014) have focused on spatial and temporal travel regularity from SC data and then use it for transit passenger segmentation. Their study first defines spatial regularity and defines a metric, independent of the actual geography for each user. Similarly, it defines a temporal regularity metric independent of the actual time of day. The methodology presented in Section 3 can be extended to capture spatial and temporal regularity simultaneously over multiple days for each user and can incorporate segmentation in continuous space-time.

Section 4 explores the impact of service characteristics including reliability and occupancy on route choice made by travelers in public transit systems. A large body of literature is available regarding studies related to the causes and measurement of transit service unreliability. It is still not common practice, however, to include service reliability as an explanatory variable in transit route choice models (Outwater and Charlton 2006). In the corridor under analysis, the headway ranges from 15 minutes to 1 hour and therefore, the scheduled time is likely to govern waiting time (Welding 1957). As a result, the common bus line approach (Chrique et al. 1975) may not be applicable in this case.

The paper finally concludes in Section 5 with a discussion of the results and possible improvements for the discussed work.

The data used in this paper are from the farecard system of Brisbane. Fare collection in Brisbane is done for buses, trains and ferries using Go Cards. The AFC system in Brisbane is a closed system, i.e. passengers are required to tap-on and tap-off during their journeys. The dataset is therefore able to provide information regarding boarding and alighting locations and timestamps. It also provides encrypted but consistent IDs of the cards used on the trips. The penetration rate of the farecards is 85-90% of all journeys in Brisbane, thus generating a nearly complete universe of journeys within the transit network.

2 Trip Identification and Schedule Matching using Farecard Data

2.1 Problem description

As discussed, AVL technology provides information about vehicle movements during operations. Yet, in the absence of a strong complementary AVL system, it might be possible to use the farecard data alone to be able to identify these vehicle trips and their mapping to the scheduled operations. AFC systems are, however, prone to a much broader range of data errors. Robinson et al. (2014) have classified these into four groups- hardware, software, user and data, and have explored the causes and consequences of each.

The data, therefore, need to be cleaned with an array of filters and handled with care. Data are available in the form of individual passenger transactions. The route number and an identifier for the transit vehicle are also present. Since a transit vehicle can perform multiple trips on a route, there is a need to segregate the series of observations into individual trip-buckets. Since transactions need not occur at every stop along the route, identified trips are often constituted by arrival and departure times at a subset of stops along the route. Section 2.2, discusses a methodology to identify trips from farecard data.

After trip identification, a correspondence between identified trips and scheduled trips in the GTFS needs to be established. Trips scheduled on a route can follow different stopping patterns. Cut trips are a common observation based on the time of day. The set of stops on a cut trip are a subset of stops on through trips. This, coupled with the fact that the set of stops on identified trips can be a subset of all the stops on a route, can cause observed counterparts of through trips to combine with scheduled cut trips.

When service headways are large, matching can be an intuitive. Deviation at the starting terminal was used by Bullock et al. (2005) to match trips. Information may not always be available at the starting terminal with both AVL and AFC data.

Thus, for low headways with the possibility of missing information at the starting terminal, and the existence of multiple patterns on a route, there is a need for a much stronger methodology for schedule matching. Such a methodology is presented in Section 2.3. Section 2.4 compares the results of this methodology with other attempts after which Section 2.5 discusses potential application. Section 3 goes into exploring spatio-temporal analysis of journeys.

2.2 Identification of trips from farecard data

Figure 1 outlines the proposed methodology for trip identification. It consists of two segments. In the first segment (pre-processing), a series of filters are applied to the available data to discard faulty observations, which are likely to interfere with the process of trip identification. The second segment (trip identification) details the core logic for trip identification.

2.2.1 Pre-processing

This step involves two tasks. The first is a reduction from the entire data available for a route to the logically admissible (valid) data. The second task evaluates the feasibility of occurrence of each transaction in keeping with the characteristics of the route they are observed on (feasible observations). Both the transitions consist of a series of filters, which have been listed below:

Data Validation: In this process, observations that meet items 1 and 2 and do not meet items 3 and 4 from the following are discarded-

1. Boarding stop and time are not null
2. Alighting time and stop are not null
3. Boarding stop is different than the alighting stop
4. Alighting time is after boarding time

Data Evaluation: The GTFS is used to derive the following information for the route under consideration-

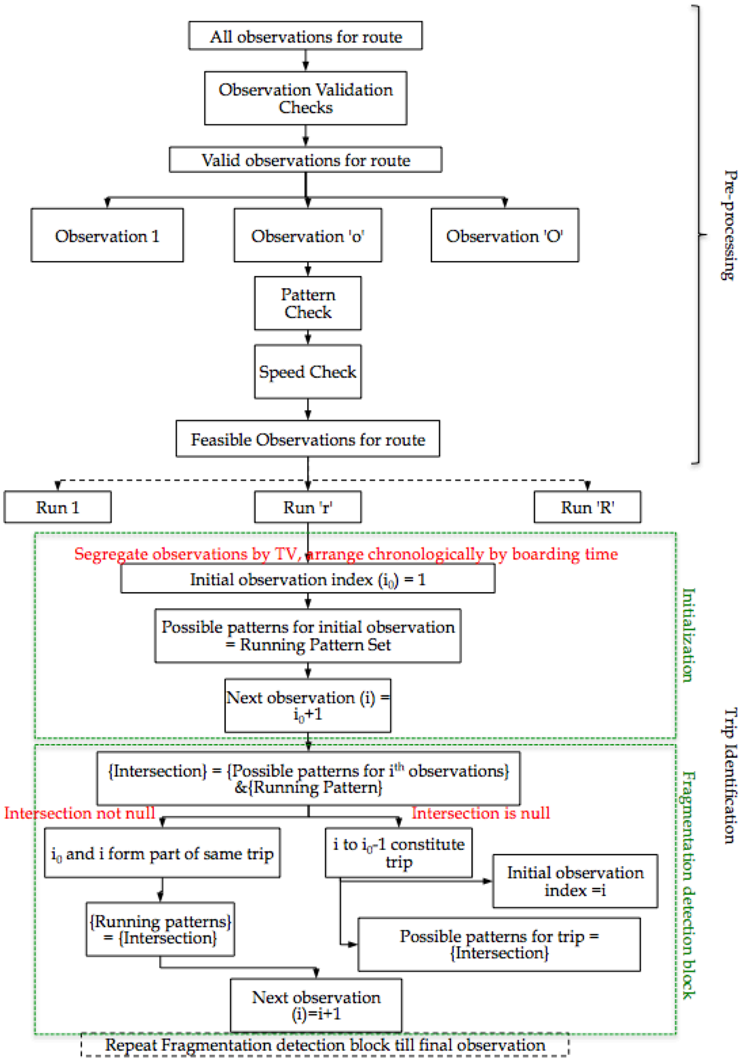
1. The patterns on the route
2. Inter-stop distances

Using this for each observation it is ensured that-

1. The OD pair indicated in the pattern occurs sequentially on at least one of the patterns for the route.

2. The speed indicated by the observation is between 3 and 100kmph. These limits have been arrived at heuristically by examining valid results from a few routes.

An observation that does not meet any of these two criteria is also discarded. The remaining observations are assumed to be feasible on the route under usual operating conditions.



All observations for route
Observation Validation Checks
Valid observations for route
Observation 1
Observation 'o'
Observation 'O'
Pattern Check
Speed Check
Feasible Observations for route
Pre-processing
Run 1
Run 'r'
Run 'R'
Segregate observations by TV, arrange chronologically by boarding time
Initial observation index (i_0) = 1
Possible patterns for initial observation = Running Pattern Set
Next observation (i) = i_0+1
Initialization
{Intersection} = {Possible patterns for i^{th} observations} &{Running Pattern}
Intersection not null
Intersection is null
i_0 and i form part of same trip
i to i_0-1 constitute trip
Initial observation index =i
{Running patterns} = {Intersection}
Possible patterns for trip = {Intersection}
Next observation (i)=i+1
Fragmentation detection block
Trip Identification
Repeat Fragmentation detection block till final observation

Fig.1: Flow chart for trip identification

After completing these pre-processing steps, the data are segregated based on their ‘run number’. The run number is unique to a transit vehicle. Each segregated set is then fed into the trip identification segment.

2.2.2 Trip Identification

The data for each transit vehicle (run) are then sorted chronologically. These data contain observations for each trip performed on the run. They are, however, connected at the ends. The objective of this module is to determine the point of fragmentation between observations that belong to tow distinct but consecutive trips. The functionality of this module can be divided into two blocks.

The first block (Initialization) initializes the process of identification of a new trip and the second (Fragmentation Detection) determines the set of observations that form part of an initialized trip.

Initialization

This block is invoked every time the starting observation of a trip is identified. By default, the trip identification process for a run commences by feeding the first observation in the chronological dataset as input to this block.

This block performs the following functions-

1. Define two sets- 'Running Patterns' and 'Possible Patterns'. 'Running Patterns' stores the set of patterns that a sequence of observations previously analyzed can possibly follow. 'Possible Patterns' stores the set of potential patterns to which the given observation could belong. Initially both of these are null sets.
2. It processes the input observation to determine corresponding feasible patterns and stores them in the 'Possible Patterns' set. Since this is the only observation in the trip thus far, the 'Running Patterns' set is also populated with these patterns.

It calls the next observation in chronological sequence and feeds it into the 'Fragmentation Detection' block

Fragmentation Detection

Both the initialization block and another fragmentation detection block can invoke a Fragmentation Detection block. The inputs to this block include the observation to be analyzed, the 'Possible Patterns' set and the 'Running Patterns' set.

The flow of logic in this block is as follows-

1. It determines the set of patterns that could correspond to the input observation and sets this as 'Possible Patterns'.
2. It finds the intersection of 'Possible Patterns' with 'Running Patterns'.

3. If the intersection is not a null set, all observations from the initializing observation to the current observation are identified to belong to the same trip and the trip is considered ongoing until termination occurs in step 5 below.
4. The set of intersecting patterns is set as 'Running Patterns'.
5. It now determines if the intersection is a null set or observations for the run have been exhausted.

6. The next observation in chronological sequence is used to invoke another Fragmentation Detection Block.

3. If the intersection is a null set, all observations from the initializing observation to the observations upto but not including the input observation are said to constitute a complete trip.
4. 'Running Patterns' is the set of possible patterns for this trip
5. The input observation is used to invoke the Initialization block to initiate the determination of the subsequent trip.

Once all observations for the current run have been exhausted, this process of trip identification is repeated for subsequent runs.

2.2.3 Post-processing

Once route-wise data have been segregated into trip-wise data, it becomes possible to trace individual trip trajectories. The last fragmentation block for each trip in the process of trip identification determines the set of possible patterns that a trip can associated with. This facilitates identification of the sequence of occurrence of stops on the trip. The trip data are then split into boarding and alighting events and are stored in buckets associated with each stop. For each bucket, it can determine the number and timestamp of each boarding and alighting event. This enables the computation of occupancy, arrival time and departure time at each stop, which in turn opens up opportunities for computing reliability and average occupancy metrics (discussed in Section 2.4).

Within the data available in each stop bucket, timestamps from boarding taps are considered reliable. If an alighting tap at a stop is found to more than 2-minute distant from the closest boarding tap, it is not considered during the estimation of boarding and alighting times.

The smallest and largest timestamps in the bucket corresponding to a stop are taken as the arrival and departure time of the bus at/from the stop.

2.2.4 Output and Evaluation

The outputs from the processing stage are trip trajectories and the data corresponding to each trip. For the purpose of this analysis, farecard observations from Route 200 over a period of 30 days are used. The trips identified for March 5, 2013 for an outbound pattern are shown in Figure 2. Similar plots are available for other patterns.

In order to evaluate the quality of schedule matching, the number of trips determined by our algorithm and the number of trips as per the GTFS schedule are compared. The fractions of data utilized at different steps of the matching process are also determined. These are shown in Figure 3 and Figure 4.

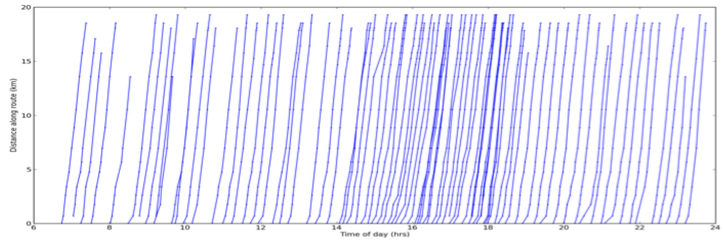


Fig. 2: Trips determined using farecard data on a particular day

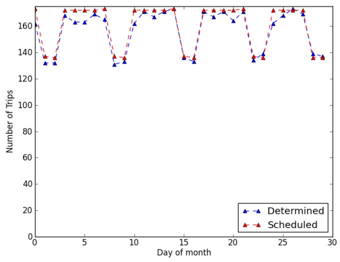


Fig. 3: Trip count from the proposed trip detection algorithm vs. scheduled trip count

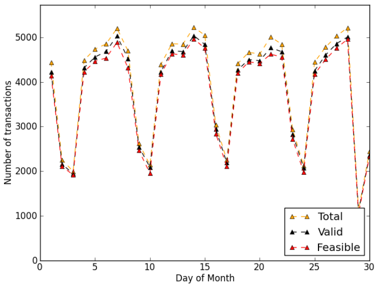


Fig. 4: Data utilization at different stages of trip detection

In Figure 3, it is observed that the determined and observed trip counts are close to each other, with the determined trip count usually being a bit lower (<8% over 30 days). This lower count might be justified by absence of farecard transactions on trips scheduled early in the morning or late at night. Or they could be explained by deviation from scheduled operations, which are likely to manifest as cancellation of scheduled trips rather than introduction of new ones.

In Figure 4, it is seen that only a small fraction (about 3.6%) of the observations are invalid (in complete or not logical). Of the valid observations, almost all observations (97.7%) pass the route-specific filters during the feasibility checks and are therefore incorporated into trips.

2.2.5 Comparison with other known attempts

In a study using BDS and APC data, Strathman (2002) found that about 97% of the records could be matched to the scheduled database and valid APC data were recovered on about 75% of the trips. The methodology proposed here performs equally well on the former metric and significantly better on the latter.

Bullock et al. (2005) in their attempt at schedule matching using GPS data in Sydney reported an overall undetected/misclassification percentage of trips at 13.9%. In comparison, our measure is significantly lower at below 8%.

In Singapore (Lee et al. 2012), data are pre-processed according to journey definition by the Land Transport Authority prior to making them available to researchers. Since pre-processing takes care of inconsistent timestamps, they are readily usable for trip definition.

El-Geneidy et al. (2011) while evaluating service performance using AVL and APC data 'subjected the data to detailed observation to remove extreme travel time'. Our methodology, on the other hand discards the need for detailed manual inspection.

2.3 Schedule Matching

A typical instance of schedule matching problem is shown in Figure 5.

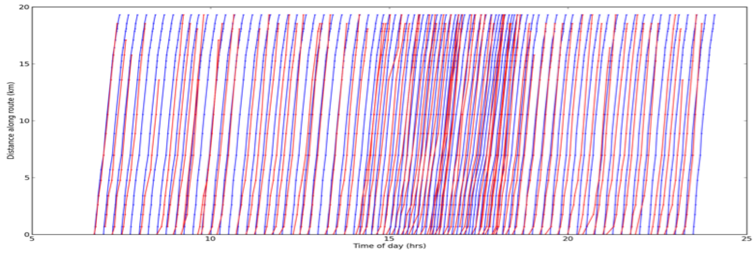


Fig. 5: The Matching Problem- Scheduled trips are in blue and determined trips in red

Section 2.1 has already discussed some of the challenges in schedule matching. A single metric can be defined to compare any scheduled trip-determined trip pair. In this paper, the dissimilarity between a trip pair has been quantified using the mean (per stop) absolute arrival/departure time deviation. This is appropriate since different determined trips have observations available at different number of stops.

Two approached are explored in this paper to automate the process of schedule matching. The first is a heuristic approach to develop a matching algorithm and the second involves the Hungarian algorithm, an algorithm for the minimum cost assignment problem with a strictly polynomial time complexity of $O(n^4)$.

2.3.1 Heuristic Matching

While developing a heuristic approach for matching, the following set of rules are defined for the matching process:

1. For a matching to be valid, the absolute deviation per stop in terms of the departure and arrival times should not be greater than 10mins/stop.
2. When multiple patterns are available on a route, matching for shorter patterns should be performed prior to that for longer ones.

3. Matching should be performed chronologically. This is useful when bus bunching occurs during peak periods.
4. The matching has to yield a one to one mapping i.e. multiple determined trips cannot be matched to the same scheduled trip or vice versa.

The result for this process for one of the outbound patterns is shown in Figure 3.6.

For the case under consideration here, there were a total of 173 scheduled trips. The trip determination process yielded 163 trips. The heuristic approach was able to match 149 of the determined trip to some scheduled trip. The total per-stop deviation was 9.52 hours, and the corresponding average per-stop deviation per trip was 0.064 hours or 3.83 minutes.

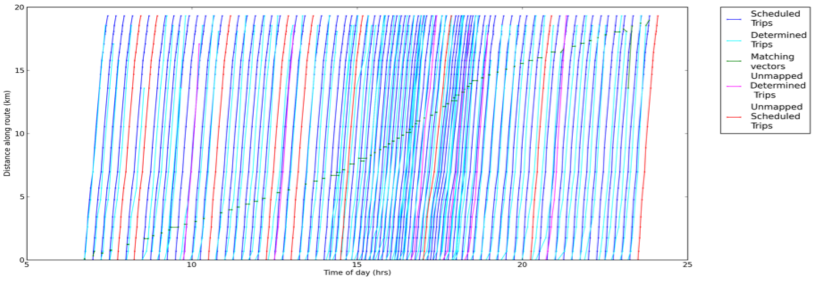


Fig. 3.6: Trip matching using the Heuristic approach

The threshold of 10mins/stop was selected by exploring a range of threshold values ranging from 0 to 30mins/stop.

2.3.2 Matching using Hungarian Algorithm

The Hungarian algorithm is often used for the allocation of *n* tasks to *n* people when the cost of getting each task done by each person is available as a cost matrix. For schedule matching, this matrix was constructed by substituting tasks and people with scheduled and determined trips and the cost for performing the task with the mean absolute deviation between the scheduled and determined trip pair. A provision was included to match unequal number of scheduled and determined trips.

As with the heuristic approach, an upper limit for the deviation between matched trips is defined. This is incorporated by setting the cost of trips with mean absolute deviations higher than the threshold to a very high value (equal for all pairs outside the threshold) and discarding all matched pairs, if any, that are observed in the output from the Hungarian matching. Since, all trip pairs outside the threshold have equal costs associated with them, all pairs with a matching within the threshold can proceed without interference even if a high value matching is observed from the algorithm.

In this implementation of matching, all per-stop deviations in the cost matrix that have a magnitude greater than 15 are set to a high value.

The results for this implementation are shown in Figure 7. All 159 determined trips were matched; the total per-stop deviation was 11.74 hours, and the corresponding average per-stop deviation per trip was 0.074 hours or 4.43 minutes.

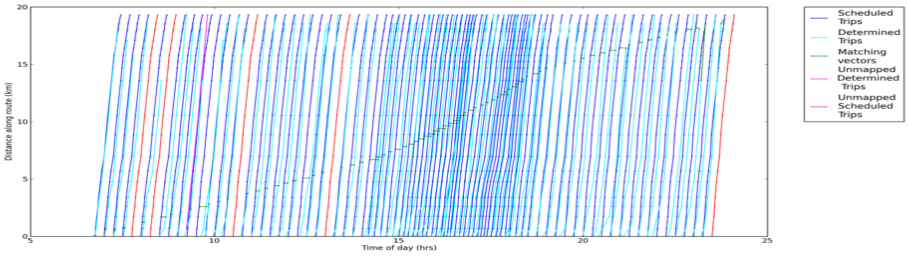


Fig.7: Trip Matching using the modified implementation of the Hungarian Algorithm

2.3.3 Comparison of the two approaches

The results for matching from each of the three methodologies were inspected manually. There is a need for defining a stricter threshold with the heuristic approach since it is not an optimization algorithm and can end up matching all trips with highly deviant pairs so long as they lie within the defined threshold. While it matches a high fraction of trips, some easily available matching pairs were missed.

The results yielded by the implementation of the Hungarian algorithm were in close conformity with those obtained by manual inspection.

The thresholds prescribed in both the approaches can be modified based on the average headway of the service on the route under analysis.

2.4 Computation of reliability and occupancy metrics

This section focuses on computing average reliability and occupancy metric for weekday trips. The GTFS is structured such that trip IDs usually remain same for trips occurring on the same route at the same time of day across weekdays (except holidays). Hence, historical training for a scheduled trip becomes possible since sufficient data points can be obtained in the preceding 2-4 weeks.

Figure 8 demonstrates the results for on-time reliability at each stop on a route from schedule matching with a training period of 4 weeks. This plot can be used to compute an average reliability metric for the trip at each stop. It can also be used to update scheduled in segments where the inter-stop travel times have been wrongly approximated. Figure 9 shows an occupancy profile from a particular day. Such profiles can be obtained for each day in the training period to arrive at an average occupancy metric for each stop.

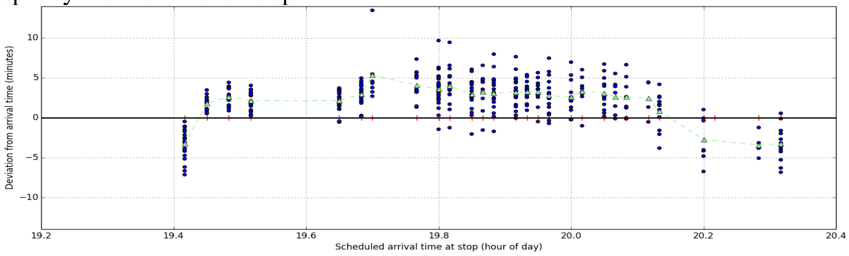


Fig. 8: Training for average reliability using multiple iterations of the same trip

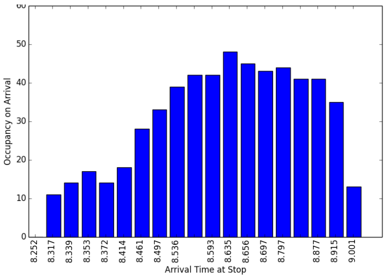


Fig. 9: A typical occupancy plot. Each bar corresponds to a stop. For stops without transactions, the timestamp is not available.

Both these measures can prove to be useful in traveler information systems. Mean values from such plots shall be used to model route choice in Section 4.

3. Spatio-temporal analysis of journeys

This section focuses on a generalized spatial-temporal analysis framework for journeys in any transportation network. High-resolution travel datasets facilitate analysis in a space-time continuum. When these data are organized in a 3-dimensional framework (Figure 10), they generate a cloud of journey data. Each region in this cloud can be associated with its own density, which can be defined differently based on know characteristics of the journeys. The greater the amount of information associated with each journey vector, the broader can be the definition of the density function. After defining the density function, the approach utilizes different clustering approaches to determine similar regions in space. Since each dimension associated with a journey represents a different characteristic (such as travel time, distance, age of person, time of day etc.), the approach needs to meaningfully compare each pair of journey characteristics. The definition of such a comparison metric should incorporate human perception towards each characteristic. This forms the basis for the methodology described in this section.

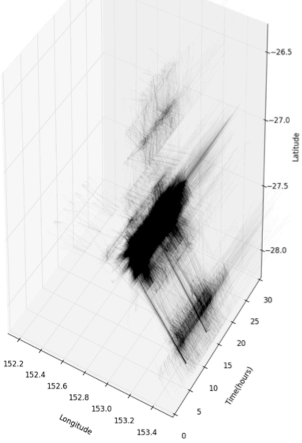


Fig. 10: Journey-vectors (90% transparency) from one day (March 1 2013) in a 3-dimensional space-time framework (total 359,275 observations)

Each journey has been represented as a vector in Figure 10. Though only the OD-locations (latitude-longitude pair) and time stamps are used for this purpose, as demonstrated in Section 2.4, more characteristics can be derived for each journey. These can include journey length, mean experienced occupancy, mean experienced reliability, journey cost and passenger characteristics among others. Hence, each journey can become a multi-dimensional vector. The discussion in this section is confined the application of machine learning clustering techniques on 7-dimensional journey vectors (2 dimensions each for OD locations, one apiece for OD timestamps and the last one for journey length, i.e. the sum of lengths of the trips constituting the journey) and their representation in 3-d space-time. Since different dimensions of a journey vector possess different units (latitude and longitude in degrees while timestamp is measured in hours), a means of cross-dimensional comparison should be defined. The approach for this depends on the phenomenon under analysis.

This methodology is illustrated through two applications-

1. Identification of important flows in the transit network
2. Capturing OD-perception of travelers from revealed travel patterns

3.1 Identification of important flows in the transit network

Important flows in the network can be considered as regions that possess a high concentration of journeys in space-time. Two journeys can be considered close if the difference between the space and time coordinates is reasonably low. Since the study is interested in the difference of latitude and longitude, they can be converted into kilometers by considering 1 degree approximately equal to 111.2km.

To compare distance and time dimensions, Distance Time Equivalence (DTE) is defined. To determine the Distance Time Equivalence (DTE), the headway characteristics of the route under analysis can be used. Most services operate with headway of about 15 minutes in the system. Suppose in any region of the journey cloud, flows are required to be identified when in a 1km radius (2×acceptable walking distance at the origin and destination each, since the mean distance to either stop from any destination in between would be 500m) and a half hour deviation combined at the origin and destination (corresponding to duration of unit headway at the origin and destination) of a journey vector, at minimum number of journey vectors need to be present; the scanning radius corresponds to 2km (=2km-equivalents) and half an hour interval. For a half hour interval to correspond to 2km-equivalent, 1hour has to equal 4km-equivalents. With these assumptions, the DTE equals 4km-equivalents/hour.

With a scanning radius (ε) of 2km-equivalent, these transformed vectors are input to the DBSCAN algorithm with different minimum cluster sizes. The results from a range of minimum cluster sizes are shown in Table 1.

For DTE=4 and ε=2, results are shown in a 3-D framework in Figure 4.2. For the plots corresponding to low permissible densities (minimum cluster size=50 or 100), the cluster in the downtown (CBD) region blows up and encapsulates 100-600 times more journeys than the next largest cluster. These iterations are, however able to capture major inter-city flows, which are missing in iterations with higher density thresholds.

Table 1: Results of DBSCAN for flow determination using the DTE=4 and ε=2

DTE	Min. Size	ε	Cluster Count	Fraction of journeys classified
4	50	2	58	48.62
4	100	2	22	39.12
4	200	2	9	25.71
4	300	2	11	16.43
4	500	2	9	5.54

With higher density thresholds (200, 300 and 500), the CBD cluster is split into the morning and evening peak periods. The order of magnitude of these clusters is less than 10 times that of other clusters. Therefore, they capture inner city flows effectively. The iteration with minimum cluster size of 200 captures important flows in the early morning and late night periods. A cursory inspection of the iteration with minimum size 300 reveals important mid-day flows and that with 500 captures significant intra-city flows in the peak period.

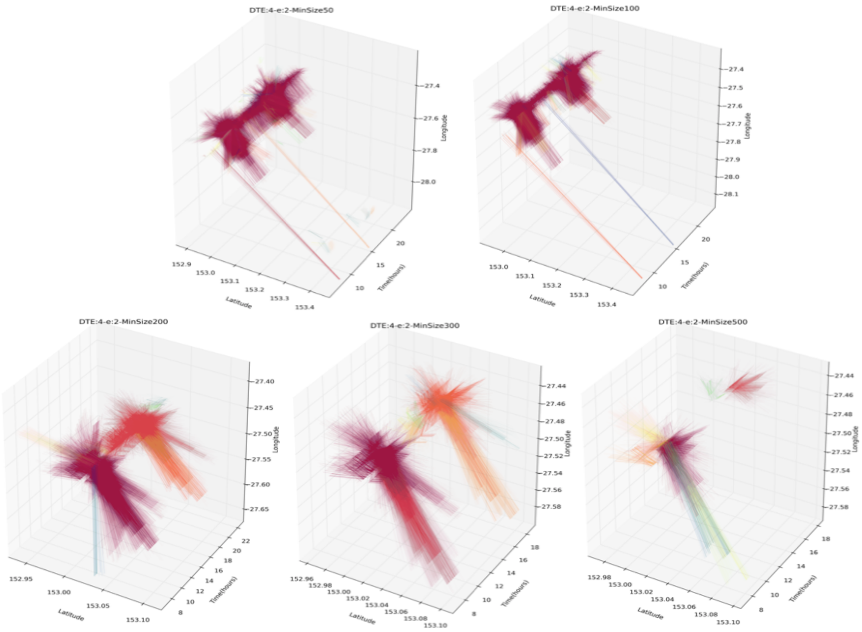


Fig. 11: Results of DBSCAN for the determination of important journey flows in the transit network. The five subfigures differ in the minimum cluster size

Thus, combining the observations from the 6 simulations in Figure 11 can help in generating Origin-Destination-Time matrices relevant to different objectives.

3.2 Capturing OD-perception from revealed travel patterns

This section determines the set of stops that users are likely to perceive as the same origin and destination region for travel. The 6 basic characteristics of each journey vector are considered for this application- latitude, longitude and timestamp at the origin and destination.

As before, the latitude and longitude are converted into kilometers. A large body of literature (Atkinson 1993, Lam and Morrall 1982, Peterson 1968, Shortreed and Maynes 1977, Morency et al. 2011) is available regarding acceptable walking distance to transit stations for passengers. Based on this, an acceptable walking distance of 400m has been assumed to search for viable alternatives. Hence, two journeys are assumed to be similar if the separation between their origin and destination stations is less than 400m each. Since perception can be captured only over a long period of time, a liberal temporal equivalence, 1hr to 1km (DTE=1), is considered permissible for journeys to be considered similar.

Since OD pairs considered interchangeably should have a minimum concentration of flows, a density based clustering technique, DBSCAN, can be used to study this phenomenon.

The results from DBSCAN with DTE=1, epsilon=0.8 and minimum cluster size=50 are shown in Figure 12. Expansion of an example cluster is shown in Figure 13. From Figure 13, it can be seen that the set of stops constituting a cluster are close enough to be considered interchangeably by passengers. This was confirmed by looking into the behaviors of passengers traveling between these OD regions.

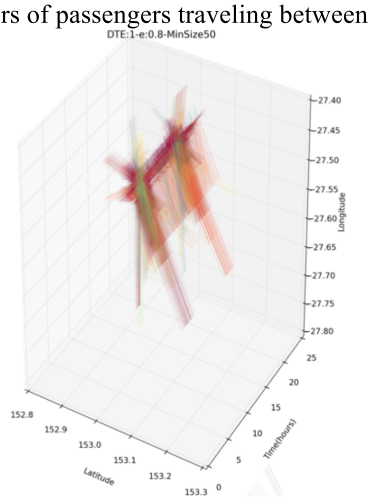


Fig. 12: Capturing OD-perception using the proposed methodology

3.3 Summary

In this section, a framework for spatio-temporal analysis has been successfully developed and applied to two applications. A variety of phenomena can be incorporated into the framework by changing the DTE used for comparing distance and time. Other clustering techniques can also be used when found appropriate for the application under analysis. Other characteristics associated with journey vectors can also be incorporated and analyzed through this framework. This can be extremely useful for analysis of occupancy and reliability characteristics of the network from the metrics developed in Section 2. Section 4 now discusses route choice between the OD pairs shown in Figure 13

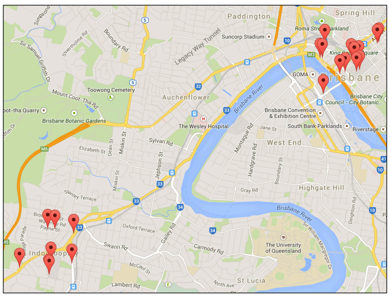


Fig.13: OD-perception captured by the proposed methodology between Indooroopilly and Queen Street Segmentation of journeys

4. Analysis of impact of occupancy and reliability characteristics on transit route choice

OD perception between the Indooroopilly region and the CBD has been captured in Section 3. This section looks at the routes operating between these regions and attempts to model route choice for the same.

Several studies have revealed that crowding and reliability are important factors for passengers while choosing transit services (Petersen and Vovsha 2006, Hollander 2006, Tirachini 2013, Cheng 2010, Griffiths 2006, Fletcher and El-Geneidy 2013, Katz and Rhman 2010, Lundberg 1976). The overall objective is to determine whether, in the absence of corresponding data, passengers are able to make good decisions regarding occupancy and reliability characteristics.

The set of routes operational between these regions during weekdays include route number 425, 430, 435, 444, 453, 454 and 460. All routes passing through the Indooroopilly region have scheduled stop at Indooroopilly Shopping Centre. In the CBD, all routes, except route 444, stop at Queen Street Bus Station. Route 444 terminates at King George Square Station, which is at a distance of approximately 300 meters from Queen Street Bus Station. The bus line for the other routes is common. The headway for routes 425, 453, 454 and 460 is 30 minutes. That for routes 430 and 435 is 60 minutes while 444 is a high frequency BUZZ route with headway of 15 minutes.

For this analysis, the choice set actually considered by a user is not available. It has been assumed that the choice set considered by travelers lies in an interval between 5 minutes before and 15 minutes after the observed choice of travelers. Within the 5-minute period preceding the observed choice, it is assumed that a passenger is likely to have skipped a maximum of 1 bus service, if any were available in the window. With this assumption, a total of 1151 transit route choice tasks were identified during the training period. An example choice task is shown in Table 2.

Table 2: Example choice task curated from the Go Card data

Route	Mean Occupancy	SD Occupancy	Mean reliability	SD reliability	TT (min)	Frequency (hourly)
425	10.8	2.37	3.86	2.89	25	2
435	9.47	2.78	1.10	0.77	22	1
444	15.8	6.77	1.54	0.85	21	4
452	8.87	3.03	2.67	1.57	24	2
460	7.6	3.36	1.99	1.30	21	2

Section 4.1 explores characteristics of the determined choice task. Route choice models that were considered have been discussed in section 4.2. Section 4.3 provides concludes section 4 with a discussion on potential to improve the modeling approach.

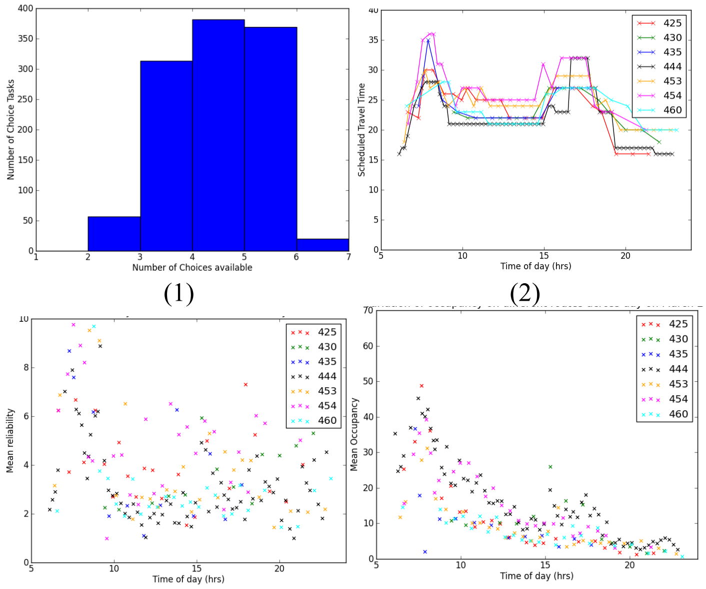
4.1 Exploratory analysis

Based on the chosen time window, services from each of the routes were not necessarily available as alternatives in each choice task. Figure 14.1 shows the histogram of the number of alternatives available in each choice task. Figures 14.2 through 14.4 show the distribution of travel time, reliability and occupancy on each of the routes through the day.

In these profiles a clear peak in the morning period is observed. The evening peak is dampened since the direction of peak travel is outbound- opposite to the direction under analysis. To ascertain that a trade-off occurs between the modeled characteristics while making a choice, the count of dominant, non-dominant and dominated alternatives was determined. It was observed that out of 1151 tasks, 1095 saw non-dominant alternatives being chosen. Dominant alternatives with respect to occupancy-travel time and reliability-travel time were available often (200 and 350 times respectively) but overall non-dominant alternatives were chosen, indicating a tradeoff between the modeled variables.

To capture users' perception, the occupancy, travel time and reliability variables are convertedfrom the continuous scale to categorical. Each of these is converted into a 3-level scale based on likely user perception with the reference set at the first. The levels are listed below-

1. Occupancy: Low (0-20), Medium (21-40) and High (>40)
2. Reliability: On-Time (1 minute early to +2 minutes late), Slightly Delayed (2-5 minutes late) and Significantly Delayed (>5 minutes late)
3. Travel Time: Free Flow (<24 minutes with 20-24 being the common off-peak range), Mildly congested (24-28 minutes) and Highly Congested (>28 minutes)



(1) (2)

(3) (4)

Fig.14: (1) Distribution of number of alternatives available in each choice task, (2), (3) and (4) Variation of travel time, reliability and occupancy on each of the routes over the period of a day.

4.2 Route Choice Models

Multinomial logit modeling, with travel time, occupancy and reliability as factors, yielded statistical significance for travel time and occupancy. The coefficient for travel time, as per intuition turned out to be negative. Occupancy, on the other hand, contrary to intuition, was seen to possess a positive coefficient for both medium and high crowding levels.

Table 3: Significant results from model building

	Estimate	Std. Error	t-value	Pr(>|t|)
Medium Occupancy	0.52	0.10	5.25	1.50E-07
High Occupancy	1.13	0.26	4.31	1.64E-05
TT- Mild Congestion	-0.18	0.08	-2.27	0.02
TT- High Congestion	-0.30	0.14	-2.04	0.04

When an alternative specific constant for service frequency was added to the modeled parameters (Table 4), it was seen to possess strong statistical significance and travel time was noted to lose its significance. Occupancy was still significant albeit with a positive estimate for its coefficient.

Table 4: Results from model building with frequency (Log-likelihood: -1544.6)

	Estimate	Std. Error	t-value	Pr(>|t|)
Medium Occupancy	0.31	0.11	2.87	0.004
High Occupancy	0.74	0.27	2.78	0.005
Frequency	0.16	0.03	5.09	3.54E-07

4.3 Discussion

The models discussed above suggest that frequency is a very important factor in route choice. The utility of a service reduces with increasing travel time. The analysis also suggests that people are not making good decisions with regards to occupancy and also that service reliability is not a significant factor in the case under consideration.

Studies cited from literature have revealed that occupancy and reliability are important with regards to route choice.

There could be several possible explanations for lack of significance associated with these variables-

1. Information regarding occupancy and reliability are not made available to travelers
2. Passengers might be opting for the service arriving first after their arrival irrespective of crowding characteristics
3. The 7 routes available to the OD regions under analysis provide high frequency travel options. With such high frequency services, reliability can be expected to lose significance. This is because with high frequency services, passenger arrivals are more likely to resemble random events rather than schedule driven events.
4. A rail route is available from Indooroopilly (about 400m from Indooroopilly Shopping Center) to the CBD (about 700m from Queen Street at Roma Street). Railway services are known to have higher reliability and more capacity. Passengers sensitive to these characteristics might be opting for rail as their mode of travel.
5. The process of schedule matching is open to some data errors listed in the section on schedule matching.

It might also be possible that for the chosen OD regions under the prevailing conditions, reliability and occupancy do not play a significant role towards route choice. This can, however, be concluded only after a more thorough study.

5. Conclusion

This study has been able to leverage data from an AFC system to provide insights into both operational characteristics of the system as well as the behavioral characteristics of the user. In the end, it also attempts to model route choice based on learning from these insights.

Schedule matching mapped observed traces in the form of passenger transactions to scheduled vehicle traces. Characteristics relevant to service reliability, work utilization, feasibility of transit schedule, driver behavior and several other factors can be studies through this process.

Of special interest to transportation planners can be the determination of locations and time intervals wherein different phenomena occur in the network. For this, the study has suggested a very generalized approach using spatial-temporal analysis. It has also demonstrated the application of this approach to various phenomena and established the validity of the results obtained.

The final section on route choice analysis, though open to future improvement, tends to suggest that information provision with respect to service occupancy might be able to improve travelers' experience on the transit network.

Overall, the outcomes from the study open up pathway for analyzing several interesting phenomena occurring in the transit network. The frameworks suggested for schedule matching and spatio-temporal analysis are extensible to data emanating from other sources as well.

References

Atkinson, W.G. (ed.). Canadian Transit Handbook, Third Edition, Canadian Urban Transit Association, Toronto, 1993.

Bertini, R., and A. El-Geneidy (2003). Generating transit performance measures with archived data. Transportation Research Record 1841, Transportation Research Board, Washington, DC, 109–119.

Bullock, P., Jiang, Q., Stopher, P.R., Using GPS Technology to Determine On-Time Running of Scheduled Bus Services. Journal of Public Transportation, Vol. 8, No. 1, 2005, 21-40

Cheng. Y.H. Exploring passenger anxiety associated with train travel, Transportation, 37 (6) (2010), pp. 875–896

Chriqui, Claude and Pierre Robillard (1975) Common Bus Lines. Transportation Science, 9, 115-121

Chu, K., and R. Chapleau (2008). Enriching archived smart card transaction data for transit demand modeling. Transportation Research Record 2063, Transportation Research Board, Washington, DC, 63-72.

Cox, T., J. Houdmont, A. Griffiths. Rail passenger crowding, stress, health and safety in Britain, Transportation Research Part A: Policy and Practice, 40 (3) (2006), pp. 244–258

Dodge, S., Weibel, R., & Lautenschütz, A. K. (2008). Towards a taxonomy of movement patterns. Information Visualization, 7(3-4), 240-252.

Dykes, J. A., & Mountain, D. M. (2003). Seeking structure in records of spatio-temporal behaviour: visualization issues, efforts and applications. Computational Statistics & Data Analysis, 43(4), 581-603.

Fletcher, G., A. El-Geneidy. The effects of fare payment types and crowding on dwell time: a fine-grained analysis. In: 92nd TRB Annual Meeting, Washington, D.C.

Furth, P., B. Hemily, T. Muller, and J. Strathman (2006). Transit Cooperative Research Program (TCRP) Report 113: Using Archived AVL-APC Data to Improve Transit Performance and Management. Transportation Research Board: Washington, DC.

Geurs, K. T., & Van Wee, B. (2004). Accessibility evaluation of land-use and transport strategies: review and research directions. Journal of Transport Geography, 12(2), 127-140.

Google (2014). Google Transit. https://developers.google.com/transit/, Accessed August 31, 2014.

Hollander, Y. Direct versus indirect models for the effects of unreliability. Transportation Research, Part A: Policy and Practice, Vol. 40, No. 9, 2006, pp. 699-711.

Katz, D., M.M. Rahman. Levels of overcrowding in bus system of Dhaka, Bangladesh, Transportation Research Record, 2143 (2010), pp. 85–91.

Kieu, L. M., Bhaskar, A., & Chung, E. (2014) Transit passenger segmentation using travel regularity mined from Smart Card transactions data. Paper presented at the Transportation Research Board 93rd Annual Meeting, January 2014, Washington, DC.

Kobayashi, T., & Miller, H. (2014). Exploratory visualization of collective mobile objects data using temporal granularity and spatial similarity. In Data Mining for Geoinformatics (pp. 127-154). Springer New York.

Kwan, M. P. (1998). Space-time and integral measures of individual accessibility: a comparative analysis using a point-based framework. Geographical Analysis, 30(3), 191-216.

Lam, W. and J. Morrall. Bus Passenger Walking Distances and Waiting Times: A Summer-Winter Comparison. In Transportation Quarterly, Vol. 36, No.3, 1982.

Lee, D.-H., L. Sun, and A. Erath (2012). Study of bus service reliability in Singapore using fare card data. In Proceedings of the 12th Asia-Pacific Intelligent Transportation Forum.

Lundberg, U. Urban commuting: crowdedness and catecholamine excretion, Journal of Human Stress, 2 (3) (1976), pp. 26–36.

Miller, H. J. (1991). Modelling accessibility using space-time prism concepts within geographical information systems. International Journal of Geographical Information Systems, 5(3), 287-301.

Morency, C., M. Trepanier, and M. Deners. Walking to Transit: An Unexpected Source of Physical Activity. In Transport Policy, Vol. 18, 2011, pp. 800-806.

Navick, D., and P. Furth (2002). Estimating passenger miles, origin-destination patterns, and loads with location-stamped farebox data. Transportation Research Record 1799, 107-113.

Neutens, T., Schwanen, T., & Witlox, F. (2011). The prism of everyday life: towards a new research agenda for time geography. Transport Reviews, 31(1), 25-47.

Outwater, M. L. and B. Charlton. The San Francisco Model in Practice. Validation, Testing and Application. In Conference Proceedings 42. Innovations in Travel Demand Modeling. Summary of a conference. Transportation Research Board of the National Academies, Austin, Texas, 2006, pp. 24-29.

Páez, A., Scott, D. M., & Morency, C. (2012). Measuring accessibility: positive and normative implementations of various accessibility indicators. Journal of Transport Geography, 25, 141-153.

Petersen, E. andP. Vovsha. Directions for Coordinated Improvement of Travel Surveys and Models. In Conference Proceedings 42. Innovations in Travel Demand Modeling. Summary of a conference. Transportation Research Board of the National Academies, Austin, Texas, 2006, pp. 85-88.

Peterson, S.G. Walking Distances to Bus Stops in Washington, D.C. In Traffic Engineering, Vol. 39, No.3, 1968.

Robinson, S., Narayanan, B., Toh, N., & Pereira, F. (2014). Methods for pre-processing smartcard data to improve data quality. Transportation Research Part C: Emerging Technologies, 49, 43-58.

Shortreed, J.H., and D. Maynes. Calibration of a Transit Demand Model for Kitchener Waterloo. Project Report WRI 606-11, Ontario Ministry of Transportation and Communications, Toronto, 1977.

Strathman, J. (2002). Tri-Met's Experience With Automatic Passenger Counter and Automatic Vehicle Location Systems, Appendix A- TCRP Web Document 23 (Project H-28), TRB

Sun, L., D.-H. Lee, A. Erath, and X. Huang (2012). Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system. Proceedings of the ACM UrbComp '12 Conference, Beijing, August 12, 2012.

Tirachini, A. Crowding in public transport systems: Effects on users, operation and implications for the estimation of demand, Transportation Research Part A: Policy and Practice, Volume 53, July 2013, Pages 36–52

Trépanier, M., C. Morency, and B. Agard (2009). Calculation of transit performance measures using smartcard data. Journal of Public Transportation, 12(1), 79-97.

Welding, P.I. The instability of a close interval service. Operational Research Quarterly, Vol. 8, No.3, 1957, pp.133-148.

Wu, Y. H., & Miller, H. J. (2001). Computational tools for measuring space-time accessibility within dynamic flow transportation networks. Journal of Transportation and Statistics, 4(2/3), 1-14.