Australasian Transport Research Forum 2016 Proceedings
16 – 18 November 2016, Melbourne, Australia
Publication website: http://www.atrf.info

How close the models are to the reality? Comparison of Transit Origin-Destination Estimates with Automatic Fare Collection Data

Ahmad Tavassoli[1], Azalden Alsger[1], Mark Hickman[1], Mahmoud Mesbah[1]

[1]School of Civil Engineering, The University of Queensland, Brisbane, Australia

Email for correspondence: a.tavassoli@uq.edu.au

Abstract

There is a consensus on the importance and value of automatic fare collection (AFC) data in analysing different aspects of public transport. As such combining other data sources such as the General Transit Feed Specification (GTFS) can greatly improve the quality of the analyses and ultimately provide a better understanding of public transport performance. This paper presents a methodology for data processing and analysis to acquire a public transport Origin Destination (OD) matrix. The case study uses a very large dataset on passenger boarding and alighting of all three transit modes, namely bus, rail and ferry, in South-East Queensland (SEQ). The OD trip matrices are estimated for both the AM and PM peak periods for five weekdays. Also, the estimated public transport demands for the same periods from the SEQ strategic transport model (SEQSTM) are employed. This approach enables not only the comparison of OD matrices over time to determine changes in travel patterns but also investigates the similarity between the demands from the SEQSTM procedure and those from AFC data. A number of statistical measures, namely R^2, GEH, %RMSE and an eigenvalue-based measure, are utilized to determine the level of similarity of these OD matrices. The results highlight the similarity of the trip pattern between four workdays (Mon-Thu). However, trip patterns on a Friday are slightly different from the other weekdays, particularly in the PM peak period. Also, the demand from SEQSTM for both time periods is not analogous to any of the AFC patterns.

Key words: Smart card scheme; Public transport, OD matrices, OD matrix similarity

1. Introduction

Smart card data are increasingly used for transit network planning, passengers' behaviour analysis and network demand forecasting. The primary advantage of using smart cards, in addition to their original use as a valuable payment option, is to provide a high quality and plentiful source of information for transit agencies and researchers (Pelletier et al., 2011). In addition, smart card data can be used to better understand passenger travel behaviour and measure trip habits (Lee and Hickman, 2011; El Mahrsi et al., 2014), improve strategic planning and manage the demand through the network (Frumin, 2010; Sun et al., 2012) and estimate missing information such as alighting locations and OD trips (Gordon et al., 2013; Alsger et al., 2015). The accuracy level of smart card data greatly influences the extracted information and the successful estimation of OD matrices (Pelletier et al., 2011).

An OD matrix is an important input to transport models to assess new transport policies. There have been a few attempts to evaluate and assess the accuracy and similarity of OD matrices, using a number of statistical measures. Ye et al. (2012) used the Chi-squared test as a goodness-of-fit measure to compare synthetic matrices. The Chi-squared test ignores the correlations between cells and deals with cells independently. Alsger et al. (2015) used

the Geoffrey E. Havers (GEH) statistic to evaluate the accuracy level of a set of estimated matrices with a base OD matrix.

Djukic et al. (2013) presented the Mean Structural SIMilarity (MSSIM) as a more appropriate comparator of matrices. This method compares OD matrices as images based on pixels equating to individual OD cells. The authors showed a degree of correlation between the neighbouring cells (pixels) just as in images. Later, Day-Pollard and van Vuren (2015) investigated the comparison of OD matrices based on the MSSIM and other comparison techniques, namely R^2, GEH and RMSE. The authors concluded that the MSSIM approach requires further refinement for use with OD matrices. Ruiz de Villa et al. (2014) introduced a measure for comparing OD matrices, Wasserstein metric, unlike the methods that only based on the cell by cell comparison. The suggested method is based on the topology of the network and considering travel time between all OD pairs. However, this approach is impracticable for large networks due to the huge number of calculations required.

The accessibility and quality of data required for evaluation of estimated OD matrices have usually been a big challenge. In the current research, a unique smart card (automated fare collection, or AFC) dataset, known as GoCard and obtained from TransLink[1], is used to evaluate the estimated OD matrices. The important advantage of this dataset is that it includes both boarding and alighting times and locations for each passenger of the public transport services that comprise buses, trains and ferries.

In addition to the experimental data, this paper compares the results with a synthetic regional transport model. The South-East Queensland Strategic Transport Model (SEQSTM) is a four-step strategic transport model developed by the Queensland Department of Transport and Main Roads (TMR). This model is developed in the EMME/4 modelling platform[2] to serve as a long-range planning tool. The model was already calibrated and validated by TMR (TMR, 2011). The model is comprised of 1394 traffic zones, and this zoning system is used for estimating OD flows based on the AFC data. This model takes advantage of a mode choice model that includes seven modes: car driver, car passenger, walk to public transport, park and ride, kiss and ride, cycle and walk. The transit type includes three main modes: bus, rail, and ferry. The model forecasts demand for a 24-hour period and applies fixed time-period factors to allocate trips to the AM peak, inter-peak, PM peak and off-peak. Demand is segmented by eight resident trip purposes at trip generation, trip distribution and mode choice stages (Hunkin, 2009). The transit demand was obtained by aggregating all demands from different trip purposes after the mode choice step. The results of mode choice were calibrated and validated using the SEQ travel survey on the base year (2011) by TMR (Joycey and Ryan, 2008).

The objective of this paper is threefold:

- to investigate on the level of accuracy of the AFC data;
- to evaluate the travel pattern changes over time by comparing OD matrices based on AFC data on multiple days and in different periods of time; and,
- to compare the travel pattern obtained from the two sources of AFC and SEQSTM in both the AM and PM peaks.

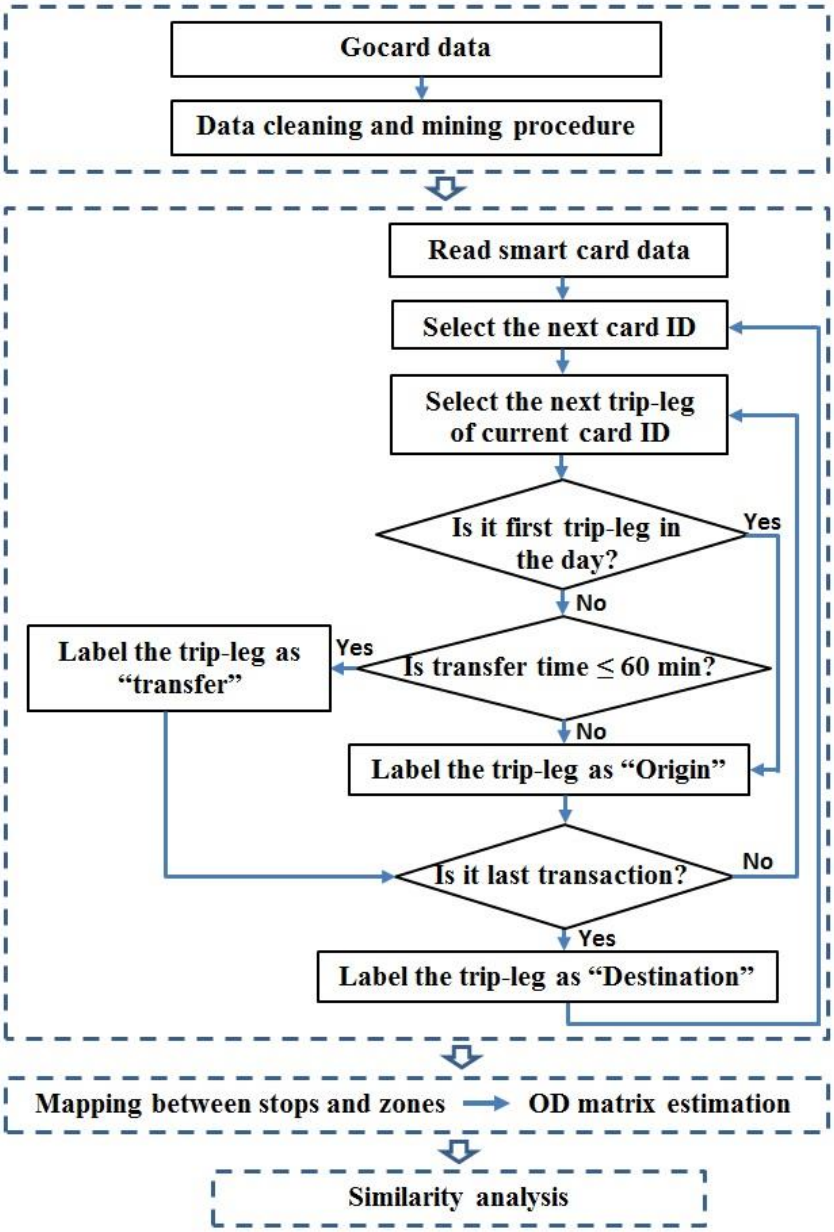
The remaining sections are organised as follows. The next section explains the research methodology, comprising the data description, the preparation and cleaning procedure, the trip-chaining method, and the OD estimation algorithm. The results of the OD estimation matrices for different weekdays are provided in the third section. These results are then used to conduct the similarity analysis and evaluate the accuracy of these matrices for different days and also for the SEQSTM using different measurements. Finally, conclusions and suggestions for future work are presented.

[1] The public transport authority of South-East Queensland (SEQ), Australia
[2] A commercial software package that is distributed by INRO in Canada

2. Methodology

Our study framework is shown in Figure 1. This framework encompasses four stages, namely: 1) AFC data processing, 2) application of a trip-chaining method, 3) OD matrix estimation, and 4) similarity analysis.

Figure 1: Study framework



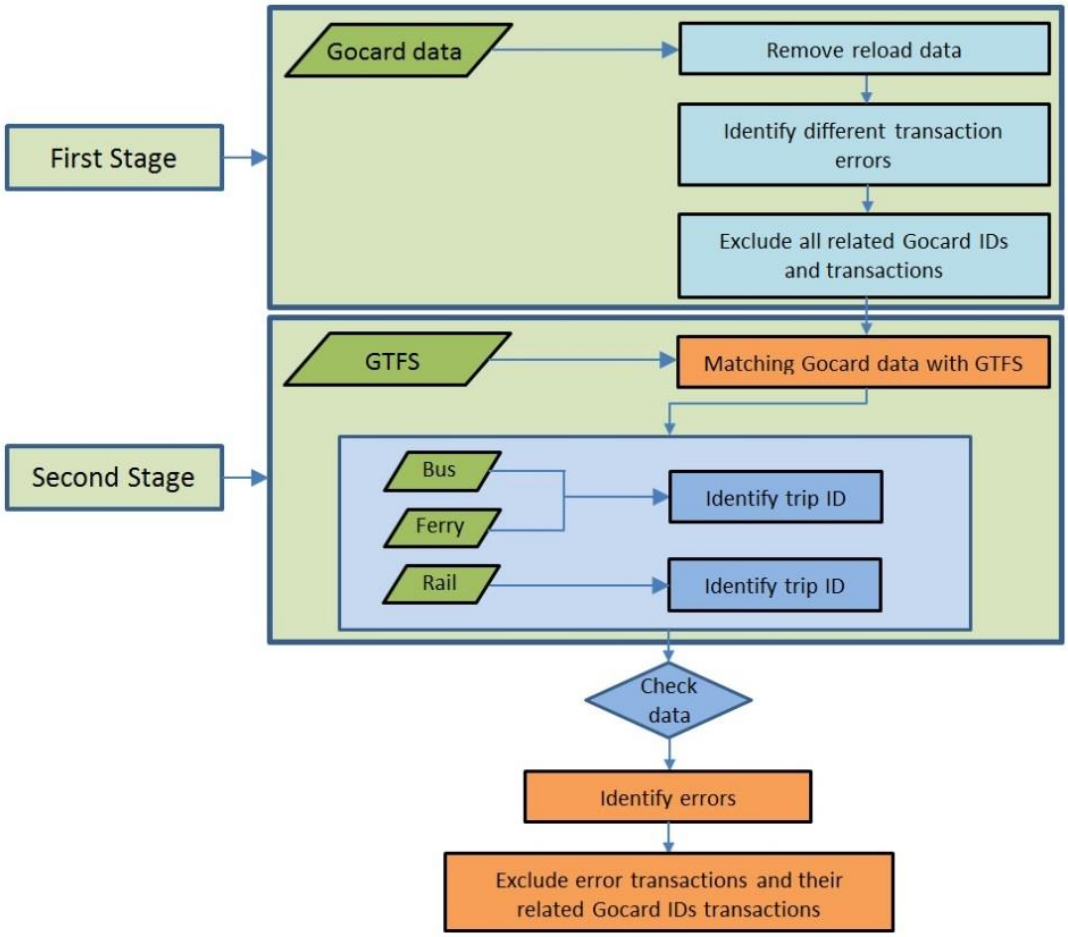
2.1. AFC data

Stage one relates to the GoCard dataset and the process of data cleaning. Robinson et al. (2014) highlight that the level of accuracy of AFC data may vary, and the data are affected by various types of errors. These errors may affect the accuracy of individual journeys and trip chains. In this regard, data validation and journey validation as described in Pelletier et al. (2011) are essential to ensure the quality of data for the purpose of this research. To perform the data cleaning, a framework is proposed including two stages as shown in Figure 2. Since the nature of data usually contains some errors caused by system failure or human error, the data are filtered with some transactions excluded, such as duplicate transactions and transactions when no boarding or alighting stops are recorded. In addition, all reload

transactions, which are related to GoCard credit top-ups and are not transactions related directly to public transit use, are excluded from the data.

The second stage involves mapping the GoCard data into the GTFS network. This stage facilitates our investigation of the validity of boarding and alighting stops and the associated route and direction. The approach includes finding all possible trip-IDs for each trip leg based on the boarding and alighting stops in conjunction with their associated times. Then, the best trip-ID is selected based on minimising the differences between the scheduled and actual transaction times calculated for both boarding and alighting stops. Ultimately, transactions with errors in defining origin or destination stops and/or times are identified and excluded from the analyses.

Figure 2: Framework of GoCard data mining and cleaning process



If any transaction of a card ID is excluded, the rest of the transactions for that card ID are also excluded for the given day, as the transactions on a single day must be in sequence to be chained into a tour for the given day.

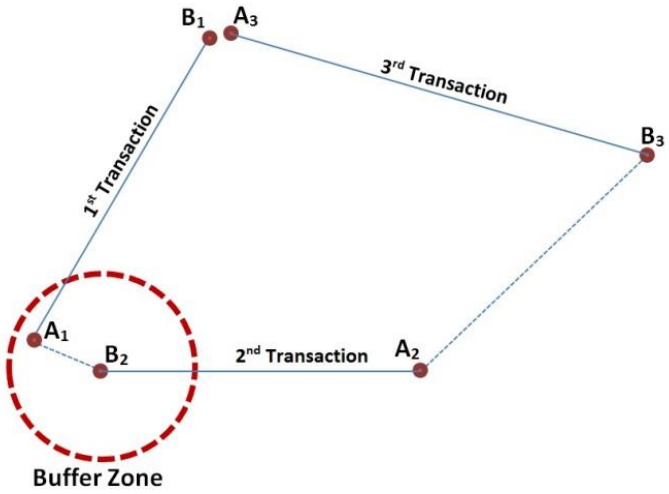
2.2. Trip-chaining method

The main purpose in the trip-chaining method is to connect transactions of a passenger to infer full passenger journeys. Figure 3 shows an example of the trip-chaining method.

A passenger has his first transaction from first boarding (B_1) to first alighting (A_1), and then walks to the next boarding stop to start the second transaction from second boarding (B_2) to second alighting (A_2). To complete a passenger's travel sequence, allowable transfer time has to be assumed. This time threshold is used to merge transactions into a single journey. If the transfer time exceeds this threshold, the next boarding is a new OD trip (Munizaga et

al., 2014; Alsger et al., 2015). It should be noted that the time between (A_1) and (B_2) could be short, with enough time only for a transfer, or long, with enough time for an intermediate activity.

Figure 3: Example of the trip-chaining method



Note: The first boarding transaction in a day is identified by B_1 and the first alighting transaction as A_1, the time between B_1 and A_1 is the in-vehicle time. (Adapted from: Alsger et al., 2015)

The main function of this method is to detect transfers, so that trip-legs can be merged to obtain journeys. In this regard, the first transaction of the day is the boarding for the first trip-leg, for a unique card ID. The remaining trip-legs for the same card ID will be transfers if the transfer time is less than the allowable transfer time (threshold). The allowable transfer time is set as 60 minutes, to be compatible with TransLink's 60-min transfer time allowance (Alsger et al., 2015). If the transfer time exceeds this value, the alighting location of the prior trip-leg is the destination of the passenger's journey, and the next transaction is the boarding location of a new journey. If an alighting transaction is the last transaction of the day for the current card ID, another card ID is chosen, and the algorithm continues searching to create passenger journeys.

2.3. OD matrix estimation

The next step is to estimate the OD matrix from the passengers' journeys based on the GoCard dataset. In this process, stop-to-stop OD journeys should be converted to zone-to-zone OD journeys. For this purpose, the same traffic analysis zones (TAZs) in the SEQSTM are utilized as the level of aggregation. A journey from any stop located within the TAZ is counted as a journey originating in that TAZ; similarly, a journey ending at any stop within a TAZ is counted as a journey destined for that TAZ. Using the same procedure, OD matrices can be estimated for different days of a week and also different time periods in each day (AM peak and PM peak).

2.4. Similarity analysis

Different measurements are utilized to compare OD matrices and determine the level of similarity over time and between sources. These measures are described below.

2.4.1. R-squared (R^2)

The R-squared (R^2), as one of the most commonly and widely used (Washington et al., 2011), is a statistical measure of how close the data are to the fitted regression line, and it is used for comparing between origin-destination pairs of two OD matrices. R^2 values range

from 0 to 1, with higher values indicating less difference between OD matrices. A general formula for calculating R^2 is:

$$R^2 = 1 - \frac{SS_E}{SS_T} \quad (1)$$

$$SS_E = \sum_i (OD^1_{i,j} - OD^2_{i,j})^2 \quad (2)$$

$$SS_T = \sum_i (OD^1_{i,j} - \overline{OD})^2 \quad (3)$$

where: $OD^1_{i,j}$ is the pair *i,j* of the first demand matrix and $OD^2_{i,j}$ is the pair *i,j* of the second demand matrix $\overline{OD}$ is the mean of the OD^1 pairs.

Along with considering higher value of R^2 as a higher level of similarity, the regression line should be close to a 45-degree line through the origin. In this condition, the coefficient of the line should be closer to one and the intercept should be closer to zero. The lower and greater coefficient values indicate the tendency of the pattern to overestimate or underestimate values in the reference OD matrix.

2.4.2. Geoffrey E. Havers (GEH) statistic

The GEH statistic is used to evaluate the level of closeness between origin-destination pairs of two OD matrices. The GEH is applied to every pair in the two matrices, with a GEH of less than 5 indicating a good fit (Hollander and Liu 2008). The GEH formula is:

$$GEH = \sqrt{\frac{2(OD^1_{i,j} - OD^2_{i,j})^2}{OD^1_{i,j} + OD^2_{i,j}}} \quad (4)$$

Then, the percentage of OD pairs that have a GEH equal to or less than 5 is calculated to indicate the level of closeness between two OD matrices.

2.4.3. Root Mean Square Error (RMSE) and Percent Root Mean Square Error (%RMSE)

The root mean squared error (RMSE) and accordingly the percent root mean squared error (%RMSE) are used to evaluate the closeness of the matrices. The %RMSE is where the variability of the demand is most evident: if two demand matrices were identical, the %RMSE would be equal to zero. The (RMSE) and (%RMSE) are:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (OD^1_{i,j} - OD^2_{i,j})^2}{N}} \quad (5)$$

$$\%RMSE = \frac{RMSE}{\left(\frac{\sum_{i=1}^{N} OD^1_{i,j}}{N} \right)} \times 100 \quad (6)$$

where: $OD^1_{i,j}$ is the pair *i,j* of the first demand matrix and $OD^2_{i,j}$ is the pair *i,j* of the second demand matrix.

2.4.4. Eigenvalue-based measure (EBM)

This measure is based on the concept of eigenvectors and is introduced for comparison of OD matrices in this study. OD_1 *is* similar to OD_2 if there exists a matrix P such that $OD_2 = P^{-1}OD_1P$.

$$\begin{aligned}\det(OD_2 - \lambda I) &= \det(P^{-1}OD_1P - \lambda P^{-1}P) = \det(P^{-1}[\mathrm{OD}_1 - \lambda\,\mathrm{I}]P)\\ &= \det(P^{-1})\det(\mathrm{OD}_1 - \lambda\,\mathrm{I})\det(P)\\ &= \det(P^{-1})\det(P)\det(\mathrm{OD}_1 - \lambda\,\mathrm{I}) = \det(P^{-1}P)\det(\mathrm{OD}_1 - \lambda\,\mathrm{I})\\ &= \det(\mathrm{OD}_1 - \lambda\,\mathrm{I})\end{aligned} \quad (7)$$

OD_1 and OD_2 have the same characteristic polynomial and therefore the same eigenvalues. On this basis, two similar matrices have the same eigenvalues (Ford, 2014). Based on this approach, two OD matrices are similar if their eigenvalues are the same. The sum of absolute error (SAE) is a promising technique to determine the closeness of model results to the actual data and has been used in a number of studies (Chowdhury and Saha, 2011; Purdy, 2012). This technique is employed in this study to establish a measure to show the similarity of ODs by comparing vectors of the eigenvalues of the OD matrices; the lower the value, the greater is the similarity.

$$EBM = SAE(\mathrm{eig}(\mathrm{OD}_1) - \mathrm{eig}(\mathrm{OD}_2)) \quad (8)$$

where: OD_1 and OD_2 are the demand matrices, and eig(.) is a vector containing the eigenvalues of a square matrix.

3. Data description and analysis

The GoCard dataset for five weekdays is analysed over the SEQ network, considering all modes, namely bus, rail, and ferry, for Monday 18 March to Friday 22 March 2013. These weekdays were selected as there was no public holiday and had normal weather conditions. In the SEQ network, a transaction record is generated each time a passenger boards and alights. Each transaction contains information comprising: the operation date, run, route, direction, ticket number, smartcard ID, boarding time, alighting time, boarding stop and alighting stop. However, transferring activities are not directly obtained.

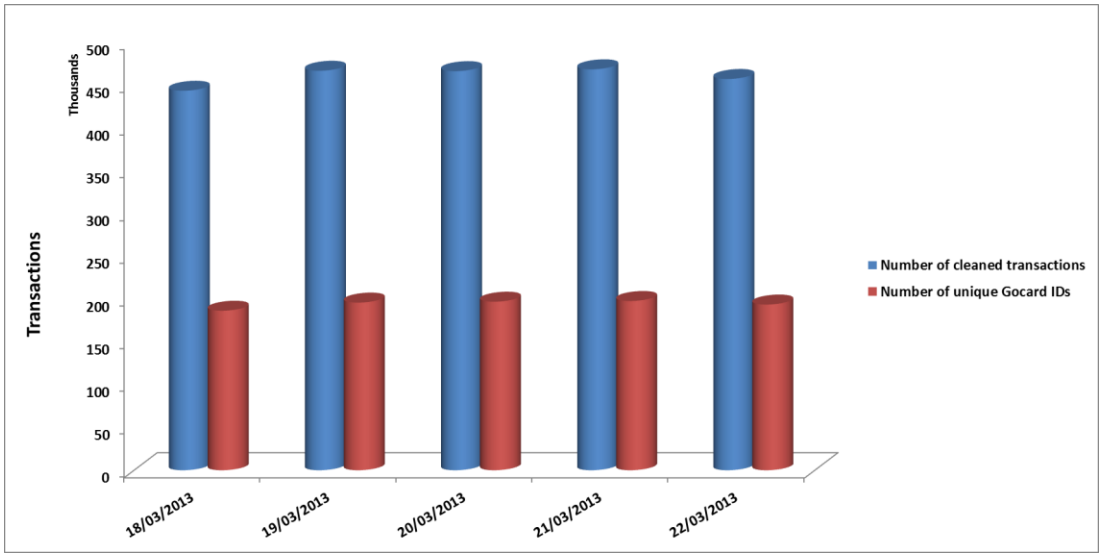
Table 1 presents the results of the data in different stages of this study as described in Section 2. First, the general data description is given, including the total number of transactions per day before cleaning, the reload transactions, and the available trip legs per day after excluding these errors. Then, a summary of these errors is given for each day: the situations that boarding or alighting locations are missing or are not recorded, boarding or alighting times are missing, the boarding stop location is the same as the alighting stop, and boarding time is later than or equal to the alighting time. The next step in data cleaning is identifying and excluding stop mapping errors where the algorithm could not find a match between the stops in both the GoCard data and GTFS.

The next part of Table 1 presents the number of trip legs after performing the data cleaning, the number of errors in trip legs in the second stage, the number of treated trip legs, and ultimately the number of available trip legs after all data cleaning. The results indicate that about 68% of the initial data can be used for the current research. This available GoCard data is large enough to provide a reliable sample size, considering the fact that more than 82% of transit trips are made by GoCard users in SEQ (Moore, 2015). After applying the trip-chaining method to obtain passengers' journeys, the last section of the table shows the information related to the number of journeys per day and also the unique number of GoCard IDs related to these journeys. Figure 4 shows the total number of cleaned transactions and the corresponding number of GoCard IDs for the selected weekdays.

Table 1: Data description and statistics in different stages of the study

Description	Date				
	18/03/13 Mon	19/03/13 Tue	20/03/13 Wed	21/03/13 Thu	22/03/13 Fri
General data description					
Total transactions	609,509	635,841	628,479	633,106	613,481
Number of reload transactions	16,382	17,034	16,999	17,621	15,921
Available trip legs per day	593,127	618,807	611,480	615,485	597,560
General errors					
No boarding stop	18,005	18,683	16,353	16,916	14,915
No alighting stop	11,931	11,698	11,780	11,430	10,504
No boarding time or alighting time	29,936	30,381	28,133	28,346	25,419
Boarding time >= alighting time	2,776	2,846	3,025	2,836	3,090
No boarding stop or alighting stop	29,936	30,381	28,133	28,346	25,419
Boarding stop = alighting stop	12,017	12,242	12,035	11,852	12,596
Stop mapping errors					
Stop mapping error in boarding	12,223	12,684	11,989	12,231	11,559
Stop mapping error in alighting	11,988	12,297	11,680	11,939	11,296
Trip legs					
Number of available trip legs after first data cleaning	443,087	466,438	465,988	468,119	456,842
Number of errors in trip legs	45,836	46,651	44,973	46,601	44,395
Number of treated trip legs	6,306	6,635	6,324	7,109	7,813
Number of available trip legs after all data cleaning	395,633	418,823	418,098	418,424	409,296
Ratio of available data to all data	66.7%	67.7%	68.4%	68.0%	68.5%
OD trips information					
Number of Go Card IDs	182,799	192,501	193,433	194,151	190,187
Number of journeys	324,091	343,000	342,821	373,796	334,287

Figure 4: Number of cleaned transactions and GoCard IDs for the selected weekdays



OD matrices for weekdays based on the passengers' journeys are then generated. The next section introduces the results of the estimation of the OD matrices for both weekdays and for the SEQSTM.

3.1. OD matrices estimation

Transit travel demand mostly follows a non-uniform time-of-day distribution and includes two main peak periods, AM peak and PM peak. To understand the daily behaviour of demand, journeys were aggregated based on the start time during each 15-min interval. Figure 5 shows the public transport time-of-day demand based on the Go Card data for weekdays.

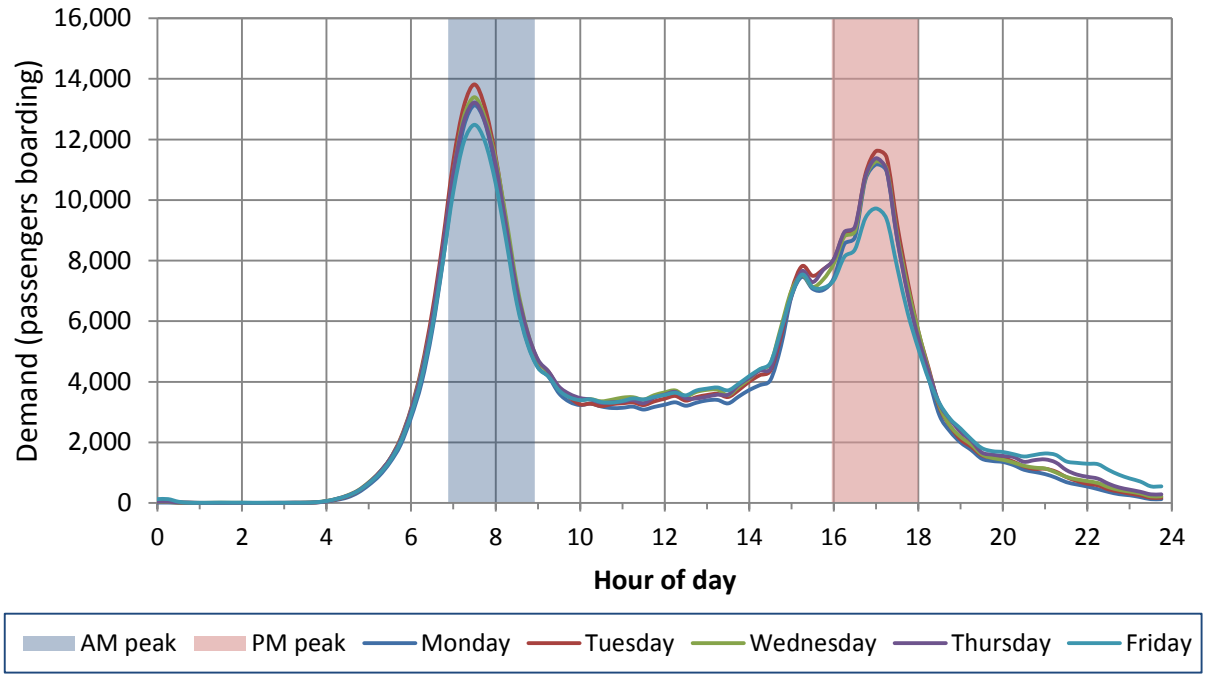
Since demand during each time unit is averaged over 15-min intervals, this could cause variation of demand. As a result, there is a potential risk of overestimating or underestimating the demand profile. The moving average technique is employed to minimize the effect of this variation and to smooth the demand profile. The moving average of three sequential demands can be calculated as:

$$MDS_t = (DS_{t-1} + DS_t + DS_{t+1}) / 3 \quad (9)$$

where

DS_t = the moving average of three sequential speeds;
DS_{t-1} = the demand at one time-interval before t;
$DS_{,t}$ = the demand at one time-interval at t;
DS_{t+1} = the demand at one time-interval after t;

Figure 1: Illustrative Time-of-Day Variations in Transit Demand for weekdays



As can be seen from Figure 5, all weekdays follow a similar trend including morning and afternoon peak periods. The afternoon peak is lower and more broadly spread out compared with the morning peak. In addition, the demand on Friday is slightly different from that of

other weekdays and lower in both peak periods. To use similar peak periods as those in the SEQSTM, AM peak and PM peak are defined as 7:00AM-9:00AM and 4:00PM-6:00PM, respectively. Based on the start time of the journeys, OD matrices were calculated for weekdays and both AM and PM peak periods. In addition, public transport OD matrices from SEQSTM for the two peak periods are extracted. A summary of statistics of all OD matrices are shown in Tables 2 and 3.

Table 2: Summary statistics of demand matrices for AM peak

	Mon	**Tue**	**Wed**	**Thu**	**Fri**	**SEQSTM**
Total demand (journeys)	80,259	84,020	83,305	82,246	77,333	210,976
Number of pairs with non-zero demand	15,922	16,507	16,469	16,496	15,978	1,865,955
Maximum demand (journeys)	587	548	560	588	513	790
Average of demand (journeys)	5.04	5.09	5.06	4.99	4.84	0.11
STD of demand (journeys)	18.40	18.34	18.33	18.13	17.37	2.40

Table 3: Summary statistics of demand matrices for PM peak

	Mon	**Tue**	**Wed**	**Thu**	**Fri**	**SEQSTM**
Total demand (journeys)	73,733	76,872	75,099	75,323	66,409	139,476
Number of pairs with non-zero demand	14,560	15,100	14,957	15,047	14,440	1,865,953
Maximum demand (journeys)	562	549	561	580	437	203
Average of demand (journeys)	5.06	5.09	5.02	5.01	4.60	0.07
STD of demand (journeys)	18.31	18.22	17.84	17.92	15.44	1.00

The results indicate that the demand based on the GoCard data in the PM peak always is lower than that during the AM peak. Friday has the lowest demand and Tuesday has the highest demand among weekdays. On average, the number of OD pairs with demands more than zero are about 1000 pairs more in the AM peak. Nonetheless, the maximum demand, average demand, and STD of demand are very similar, comparing both peak periods.

Comparing the demands from GoCard data and SEQSTM in both peak periods reveals that the SEQSTM demand is significantly higher than the GoCard weekday demand in both the AM and PM peaks. This might be due to the fact that during the data cleaning about 32% of the data was excluded from the analysis. In addition, there were some passengers who used paper tickets rather than the GoCard and therefore were not considered in the analysis. On this basis, all demand matrices were normalised according to their total demand to analyse the similarity between matrices. Furthermore, the fairly low average demand and STD of demand, and the high number of number of OD pairs with demand greater than zero in the SEQSTM, indicate that there are many OD pairs with demand less than 1 journey, particularly for the PM peak. This may cause big discrepancies between the GoCard weekday demand and the SEQSTM demand. In this regard, analysis was performed using different threshold measures including 0.5, 1, 2 and 3 trips, in order to identify the impacts of demand pairs with a low value in the results. The results show that “Number of pairs with non-zero demand” decreased to 51,360, 28,653, 15,009 and 9,918 pairs in the AM peak, respectively, for the four thresholds. In PM peak, this measure decreased to 43,337, 22,310, 10,629 and 6,719 pairs, respectively, for the four thresholds.

4. Similarity analysis

The statistical measures from Section 2.4 are utilised to assess the similarity between demand matrices of the GoCard and the SEQSTM in AM and PM peak periods. The results of the analysis based on these measures are presented in Table 4 and Tables 5 for the AM and PM peak periods, respectively. The GEH measures for almost all comparisons are more than 99%, indicating a fairly good similarity, unlike the results from the other measures. This might be related to the scale of the demand pairs. On this basis, the GEH statistic was determined not to be a suitable measure for this analysis and was excluded from the considered measures for comparison. It is only indicative when a threshold is defined for the 'negligible' OD demand; for an example of this approach, see Alsger et al. (2015). As discussed in the previous section, different threshold measures were used to analyse the demand from the SEQSTM. The results indicate that in all conditions in both the AM peak and the PM peak periods, transit travel patterns from the SEQSTM are not similar to the weekday demand from the GoCard. Accordingly, the results in this section are based on the not excluding any demand from the SEQSTM.

The demand matrices are almost the same in the AM peak on weekdays, as the R^2 measure is more than 0.97 and the coefficient of the lines are quite close to one (more than 0.910) and the constants are close to zero (less than 0.27), as shown in Tables 4a and 4b. In addition, the %RMSE measures are within the same range and the EBM measures follow a similar pattern. However, the OD matrix on Friday is slightly different from the other weekdays, as the R^2 is lower on Friday, the coefficients and constants have more distance from the ideal measures on Friday, and also %RMSE and EBM measures are higher on Friday. This suggests different timing and scale of demand in the Friday PM peak period, as some people may leave work early or may prefer to do a social activity before going home.

Comparing normalised demand matrices between the GoCard and the SEQSTM indicates that transit travel patterns from the SEQSTM are not similar to the weekday demand from the GoCard in both AM and PM peaks. R^2 measures are clearly lower compared to the same measure between weekdays, with a fairly large distance of the coefficients and constants from the ideal measures. On average the coefficient is about 0.2 away from one and the constant is about 1.15.

The EBM measures are also about 250,000 on average for similarity between the SEQSTM matrix compared to an average of 65,000 for weekdays in the AM peak period as shown in Table 4. The same trends can be seen in the PM peak period. The %RMSE measure is also confirming the dissimilarity of the SEQSTM matrix with the weekday matrices based on the GoCard data. This suggests the need to re-evaluate and calibrate the demand within the strategic transport model with the demand from actual comprehensive datasets.

Table 4: Results of similarity measures between demand matrices for AM peak *

a) R^2 measure

	Mon	Tue	Wed	Thu	Fri	STM
Mon	1	0.98	0.97	0.97	0.97	0.21
Tue		1	0.98	0.97	0.97	0.19
Wed			1	0.98	0.98	0.20
Thu				1	0.98	0.20
Fri					1	0.21
STM						1

b) Parameters for the best fitting line **

	Mon	Tue	Wed	Thu	Fri	STM
Mon	1 0	1 0.27	0.99 0.23	0.99 0.2	0.93 0.21	0.20 1.60
Tue		1 0	0.98 0.04	0.97 -0.002	0.91 0.03	0.19 1.57
Wed			1 0	0.977 0.04	0.92 0.06	0.20 1.56
Thu				1 0	0.93 0.08	0.20 1.58
Fri					1 0	0.21 1.58
STM						1 0

c) %RMSE

	Mon	Tue	Wed	Thu	Fri	STM
Mon	0	49	51	52	53	351
Tue		0	49	50	55	346
Wed			0	47	49	348
Thu				0	47	349
Fri					0	345
STM						0

d) Eigenvalue-based measure

	Mon	Tue	Wed	Thu	Fri	STM
Mon	0	63,793	60,072	63,984	67,975	247,866
Tue		0	70,292	68,605	70,135	249,276
Wed			0	60,181	67,891	248,125
Thu				0	68,802	249,120
Fri					0	248,580
STM						0

* Demand of weekdays based on Go card data, STM: public transport demand from SEQSTM
** The top value is coefficient of the line, the below value is intercept of the line

Table 5: Results of similarity measures between demand matrices for PM peak *

a) R^2

	Mon	Tue	Wed	Thu	Fri	STM
Mon	1	0.98	0.96	0.97	0.94	0.3
Tue		1	0.96	0.97	0.95	0.29
Wed			1	0.96	0.94	0.29
Thu				1	0.95	0.3
Fri					1	0.28
STM						1

b) Parameters for the best fitting line **

	Mon	Tue	Wed	Thu	Fri	STM
Mon	1 0	0.99 0.24	0.96 0.33	0.97 0.25	0.81 0.58	0.21 1.14
Tue		1 0	0.95 0.17	0.97 0.09	0.80 0.43	0.21 1.12
Wed			1 0	0.98 0.12	0.93 0.42	0.22 1.11
Thu				1 0	0.94 0.45	0.22 1.11
Fri					1 0	0.25 1.05
STM						1 0

c) %RMSE

	Mon	Tue	Wed	Thu	Fri	STM
Mon	0	49	61	57	72	335
Tue		0	58	54	67	332
Wed			0	64	72	329
Thu				0	65	331
Fri					0	312
STM						0

d) Eigenvalue-based measure

	Mon	Tue	Wed	Thu	Fri	STM
Mon	0	89,445	62,214	86,293	101,385	138,536
Tue		0	91,617	91,788	100,427	138,665
Wed			0	86,210	99,204	140,525
Thu				0	75,658	141,131
Fri					0	148,862
STM						0

* Demand of weekdays based on Go card data, STM: public transport demand from SEQSTM
** The top value is coefficient of the line, the below value is intercept of the line

5. Conclusions

Automated fare collection systems have been widely used in public transport and have provided very large datasets. One of the main challenges of using such data is its accuracy level. This study makes use of the AFC system in SEQ, Australia, which has extensive records of passenger boardings and alightings by all transit modes. On this basis, this study presents the errors at different stages of data cleaning and presents the quality of such data for journey estimation.

From another perspective, AFC systems are a rich source of information for many transport planning applications. The proposed methodology in this study utilises the AFC data to characterise passenger journeys in order to estimate the OD matrices of transit passengers in weekdays during both AM peak and PM peak periods. In addition, a traditional four-step model, the SEQSTM, is used to compare transit OD demand in both the AM peak and PM peak periods.

Comparing demand matrices provides essential information about passenger travel patterns. This comparison may also help to avoid unnecessary surveys by suggesting the use of similar available data. This study introduces a new measure for OD matrix similarity, an eigenvalue-based measure (EBM), along with the other established statistical measures of R^2, GEH, and %RMSE. The results show that the proposed measure has good performance in terms of indicating the level of similarity between matrices. This study performs the similarity analysis between weekdays demand matrices based on GoCard data and also with the SEQSTM demand in both peak periods.

The results show that the AM peak has slightly higher demand compared to PM peak for all weekdays, and the demand fluctuations are greater across days in the PM peak. Also, the Friday demand is slightly different from other weekdays (Monday to Thursday) in the PM peak. Furthermore, the SEQSTM has larger demand compared to the GoCard weekday demand. These findings highlight that the public transport travel OD matrix according to the SEQSTM is distinct from that of the GoCard data.

Further research needs to be conducted to investigate on the similarity of the transit demand on weekends. Also, it is recommended that the effects of adverse weather on transit demand and passengers' travel behaviour be examined.

Acknowledgements

The authors acknowledge TransLink (a unit of TMR covering all of Queensland, Australia) for providing the data for this research. We also acknowledge the Department of Transport and Main Roads (TMR), Transport Strategy and Planning Branch for providing access to the South-East Queensland Strategic Transport Model (SEQSTM). The research in this paper is partially supported by the Australian Research Council through a Discovery Early Career Researcher Award (grant number DE130100205) and also by the ASTRA (Academic Strategic Transport Research Alliance) Chair at the University of Queensland.

References:

Alsger, A. A., Mesbah, M., Ferreira, L. & Safi, H. (2015) Use of Smart Card Fare Data to Estimate Public Transport Origin–Destination Matrix. *Transportation Research Record: Journal of the Transportation Research Board,* 2535**,** 88-96.

Chowdhury, S. & Saha, P. (2011) Adsorption Kinetic Modeling of Safranin onto Rice Husk Biomatrix Using Pseudo-first-and Pseudo-second-order Kinetic Models: Comparison of Linear and Non-linear Methods. *CLEAN–Soil, Air, Water,* 39(3)**,** 274-282.

Day-Pollard, T. & Van Vuren, T. (2015) When are Origin-Destination Matrices Similar Enough? *Transportation Research Board 94th Annual Meeting.*

Djukic, T., Hoogendoorn, S. & Van Lint, H. (2013) Reliability Assessment of Dynamic OD Estimation Methods Based on Structural Similarity Index. *Transportation Research Board 92nd Annual Meeting.*

El Mahrsi, M. K., Etienne, C., Johanna, B. & Oukhellou, L. (2014) Understanding passenger patterns in public transit through smart card and socioeconomic data: A case study in rennes, france. *ACM SIGKDD Workshop on Urban Computing.*

Ford, W. (2014) *Numerical linear algebra with applications: Using MATLAB*, Academic Press.

Frumin, M. S. (2010) Automatic data for applied railway management: passenger demand, service quality measurement, and tactical planning on the London Overground Network, Massachusetts Institute of Technology.

Gordon, J., Koutsopoulos, H., Wilson, N. & Attanucci, J. (2013) Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. *Transportation Research Record: Journal of the Transportation Research Board,* (2343)**,** 17-24.

Hunkin, P. (2009) *Critical review of transport modelling tools* Sinclair Knight Merz (SKM), viewed 28/09/2016, <https://bitre.gov.au/publications/2009/files/cr_001_Review_of_Transport_Modelling_Tools.pdf>.

Joycey, A. & Ryan, M. (2008) *BSTM Multi-Modal Model Development, Base Year Validation, Working Paper 16*, Queensland government- Queensland transport main roads, Queensland, Australia.

Lee, S. & Hickman, M. D. (2011) Travel pattern analysis using smart card data of regular users. *Transportation Research Board 90th Annual Meeting.*

Munizaga, M., Devillaine, F., Navarrete, C. & Silva, D. (2014) Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies,* 44**,** 70-79.

Pelletier, M.-P., Trépanier, M. & Morency, C. (2011) Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies,* 19(4)**,** 557-568.

Purdy, J. E. (2012) A Theoretical Development and Simulation-Based Comparison of Four Parameter Estimation Methods for the Spatio-Temporal Autologistic Model with Emphasis on Maximum Generalized and Block Generalized Pseudolikelihood, PhD thesis, The University of Montana, Missoula, MT.

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.

Ruiz De Villa, A., Casas, J. & Breen, M. (2014) OD matrix structural similarity: Wasserstein metric. *Transportation Research Board 93rd Annual Meeting.*

Sun, L., Lee, D.-H., Erath, A. & Huang, X. (2012) Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system. *Proceedings of the ACM SIGKDD international workshop on urban computing.* ACM.

TMR (2011) South East Queensland Strategic Transport Model (SEQSTM). SEQSTM_MM_v1 ed. Brisbane, Australia, Transport Strategy And Planning Branch, Department of Transport and Main Roads.

Washington, S., Karlaftis, M. G. & Mannering, F. L. (2011) *Statistical and econometric methods for transportation data analysis,* Boca Raton, FL, CRC Press.

Ye, X., Cheng, W. & Jia, X. (2012) Synthetic Environment to Evaluate Alternative Trip Distribution Models. *Transportation Research Record: Journal of the Transportation Research Board,* (2302)**,** 111-120.