Transfer Journey Identification and Analyses from Electronic Fare Collection Data

Markus Hofmann, Margaret O'Mahony

Abstract—Understanding the behaviour of public transport passengers is key to providing a system from which passengers will derive the maximum benefit. One method of analysing this behaviour is with the use of passenger boarding data, stored in a database. Such a database may be improved by enriching the already existing dataset by applying specific algorithms. This paper describes an iterative classification algorithm that classifies passenger boardings into two categories; transfer journeys and single journeys. The dataset used was from an urban public transport operator with a large fleet (over 1000 buses) and data of 48 million magnetic strip card boardings from 1998 and 1999. This paper details the process involved in the initial development of the iterative classification algorithm, the analysis of transfer node identification matrices, waiting time information charts and spatial first/second boarding matrices. When the algorithm is applied to the dataset it provides transport planners with valuable information with regard to passenger boardings, transfers and waiting times which can assist them in transport planning and policymaking. The purpose of this paper is to describe the automatic generation of a new data attribute that cannot be derived directly and therefore increases the future utilization of the dataset. The paper presents various analyses based on the extended and enriched database to illustrate this point.

I. INTRODUCTION

Key to the success of a public transport system is not only to have the appropriate infrastructure in place, but also to maximise this infrastructure with all modes complementing each other. This can be achieved with a system of complementary scheduling whereby transport planners are aware of passenger movements within the network and can use this information for the scheduling of new or remodelled routes or services [1]. Connecting services provided for passengers between origin and destination should be within an acceptable time period in order to minimise waiting times and therefore attract and retain customers. In order for these connecting services to be provided, the transfer nodes of the transportation network and their properties must be known. Properties of transfer nodes could be passenger volume, variation of passenger volume throughout peak/off-peak time or modes serving the transfer node. Before it is possible to analyse the transfer nodes in a transportation network, it is necessary to have data on passenger transfer journeys. One method of analysing this behaviour is with the use of passenger boarding data, stored in a database. This paper introduced an iterative classification algorithm that classifies individual passenger boardings into two categories; single journey and transfer journey (also known as linked trips or transfer trips). According to the UK Department of the Environment, Transport and the Regions (DETR) a journey is a one-way course of travel having a single main purpose [2]. For the purpose of this report a transfer journey is a journey with one bus transfer. A single journey is a journey that does not require a transfer to reach the final destination. In Fig. 1, a transfer journey is defined as a journey that consists of two individual bus boardings at A1 and B1. A2 and B1 are defined as transfer nodes. A1 is the journey origin and B2 is the journey destination.

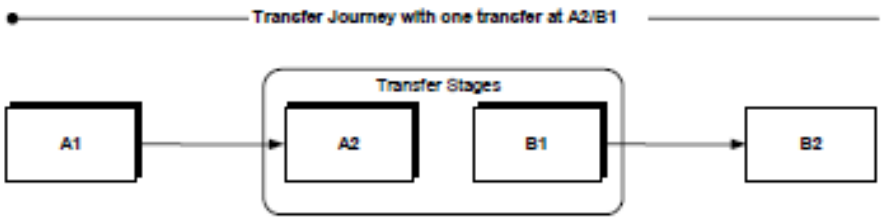

Fig. 1: Transfer journey definition diagram

One of the main aims of transfer journey analysis is to identify whether the bus services are meeting the needs of the passengers [3]. These needs can be met when the level of service (LOS) is acceptable. The LOS measures identified in the 'Transit Capacity and Quality of Service Manual' (TCQSM) [4] can be partially created by extracting information from an electronic fare collection (EFC) database. The section of the paper entitled Transfer Analysis describes some of the analyses that are possible after identifying all transfer journeys. These analyses may contribute to decision and policymaking.

This project is based on data gathered by EFC equipment, which is installed on each of the buses of the fleet. All boardings using a magnetic strip card were recorded to text files which were then stored in an Oracle database. There is no record of cash transactions in the database. The 160 million record database represents detailed information of each individual passenger boarding as well as additional spatial and time information [5].

This paper introduces an algorithm that identifies all such results can be used to improve the potential analyses of existing boarding records stored in a database. Various analyses show how the newly gained information can be used to improve decision making for operations management and policymaking. The analyses of this paper focus on both a macro level and micro level study. The generation of various transfer matrices for the purpose of transport modelling is briefly described.

II. BACKGROUND

When public transport operators use EFC data to analyse their operations or passenger behaviour, they can decrease the number of surveys they conduct and therefore decrease costs. Furthermore, it provides the analyst with a much larger sample than what could be provided using survey data. With the use of innovative technologies such as data mining or online analytical processing, new analysis methodologies and results can be generated. Bagchi et al. [3] investigated transfer journeys and single journeys for a small subset of smart card data collected in Bradford and Southport, UK. The study analysed bus-to-bus interchanges of elderly passengers and based their analysis on a sample of 98,000 journeys. The analysis produced various descriptive statistics grouped by ticket validity period, which contributed to a better understanding of transfer journeys of elderly people. This type of analysis allows the analyst to investigate travel patterns over a longer period of time. Bagchi et al. [3] suggest that sample surveys will play a complementary role for verifying the results obtained by smart card data or to provide data that is not possible to record with smart cards (e.g. reason for journey or mode choice).

Navick et al. [6] also base their analysis on EFC data from buses, which were collected in the Los Angeles area. Information where passengers board and alight is useful for effective planning and operation control. This information was used to estimate system wide passenger miles, passenger loads and origin/destination matrices [6]. EFC data generally does not show where the passenger alights. However, in the case of transfer journeys the location of A2 can be estimated (see Fig. 1) as it can be assumed that the passenger alighted very close to bus stop B2.

Many public transport service providers have to base their analysis of their operations on very small survey samples [7]. The TCQSM [4] introduced a series of service measures such as service frequency, hours of service, service coverage, passenger loads, reliability and transit time.

The project database is based on data gathered from an urban bus operator on its entire transportation network. For confidentiality reasons, the source of the database cannot be disclosed. The vast amount of transactional data from 1998 and 1999 (160 million records) has been moved from text files (one file per day) into a large relational database [5].

The iterative classification algorithm described in this paper enriched the database by an identifier which shows whether the individual passenger boarding is part of a transfer journey (two or more boardings within 90 minutes by the same passenger). This newly added data attribute facilitates further analysis of which some is presented in this paper. It is important to note that only journeys that have been carried out using a magnetic strip cards can be analysed. Cash transactions cannot be analysed.

The transfer journey identifier has been created for the months April '99, May '99, September '99 and October '99 as these months did not reflect any major abnormalities such as school breaks or summer holiday periods. All analysis is based on these four months unless stated differently. During this 4 month period over 7.9 million individual boardings were recorded. Over 1.4 million transfer journeys were identified by the classification algorithm proposed in this paper. Approximately 12.5% of all transactions (including cash boardings) were carried out using magnetic strip cards. Approximately 35% of all magnetic strip card boardings were part of a transfer journey.

III. DEFINITION OF THE ITERATIVE CLASSIFICATION ALGORITHM

The original project dataset includes data that has been recorded for each individual passenger boarding. In its original state it was not possible to identify which passenger boarding was part of a transfer journey (two boardings). In order to generate transfer passenger analysis, it has to be known which individual passenger boardings were linked together. The solution was an iterative classification algorithm that had two possible results; either the passenger boarding was part of a transfer boarding or it was a single journey. The concept of iterative classification is to solve a complex problem by dividing it into simpler smaller problems [8]. The solutions of the sub problems can then be combined to solve the complex main problem. The iterative classification algorithm was used as a classification technique to differentiate between transfer journeys and single journeys. It uses a series of decision functions or decision tests to classify the identity of an object [9].

The algorithm is based on the selection and comparison of individual passenger boardings and verifies whether certain data attributes match or variables apply. The individual passenger records were then linked together when identified as a transfer journey 'pair'. The output of the iterative classification algorithm is defined in Structured Query Language (SQL) statements, which feed the newly generated knowledge back into the relational database thus facilitating transfer journey analysis.

A. Definition 1

A transfer journey consists of two different individual passenger boardings that were recorded following these rules:

- Both passenger boardings had to be recorded on the same day
- A second passenger boarding occurred less than 90 minutes after first ticket validation
- Ticket type of the passenger had to be the same between the two passenger boarding records
- Ticket ID had to be the same for the two passenger individual boarding records. The ticket ID and ticket type ID create a unique identifier for each passenger over a certain period of time (depending on the ticket type).
- Route ID had to be different for the two individual passenger boardings

The 90-minute constraint was based on a preliminary analysis of transfer journeys, their peak times and the travel patterns of transfer passengers. This study looked at 10,000 transfer journeys and their time differences between boarding at A1 and at B1. It was found that the 90 minute constraint would be sufficient to identify all transfer journeys. As defined before, a transfer journey is a one-way course of travel having a single main purpose [2]. Following this definition and focusing on the 'single main purpose' statement it could be argued that no journey with a single main purpose will last longer than 90 minutes. This study therefore indicated the 90 minute constraint as the limit of valid transfer journeys. As the time difference between A1 and B1 is a derivable attribute, the analysis can be restricted to transfer journeys that had e.g. a time difference of 60 minutes if one wished to do another analysis using different time constraints.

B. Definition 2

The algorithm is required to produce a link between the passenger boardings if and only if all the rules of definition 1 apply. This results in a SQL statement that will later update the database and therefore enrich the data records with the information whether the transaction record was part of a transfer journey or not. The algorithm uses the following procedures (see Fig. 2):

Step 1: Open the first file - Each file (i) is processed individually as each file stores the data of one day.
Step 2: All variables are declared and initialised. A data structure is required to store the various data attribute values.
Step 3: Read boarding records x and compare with all remaining records (yn) and verify that the conditions listed in steps 4, 5, 6 and 7 are true. If NO, compare record x to next record (y=y+1).
Step 4: Compare Ticket ID number of the two records to check if the transfer journey was carried out by the same passenger. If NO, compare record x to next record (y=y+1).

Step 5: Compare matching Ticket Type as the same Ticket ID could have been issued for more than one Ticket Type but only once for the same Ticket Type. If NO, compare record x to next record (y=y+1).
Step 6: Compare the boarding time of each boarding and ensure that they occurred less than 90 minutes apart from each other (tx > ty-90 minutes). If NO, compare record x to next record (y=y+1).
Step 7: Compare Route ID of each boarding and ensure that they are different. The argument is that when the same route is taken twice within 90 minutes it is not considered a passenger transfer journey as it could be assumed that the journey does not have a single main purpose anymore. A single main purpose would be to get from home to work. If the passenger stops to go shopping and then continues the journey to work it can be seen as two individual journeys, as the main purpose of the first journey was to go shopping and the second purpose was to get to work.
Step 8: Generate SQL statement.
Two SQL statements are generated for each identified transfer journey. One SQL statement updates the already existing table that stores all individual passenger boardings. The second SQL statement inserts new data into a table which only stores data of transfer journeys.
Step 9: Compare the next transaction record (x=x+1) to all remaining records starting at step 4.

For simplicity and performance optimisation (runtime of the algorithm) the algorithm only focused on transfer journeys with one transfer point. A study on a small subset of the data indicated that less than 1% of all transfer journeys have two or more transfer points (the passenger boards three or more buses to reach his/her destination.

IV. GENERATED SQL STATEMENTS

The outcomes of the algorithm are SQL statements which enrich the database with information on whether the passenger boarding was part of a transfer journey or not. This information can then be used to carry out further statistical or other more detailed analysis. When two passenger boardings were identified as a transfer journey (Steps 4, 5, 6 and 7 were true) the program generated two SQL statements. The UPDATE statement was responsible for linking the second boarding of the transfer journey to the first passenger boarding. The INSERT INTO statement was responsible for generating a new record, which consists of Boarding ID of the first passenger boarding and the Boarding ID of the second passenger boarding.

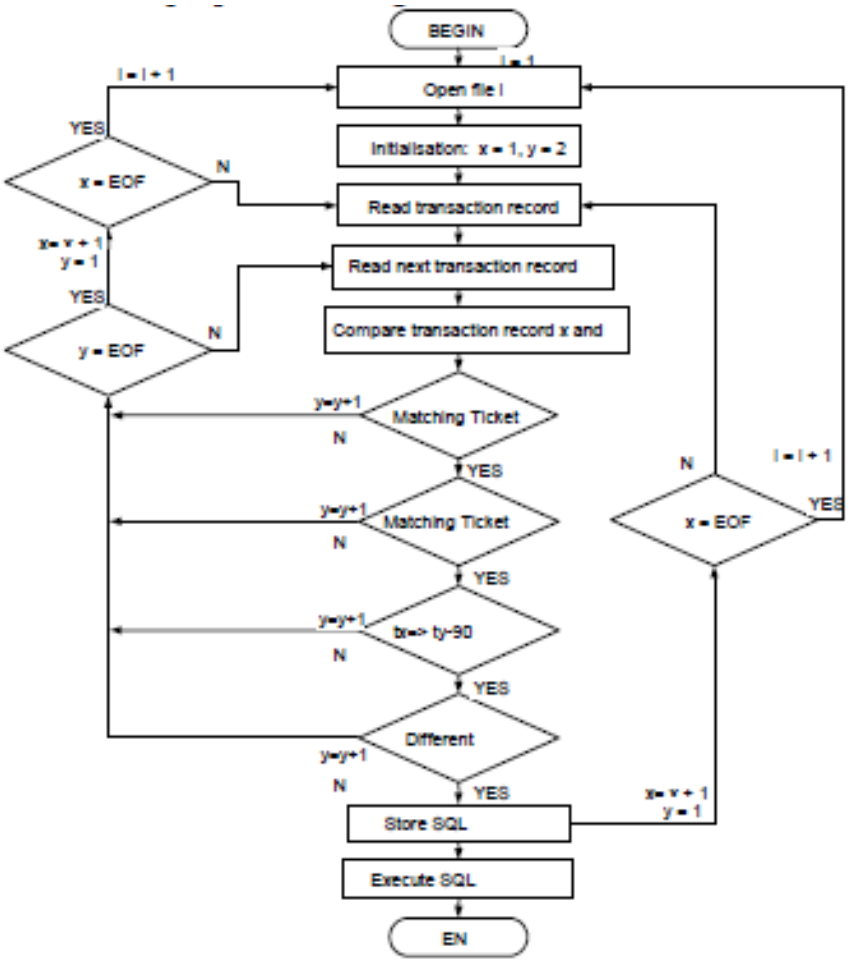

Fig. 2: Iterative Classification Algorithm

A. Final result of the iterative classification algorithm

Table I shows the final transfer journey table with the data attributes and its name and description. The newly created database table shows a new record for each transfer journey. Each of these records is based on data from two individual passenger boardings (A and B). Each transfer journey is detailed by the attributes listed in Table I. The new table facilitates much more detailed analysis about transfer journeys than the original database offered. It is now possible to generate analysis that will contribute to a clearer understanding of the behaviour of transfer journey passengers. Further analysis can be applied to improve decision support at an operational or policymaking level.

TABLE I: LIST OF AVAILABLE VARIABLES FOR STATISTIC ANALYSIS

Name	Description
ID	Unique ID for each transfer journey record in the transfer journey table
ID A	Link to individual passenger boarding A
ID B	Link to individual passenger boarding B
A Route ID	Route ID of individual passenger boarding A
B Route ID	Route ID of individual passenger boarding B
A Bus Stop	Bus Stop ID of individual passenger boarding A
B Bus Stop	Bus Stop ID of individual passenger boarding B
A Direction	Direction of individual passenger boarding A
B Direction	Direction of individual passenger boarding B
A Zone	Zone of individual passenger boarding A
B Zone	Zone of individual passenger boarding B
A Area	Area description of individual passenger boarding A
B Area	Area description of individual passenger boarding B
A Time	Boarding time of passenger boarding A
B Time	Boarding time of passenger boarding B
Ticket ID	Unique ticket ID of transfer journey
Ticket Type	Ticket Type used

V. TRANSFER ANALYSIS

The previous section described the process of linking two formerly independent boardings together by comparing the various parameters of each of the journeys. The two independent boardings built a record of a transfer journey which provides the opportunity for further analysis of the EFC data. The variables identified in Table I will be used throughout this section to create various statistics and analyses.

The new transfer journey attribute allows for many types of analyses:
1. Descriptive statistics about transfer journeys
2. Time difference between boarding A1 and B2
3. Relationships between time difference and hour of boarding at A1
4. Transfer analysis and spatial visualisation by zone
5. Transfer analysis and visualisation by suburban or city centre areas
6. Route matrix analysis of transfer journeys
7. Symmetry analysis of transfer journeys and single journeys
8. Direction combination pair analysis of transfer journeys
9. Time analysis of transfer journeys (peak/off-peak, day of the week)
10. Ticket type and ticket category analysis of transfer journeys
11. Analysis of abnormalities on key routes and substitutional routes
12. Micro-level analysis such as waiting times at transfer nodes

A. Time difference constraint

For example, some analyses could restrict all transfer journeys to a time difference between A1 and B1 to 45 or 60 minutes by simply ignoring the transfer journey records that do not conform with the set time difference.

The time difference between A1 and B1 is one of the deciding parameters in relation to whether a passenger boarding is part of a transfer journey. The maximum accepted time difference between boarding at A1 and boarding at B1 has been set to be less than 90 minutes. As this is a derived data attribute (Boarding time at B1 – Boarding time at A1) this parameter can be redefined when carrying out various analyses on the final data. Fig. 3 shows the distribution of the time differences in 5 minute intervals.

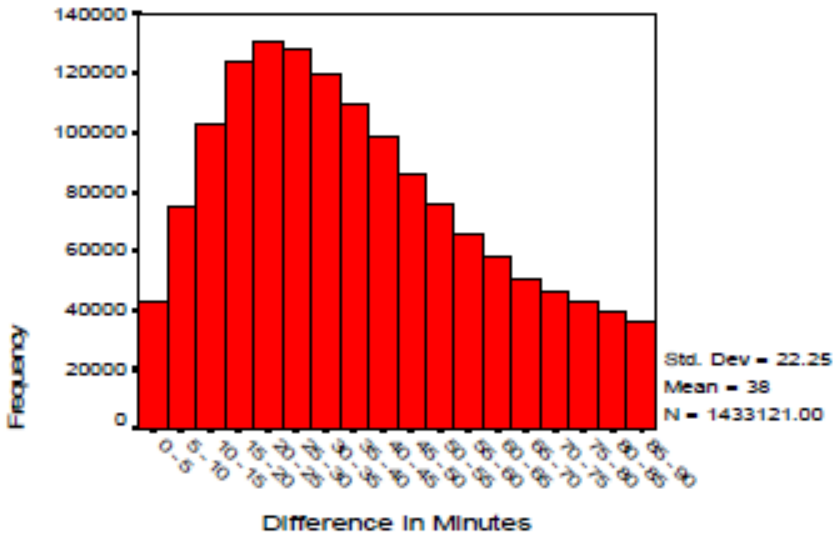

Fig. 3: Histogram of the variable 'Time Difference'

The middle 50% of all transfer journey boardings from A1 to B1 take between 20 and 53 minutes. Almost 10% of transfer passengers transfer within 10 minutes after boarding the first bus. 50% of all transfer customers transfer within 34 minutes. The highest peaks of passenger transfers are between 18 and 28 minutes after boarding the first bus (A1). Only 10% of the transfer passengers transfer between 72 and 89 minutes after boarding the first bus (A1).

B. Spatial analysis of transfer journeys

The 3-dimensional bar chart (see Fig. 4) shows the total number of transfers from the first zone (spatial zone in which the bus stop is located) (A1) to the second zone (B1). The zones are spatial classification codes of city areas that have been introduced for modelling purposes. 64% of all transfer journeys transfer in zone Z. Almost 77% of all transfer journeys started outside the city

centre zone (Z). The graphs also show that transferring within the same zone is very common. The charts highlight that a great percentage of transfer passengers transfer in the city centre zone (Z).

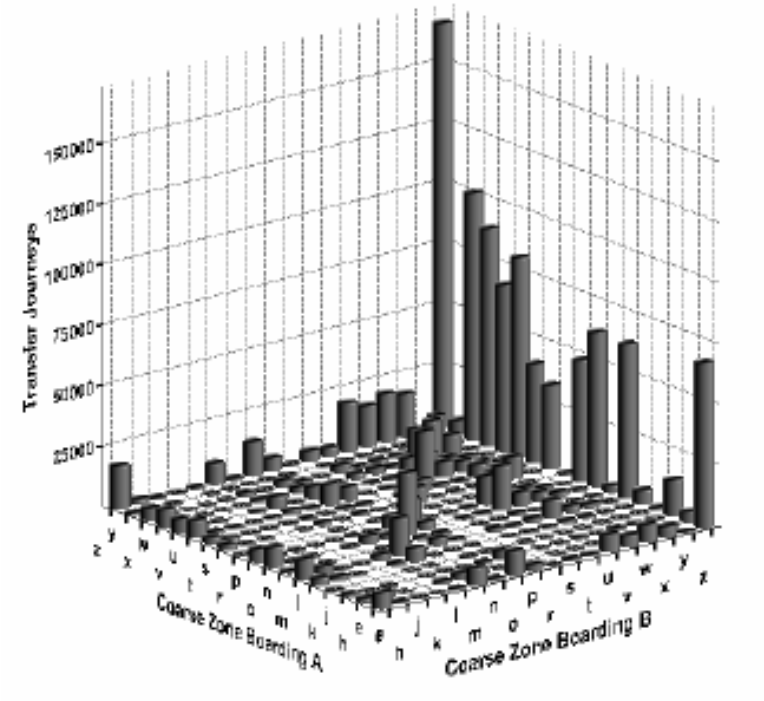

Fig. 4: Three dimensional bar chart

Such analyses can be useful when passenger behavior and travel patterns need to be examined. If geographic coordinates are known for the various locations such analysis can lead to more detailed results.

C. Waiting times

In order for the public transport system to provide a fully integrated service whereby passengers can transfer seamlessly between services, the waiting time for connecting services needs to be minimised. Table II illustrates an example of how waiting time on transfer nodes can be extracted from the database.

Two different transfer points are analysed in detail. Transfer point one is at the intersection of Route A (bus stop 6) and Route B (bus stop 8). Transfer point two is at the intersection of Route A (bus stop 30) and Route B (bus stop 36). In this example, data is taken from one particular day showing bus arrival times of Route A (inbound, arterial) and Route B (outbound, arterial but in a different direction to outbound of Route A) during the morning peak time. This analysis can be generated for all transfer points for any time period. Although it is unknown where and what time a passenger alights (because tickets are not validated on exit), the time a bus arrives at a bus stop is recorded by the EFC system. In addition to this, the exact time when the passenger validates the ticket for boarding and the number of the bus he/she is transferring to is also stored. This allows the

assumption that a passenger who transfers to another route alighted the previous bus at the nearest bus stop of the new route. Table II shows waiting times at transfer points. The arrival time of Route A at stop 6 determines the arrival time of the passenger(s) (e.g. 07:41). The passenger's boarding time (who was originally on the bus Route A) onto Route B determines the departure time of this bus. This allows the calculation of the waiting time by subtracting departure time from arrival time. The information generated could serve as a performance indicator of an integrated transport network. Scheduled arrival times could be optimised in such a manner as to minimise passenger waiting times at transfer nodes. It can also be used as a monitoring tool to evaluate the quality of service provided for the customers. The arrows in Table II specify the waiting time pairs that have to be subtracted.

TABLE II: WAITING TIME AT TRANSFER POINTS

	Transfer Point 1		
	Arrival Route A - Stop 6	Arrival Route B - Stop 8	Waiting Time in minutes
Arrival times of buses at the bus stop	07:41	07:33	5
	07:45	07:38	1
	07:55	07:46	3
	08:10	07:47	6
	08:12	07:51	4
	08:13	07:53	3
	08:26	07:58	4
	08:34	08:16	9
	08:34	08:21	9

D. Transfer journey matrices

After creating the transfer journey table it is now possible to create various matrices, which may serve as input for further statistical analysis or transport modelling. The analyst can define the aggregation level as well as the dimension of matrices. Two different sample matrices are described in the following paragraphs.

A transfer node identification matrix shows the volumes of transfers at the A2 bus stop and each B1 bus stop where a transfer journey took place. The nodes are defined by route ID and bus stop ID so the node identification 'A/10' reads bus stop '10' on route 'A'. Although the exact bus stop of A2 is unknown, it can be assumed that it is the closest bus stop to B1 as the transfer passenger boarded the bus at this bus stop. It can therefore be said that B1 is a transfer node. The significance of this transfer node depends on the daily volume of transfer passengers. In this study there are 8,270 bus stops and therefore the size of the matrix is 8,270 * 8,270. Such a matrix can be used to analyse the

effectiveness of routes at a micro level. For example, is there a need for orbital or crosscity routes where currently only arterial routes exist? The main purpose of such a matrix is for bus planning and policymaking. It could serve as input matrix for different transport models.

Another matrix can be developed to show the transfer volume of passengers that occurred between all combinations of routes. Various parameters such as peak/off-peak time, ticket types, age group or direction of travel can be applied while producing the matrix. This increases the usability and efficiency of these matrices considerably.

VI. CONCLUSIONS

The primary aim of this paper was to introduce an algorithm that automatically identifies transfer journeys from a pool of single journeys. The algorithm was applied to a dataset with almost 8 million individual passenger boardings identifying over 1.4 million transfer journeys over a four-month period. Detailed information about transfer journeys can be examined on micro and macro levels, which will (a) contribute to the understanding of transfer passengers and (b) add to the network optimisation with regard to operational and strategic planning and policymaking. The following conclusions can be made: The iterative algorithm can be applied to vast amounts of EFC data in order to identify transfer journeys within a transport network. This newly gained information can be used as the foundation for many transfer journey related analyses and therefore supports decision and policymaking. Restructuring of the database will improve data access and retrieval time. Table I showed the new table with all available attributes which can be used to execute various analyses.

As the time difference between A1 and B1 is a derivable attribute, the analysis can be restricted to transfer journeys that had a time difference of only 60 or 75 minutes. The analysis of this attribute showed that the middle 50% of all transfer journey time differences are in between 20 and 53 minutes. 50% of all transfer passengers transfer within 34 minutes. Only 10% of all transfer passengers transfer between 72 and 89 minutes.

The results of the algorithm enabled the implementation of spatial analysis for the first and second boarding, the results of which are displayed. It identified that 64% of all passengers transferred in the City Centre. The starting area (first boarding) is more dispersed among the various zones. This analysis has been based on data of an entire day but this can also be produced by time of day, age category of passenger (child, student, adult or pensioner), or by any other filter that can be applied to the dataset.

The spatial analysis is based on a matrix showing the number of transfers and the location of the two boardings (A1 and B1). This analysis also shows that the City Centre (zone 1) is the main location where transfers take place. Another observation is that there are a considerable amount of transfers within each zone, indicating that passengers use a feeder bus to one of the main routes leading into the city centre or taking a feeder bus after leaving the main route originating from the city centre.

Analyses with regard to transfer journeys can be generated on a macro level as well as on a micro level. This was shown in a waiting time analysis for two bus stops that serve as transfer nodes within the transportation network. Table II shows two transfer nodes and the associated waiting time for passengers that transferred from Route A/Stop 6 to Route B/Stop 8. The arrival times of both buses allowed the calculation of the waiting time of transfer passengers. The analysis of these two bus stops revealed an average waiting time of 6 minutes. As data of the entire network is available this analysis could be created on a network wide scale.

Transfer node matrices, route matrices and area matrices can be obtained by generating crosstabulation tables based on the data stored in Table I. Analyses of values stored in these matrices may contribute to the understanding of passenger movements, route utilization and transfer node identification.

REFERENCES

[1] Allen, J. G., Hammel, D. and T. G. Beyer. Chicago's Information and Physical Coordination Study: Transit Transfers from the Customer's Perspective. CD-ROM. Transportation Research Board, National Research Council, Washington, D.C., 2003.

[2] DETR. Transport Trends. Great Britain Department of the Environment, Transport and the Regions, The Stationery OfficeBooks, UK, 2000.

[3] Bagchi, M. and P. White. Use of Public Transport Smart Card Data for Understanding Travel Behaviour. CD-ROM. Conference Proceedings of Association for European Transport, Strasbourg, France, 2003.

[4] TCRP. Transit Capacity and Quality of Service Manual. Transportation Research Board, National Research Council, Transit Cooperative Research Programme, Washington, D.C., 1999.

[5] Hofmann, M., O'Mahony M. M. and B. Tierney. A Framework to Utilise Urban Bus Data for Advanced Data Analysis. CD-ROM. Conference Proceedings of the 10th World Congress on ITS, Madrid, 2003.

[6] Navick, D. S. and P. G. Furth. Using Location-Stamped Farebox Data to Estimate Passenger-Miles, OD Patterns, and Loads. Transit Transfers from the Customer's Perspective. CD-ROM. Transportation Research Board, National Research Council, Washington, D.C., 2002.

[7] Furth, P.G. Data Analysis for Bus Planning and Monitoring, Synthesis of Transit Practice 34, Transit Cooperative Research Programme,Transportation Research Board, National Research Council, National Academy Press, Washington, D.C. 2000.

[8] Gama, J. and P Brazdil. Linear Tree. Elsevier Inc, Journal of Intelligent Data Analysis, Vol 3, Issue 1, 1999, pp1-22.

[9] Ishwar K. S. and J. H. Yoo. Structure-Driven Induction of Decision Tree Classifiers Through Neural Learning, Elsevier Inc, Journal of Pattern Recognition, Vol. 30, No. 11, 1997, pp. 1893-1904.