Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.
UTILISATION OF ELECTRONIC FARE COLLECTION DATA OF URBAN BUS OPERATORS WITH REGARD TO TRANSFER JOURNEYS AND ORIGIN/DESTINATION ESTIMATION

By
Markus Hofmann

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
AT UNIVERSITY OF DUBLIN - TRINITY COLLEGE DUBLIN, IRELAND APRIL 2008
The author hereby declares that this thesis, in whole or in part, has not been submitted to any other university as an exercise for a degree. Except where reference has been given in the text, it is entirely the author's work. The author agrees that the library may lend or copy the thesis upon request. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Dated: April 2008

Author: Markus Hofmann
Abstract

The understanding of an urban public transport network from an operational point of view and the understanding of passenger's travel patterns become increasingly important due to the growing complexity of most networks. Large cities heavily depend on their public transport networks as urbanisation vastly increases. One way of obtaining more factual knowledge of network's performance measures and passenger travel patterns is to analyse Electronic Fare Collection (EFC) data which are recorded when buying or validating tickets. Dublin's main urban bus operator, Dublin Bus, records trip level data from passengers that use a ticket that is valid for a period of time. Such tickets have a unique identifier for the validity period which allows to analyse travel paths of individual passengers. Over 45 million individual passenger boardings were recorded over a two year period.

One of the most sought after data attributes of passengers records is the information where they alight as this would add considerably to the transparency of the network and therefore to the decision making of public transport operators and strategic planners. This thesis describes the process of developing an algorithm that estimates the final location of public transport passengers using a rule based method. This method successfully infers a destination to an average of 45% of all boarding records. Various validation techniques are used to prove the correctness and robustness of the algorithm. Several analyses are carried out that focus on the estimation of performance measures such as in-vehicle time, in-vehicle time variability and waiting time at transfer nodes and its variability.

The thesis further proposes a method to identify transfer journeys and thoroughly analyses the identified results including network symmetry, transfer node volumes, travel time and route matrices.

Furthermore, one aspect of the thesis focuses on the development of a method that allows automatic identification of substitutional routes by analysing EFC data. This was achieved with an average error rate of 2%.
To my wife

Glenda

and my son

Killian
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table of Contents</td>
<td>ix</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xvii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xxi</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>xxv</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>xxvii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introductory Remarks</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Research Aim</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Research Objectives</td>
<td>2</td>
</tr>
<tr>
<td>1.4 Dublin Bus</td>
<td>3</td>
</tr>
<tr>
<td>1.5 Electronic Fare Collection Systems</td>
<td>7</td>
</tr>
<tr>
<td>1.6 Research Problems</td>
<td>8</td>
</tr>
<tr>
<td>1.7 Research Context (Scope and Audience)</td>
<td>9</td>
</tr>
<tr>
<td>1.8 Structure of the Thesis</td>
<td>9</td>
</tr>
<tr>
<td>2 Background</td>
<td>11</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>11</td>
</tr>
<tr>
<td>2.2 Automatic Data Collection Systems</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Electronic Fare Collection Systems</td>
<td>13</td>
</tr>
<tr>
<td>2.3.1 Definition of EFC</td>
<td>13</td>
</tr>
</tbody>
</table>
# 5 Transfer Journey Analysis

5.1 Introduction .......................... 77
5.2 General Descriptive Statistics of Transfer Journeys .............. 77
5.3 Analysis of Time Difference between Boardings ............... 78
  5.3.1 Relationship between Time Difference of A1/B1 and Boarding Hour at A1 79
5.4 Transfer Matrices ............................................. 80
  5.4.1 Transfer by Coarse Zones .............................. 81
  5.4.2 Transfer Node Matrix ................................. 83
  5.4.3 Transfers by Area Description ..................... 83
  5.4.4 Transfer Route Matrix .............................. 84
5.5 Analysis of Transfer Journey by Day of Week .................. 86
5.6 Transfer Analysis by Ticket Type .................................. 86
5.7 Ticket Type Category ......................................... 87
5.8 Ticket Duration Period ....................................... 88
5.9 Time Analysis ................................................. 88
5.10 Symmetry Analysis of Transfer Journeys ......................... 89
  5.10.1 Introduction and Definitions ....................... 89
  5.10.2 Initial Analysis of Symmetry Patterns ............ 92
  5.10.3 Development of the 'Degree of Symmetry' Equation 92
  5.10.4 Results of Degree of Symmetry Equation for Transfer Journeys 95
  5.10.5 Degree of Symmetry of Single Journeys ........ 97
  5.10.6 Statistical Summary of Results .................... 99
5.11 Statistical Model to Analyse Symmetry ........................... 101
  5.11.1 Final Comments to the Degree of Symmetry ........ 104
  5.11.2 Summary of Network Symmetry .................... 105
5.12 Summary ....................................................... 105
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Introduction</td>
<td>107</td>
</tr>
<tr>
<td>6.2 The Concept</td>
<td>108</td>
</tr>
<tr>
<td>6.2.1 The Main Assumption in Greater Detail</td>
<td>109</td>
</tr>
<tr>
<td>6.2.2 Trip Chaining</td>
<td>111</td>
</tr>
<tr>
<td>6.3 The Data Source</td>
<td>113</td>
</tr>
<tr>
<td>6.4 Substitutional Routes</td>
<td>115</td>
</tr>
<tr>
<td>6.4.1 Introduction</td>
<td>115</td>
</tr>
<tr>
<td>6.4.2 The Execution</td>
<td>116</td>
</tr>
<tr>
<td>6.4.3 Statistics of Number of Occurrences and Calculated Weights</td>
<td>120</td>
</tr>
<tr>
<td>6.4.4 Finding the Optimal Cut-off Points</td>
<td>121</td>
</tr>
<tr>
<td>6.4.5 Summary</td>
<td>125</td>
</tr>
<tr>
<td>6.5 Development of the Rule Base</td>
<td>125</td>
</tr>
<tr>
<td>6.5.1 Introduction</td>
<td>125</td>
</tr>
<tr>
<td>6.5.2 Decision Tables</td>
<td>125</td>
</tr>
<tr>
<td>6.5.3 Determining the Rule Attributes</td>
<td>126</td>
</tr>
<tr>
<td>6.5.4 Identified Rules</td>
<td>131</td>
</tr>
<tr>
<td>6.6 The Initial OD Algorithm</td>
<td>132</td>
</tr>
<tr>
<td>6.7 Extended OD Extraction Algorithm</td>
<td>134</td>
</tr>
<tr>
<td>6.7.1 Introduction</td>
<td>134</td>
</tr>
<tr>
<td>6.7.2 One Boarding Record per Ticket</td>
<td>135</td>
</tr>
<tr>
<td>6.7.3 Two Boarding Records per Ticket</td>
<td>135</td>
</tr>
<tr>
<td>6.7.4 Three or more Boarding Records per Ticket</td>
<td>137</td>
</tr>
<tr>
<td>6.7.5 Summary of Algorithm Improvements</td>
<td>139</td>
</tr>
<tr>
<td>6.8 Algorithm Innovations</td>
<td>139</td>
</tr>
<tr>
<td>6.9 Algorithm Output</td>
<td>141</td>
</tr>
<tr>
<td>6.9.1 Single Journeys</td>
<td>141</td>
</tr>
<tr>
<td>6.9.2 Transfer Journeys</td>
<td>142</td>
</tr>
</tbody>
</table>
6.10 Testing of the Code and Performance of the Algorithm ............................................ 142
6.11 Validation of the Results .......................................................................................... 143
  6.11.1 Survey Results ...................................................................................................... 143
  6.11.2 Validation Using a Simple Random Sample .................................................. 146
  6.11.3 Repeated OD Pairs ............................................................................................... 148
  6.11.4 Travel Diary Study ............................................................................................... 149
  6.11.5 Validation Based on the OD Results .............................................................. 150

6.12 Limitations and Boundaries of Data and the Algorithm ............................................. 152
  6.12.1 Main Assumption ............................................................................................... 152
  6.12.2 Assumption – Subroutes ..................................................................................... 152
  6.12.3 Transfer Journeys ............................................................................................... 153
  6.12.4 Data Source – Lack of Data .............................................................................. 153
  6.12.5 Small Sample ...................................................................................................... 153
  6.12.6 Unknown Routes and Ticket Types ................................................................ 154

6.13 Summary .................................................................................................................. 154

7 OD Results and Their Application to Determine In-Vehicle Time ........................................... 156
  7.1 Introduction ............................................................................................................... 156
  7.2 Summarised Results of the OD Extraction Algorithm ........................................... 156
    7.2.1 Results by Ticket Type ......................................................................................... 157
    7.2.2 Results by OD Classes ........................................................................................ 160
    7.2.3 Summary ............................................................................................................. 163

7.3 Comparison of the Results to other Algorithms ...................................................... 163

7.4 Implementation of the OD Algorithm ........................................................................... 164

7.5 Waiting Time and In-Vehicle Time Analysis ................................................................. 165
  7.5.1 Introduction ............................................................................................................. 165
  7.5.2 Extending the Database with Arrival Times of Passenger .................................. 166
  7.5.3 In-Vehicle Time for Single Journeys on a Network Level .................................... 167
8 Discussion & Conclusions

8.1 Summary of Research and Main Contributions ........................................ 178

8.1.1 Data Migration Framework ........................................................... 178

8.1.2 Transfer Journey Classification ...................................................... 178

8.1.3 Analysis of Data Quality with Regard to Bus Driver Behaviour ...... 179

8.1.4 Transfer Journey Analysis ............................................................. 179

8.1.5 Network Symmetry ...................................................................... 179

8.1.6 Identification of Substitutional Routes ........................................... 179

8.1.7 OD Extraction of Journey Pairs ..................................................... 180

8.1.8 OD Extraction of Entire Passenger Boarding Records ................. 180

8.1.9 Validation of the OD Results ......................................................... 181

8.1.10 Calculation of In-vehicle Time, Total Transfer Journey Time and Waiting Time at Transfer Nodes ........................................ 181

8.1.11 Summary ................................................................................... 181

8.2 Can EEC Data Improve Information Need of Operators? ................. 182

8.3 Main Findings and Future Implications ............................................. 183

8.4 Suggestions for Further Research .................................................... 187

8.4.1 What Data are Really Needed? ..................................................... 187

8.4.2 Development of a Multi-modal/Multi-operator Algorithm ............. 188

8.5 Final Remarks ................................................................................. 189

Bibliography ...................................................................................... 190
Appendices

A Coarse Zone Description ..................................................... 202
B Coarse Zone Map ............................................................................................................. 203
C Substitutional Routes ...................................................................................................... 204
D OD Algorithm Flow Charts ............................................................................................. 209
List of Tables

1.1 Demographic Projections of the GDA ............................................ 4
1.2 Breakdown of Dublin Bus Fleet - Source: Dublin Bus, 2002 .......... 6
1.3 Examples of Business Questions and Scenarios ............................... 8

2.1 Differences in Data Collection Technologies .................................... 12
2.2 Survey Results of Rating Statements ............................................. 25

4.1 Passenger Boarding Records .......................................................... 52
4.2 Descriptive Statistics of Total Route Journey Time ......................... 53
4.4 Validation Results of the Simple Random Sample of Transfer Journey Records 58
4.5 Summary of Classification Error (as %) for Time Attribute Simulation Study . 60
4.6 Summary of Classification Error (as %) for Time Attribute (60 Minute Range) Simulation Study .................................................. 62
4.7 Summary of Classification Error (as %) for Time Attribute (15 Minute Range) Simulation Study .................................................. 62
4.8 Summary of Classification Error (as %) for Ticket ID Attribute Simulation Study .................................................. 63
4.9 Summary of Classification Error (as %) for Ticket Type Attribute Simulation Study .................................................. 63
4.10 List of Available Variables for Statistic Analyses ............................ 64
4.11 General Information about the Analysed Routes ............................. 66
4.12 Inbound - Outbound - Total Passenger Numbers ............................ 71

5.1 Number of Journeys Generated by Magnetic Strip Cards .................. 78
5.2 Descriptive Statistics of the Variable 'Time Difference' .................... 78
5.3 Total Number of Transfers by Zones at A1 and B1 ............................ 83
3  Solution 1 - AND ........................................................... 206
4  Solution 2 - OR ............................................................. 207
5  Solution 2 - AND ........................................................... 208
List of Figures

1.1 Map of the Greater Dublin Area (DTO, 2001) ......................................................... 4
1.2 Map of Bus Routes of Dublin Bus ............................................................................. 5
3.1 Distribution of Waiting Time during Peak and Off-peak time (Okamura et al., 2003) .................................................................................................................................. 37
4.1 4-Phase Framework of Importing Wayfarer Data into an Oracle Database .......................................................... 42
4.2 Raw Data File Records .................................................................................................. 43
4.3 Entity Relationship Diagram of the Project Database ........................................... 45
4.4 Transfer Journey Definition ........................................................................................ 48
4.5 Transfer Journey Data Diagram ................................................................................ 49
4.6 Iterative Classification Algorithm for Transfer Tourney Identification .................. 55
4.7 Monte Carlo Simulation Results ............................................................................... 61
4.8 Layout of Routes ......................................................................................................... 67
4.9 Sample Layout of a Transport Network Section .................................................. 68
4.10 Inbound - Outbound Bus Stops ................................................................................. 69
4.11 Frequency of Recorded Bus Stops ............................................................................ 70
4.12 Frequency and Location of Recorded Bus Stops ................................................... 72
4.13 Arrival Time at Bus Stops ........................................................................................ 73
5.1Histogram on Time Difference between A1 and B1 ................................................. 79
5.2 Sunflower Scatter Diagram of Time Difference and Hour of Boarding ................. 80
5.3 Three Dimensional Bar Chart ..................................................................................... 82
5.4 Start Zone of Transfer Journeys (A1) ........................................................................... 82
5.5 Patterns of Boardings by Hour (Week 18/10/99 - 24/10/99) ................................... 89
5.6 Relationship between Route Combination Pairs ...................................................... 92
5.7 Degree of Symmetry Chart - Transfer Journeys .................................................... 96
5.8 Degree of Symmetry Chart - Single Journeys ....................................................... 99
5.9 Scatter Plot of Symmetry and Recorded Boardings .............................................. 100
5.10 Histogram of Degree of Symmetry of all Individual Routes ................................. 101
5.11 Scatterplot of Route Symmetry ............................................................................ 102
5.12 Histogram of asymmetric values of $s$ ............................................................... 103
6.1 The Structure of the OD Extraction Algorithm ...................................................... 108
6.2 Visualisation of the Assumption for Single Journeys ............................................ 110
6.3 Visualisation of the Assumption for Transfer Journeys ........................................ 110
6.4 Directional Scenarios ......................................................................................... 112
6.5 Data Input Format for the OD Algorithm ........................................................... 113
6.6 Examples of Subroute Scenarios ......................................................................... 115
6.7 Relationships between Number of Occurrences and Weights ............................ 121
6.8 Time Difference of Journeys ............................................................................... 144
6.9 3-D Bar Charts of OD Pairs ................................................................................ 145
6.10 Correlation Analysis of OD Matrices ................................................................. 146
6.11 Time Difference between OD Pair Boarding Records by User Category .......... 151
6.12 Boarding Time Distribution of First and Second Boarding Time Records ........ 151
7.1 Extracting Arrival Times ..................................................................................... 166
7.2 Histogram of In-vehicle Time of Single Journeys on a Network Level ............... 168
7.3 Boarding Time Distributions ............................................................................... 169
7.4 In-Vehicle Times .................................................................................................. 170
7.5 In-vehicle Time Variability of Single Journeys .................................................... 171
7.6 Boarding Times of Transfer Passengers ............................................................ 172
7.7 Journey Length of Transfer Journeys .................................................................. 173
7.8 Transfer Journeys Waiting Time and its Variability .......................................... 174
1 Solution 1 - OR - Impact Errors of False Predictions .......................................... 206
I wish to express my gratitude to both of my supervisors, Professor Simon Wilson (Trinity College Dublin) and Professor Peter White (University of Westminster) for their continued encouragement and invaluable suggestions during this work.

I want to thank my wife Glenda for her love, encouragement and support over the past five years. I would also like to thank my son Killian for our little play breaks while being in the final stages of this research. I love you both very much.

A special thought is devoted to my parents and grand parents for their never-ending support.

Many thanks to everyone within the Civil Engineering Department for making my time in the department memorable. In particular I would like to Clare Finnegan and Brian A. Caulfield for their help, motivation and chats. Thanks also technicians and their contributions in the form of tea, biscuits and laughter.

I would also like to thank all my friends who contributed directly or indirectly to this work. In particular Bryan Duggan and Brendan Tierney whose support at key stages of the research helped me to finally hold this thesis in my hands.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADA</td>
<td>Americans Disability Act</td>
</tr>
<tr>
<td>AVAS</td>
<td>Automated Voice Annunciation System</td>
</tr>
<tr>
<td>AVL</td>
<td>Automatic Vehicle Location</td>
</tr>
<tr>
<td>CBD</td>
<td>Central Business District</td>
</tr>
<tr>
<td>CBR</td>
<td>Case-Based Reasoning</td>
</tr>
<tr>
<td>CIE</td>
<td>Córas Iompair Éireann</td>
</tr>
<tr>
<td>CTA</td>
<td>Chicago Transport Authority</td>
</tr>
<tr>
<td>DD</td>
<td>Double Decks</td>
</tr>
<tr>
<td>DTLR</td>
<td>Department for Transport, Local Government and the Regions</td>
</tr>
<tr>
<td>DoT</td>
<td>Department of Transport</td>
</tr>
<tr>
<td>DSS</td>
<td>Decision Support System</td>
</tr>
<tr>
<td>DTO</td>
<td>Dublin Transport Office</td>
</tr>
<tr>
<td>EFC</td>
<td>Electronic Fare Collection</td>
</tr>
<tr>
<td>ERD</td>
<td>Entity Relationship Diagram</td>
</tr>
<tr>
<td>FTA</td>
<td>Federal Transit Administration</td>
</tr>
<tr>
<td>GDA</td>
<td>Greater Dublin Area</td>
</tr>
<tr>
<td>GH</td>
<td>Giant Hop</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>ID</td>
<td>Identification</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>ITEA</td>
<td>Integrated Transport Economics and Appraisal</td>
</tr>
<tr>
<td>ITS</td>
<td>Intelligent Transport Systems</td>
</tr>
<tr>
<td>LH</td>
<td>Long Hop</td>
</tr>
<tr>
<td>LOS</td>
<td>Level of Service</td>
</tr>
<tr>
<td>MB</td>
<td>Mini Buses</td>
</tr>
<tr>
<td>MB</td>
<td>Mega Byte</td>
</tr>
<tr>
<td>MH</td>
<td>Medium Hop</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Form</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>MHz</td>
<td>Mega Hertz</td>
</tr>
<tr>
<td>NTS</td>
<td>National Travel Survey</td>
</tr>
<tr>
<td>NTD</td>
<td>National Transit Database</td>
</tr>
<tr>
<td>MTR</td>
<td>Mass Transit Railway</td>
</tr>
<tr>
<td>NYCTA</td>
<td>New York City Transit Authority</td>
</tr>
<tr>
<td>NYMTC</td>
<td>New York Metropolitan Transportation Council</td>
</tr>
<tr>
<td>OAP</td>
<td>Old Age Pensioner</td>
</tr>
<tr>
<td>OD</td>
<td>Origin Destination</td>
</tr>
<tr>
<td>ODMG</td>
<td>Object Data Management Group</td>
</tr>
<tr>
<td>OLAP</td>
<td>Online Analytical Programming</td>
</tr>
<tr>
<td>OODBMS</td>
<td>Object Oriented Database Management Systems</td>
</tr>
<tr>
<td>ORDBMS</td>
<td>Object Relational Database Management Systems</td>
</tr>
<tr>
<td>PAYG</td>
<td>Pay As You Go</td>
</tr>
<tr>
<td>PL/SQL</td>
<td>Procedural Structured Query Language</td>
</tr>
<tr>
<td>PMT</td>
<td>Passenger Miles Travelled</td>
</tr>
<tr>
<td>PTE</td>
<td>Public Transport Executive</td>
</tr>
<tr>
<td>QBC</td>
<td>Quality Bus Corridors</td>
</tr>
<tr>
<td>QOS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RBR</td>
<td>Rule Based Reasoning</td>
</tr>
<tr>
<td>RDBMS</td>
<td>Relational Database Management System</td>
</tr>
<tr>
<td>RTPI</td>
<td>Real Time Passenger Information</td>
</tr>
<tr>
<td>SH</td>
<td>Short Hop</td>
</tr>
<tr>
<td>SPSS</td>
<td>Statistical Package for the Social Sciences</td>
</tr>
<tr>
<td>SRIA</td>
<td>Substitutional Route Identification Algorithm</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>RBR</td>
<td>Rule Based Reasoning</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SD</td>
<td>Single Deck</td>
</tr>
<tr>
<td>TCQSM</td>
<td>Transit Capacity and Quality of Service Manual</td>
</tr>
<tr>
<td>TCRP</td>
<td>Transit Cooperative Research Program</td>
</tr>
<tr>
<td>VBA</td>
<td>Visual Basic for Applications</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Introductory Remarks

Obtaining information and knowledge about business operations and customer behaviour is a critical activity in any business in order to support decision and policymaking. In the urban public transport domain this is often carried out in the form of surveys, which aim to produce performance measures from an operational point of view and also to explore what type of passenger behaviours are present and whether there is an obvious trend with regard to public transport demand.

The use of information technology (IT) allows urban bus operators to analyse their automatic collected data that, for example, originate from electronic fare collection (EFC) systems. Such systems often store data on a trip level which means that information of each passenger boarding is recorded. In addition, magnetic strip cards or smart cards are often employed on a network wide level which may add another level to the collected data as EFC systems also store details of each individual card.

Combining this information with the information that was stored when passengers boarded the bus produces a powerful dataset that has the potential to be developed into a decision support tool for public transport planners, regulators and government departments. Information such as how many passengers boarded per period (hour, day, month, year) by route segment, route or network can easily be produced once the data are stored in an accessible database. Other examples are on-time performance by vehicle, boarding by location or total travel time of each vehicle.

Though these performance measures are easily extracted, the dataset could be used for more complex analysis. For example, transfer journey analyses for which passengers need to transfer to a second or third bus in order to reach their final destination can be explored. However, the ultimate goal of such an analysis is to obtain knowledge of the passenger's final destination.
If the location of all passengers is known, then the planning and managing of the network’s resources would be much more effective and efficient. Travel times, boarding location, alighting locations, in-vehicle time, waiting times at transfer nodes and change in public transport demand over a certain period of time could be accessible to transport planners. Knowing the origin/destination (OD) of all passenger journeys would allow a more complete analysis of the network and its operators.

This thesis proposes a method to determine transfer journey parameters and information about the passenger’s final destination by analysing historical datasets which were automatically recorded using EFC systems. This is done using various reasonable assumptions which are implemented in several algorithms. A data mining approach called rule based reasoning is used to verify records with expert domain knowledge. The results are then further enhanced by improving the algorithm and are thoroughly validated.

1.2 Research Aim
The main research aim of this project is to estimate OD pairs on a passenger trip level using EFC data of an urban bus operator. Methods that estimate destination information for rail networks already exist. Barry et al. (2002), Rahbee (2003) and Bryan (2007a) state that no method exists for urban bus networks. Further justification of this research can be found in Chapter 6. As mentioned previously, such information can be valuable when trying to understand passenger behaviour or when analysing other network performance measures such as waiting times at transfer nodes. In order to reach this aim several research objectives need to be fulfilled. These objectives are listed in the following section.

1.3 Research Objectives
Research objectives were important stages that all focused on the main aim. Often these objectives revealed new knowledge without necessarily being directly linked to the OD estimation aim. These analytical results are also presented throughout this thesis.

The research aim is achieved by meeting the following specific objectives:

- **to investigate the existing body of knowledge with the aim to produce a gap analysis.**
  This is mainly with the aim to explore existing methods and results to determine a gap analysis that justifies the research including its methods, techniques and interpretation.
to develop a transferable and universal data migration framework. The EFC data are generally stored in semi-structured and semi-encrypted text files. In order to query and analyse hundreds of millions of records, it is necessary to build an environment in which the records can be explored in a more effective and efficient manner. The correct migration procedure needs to be combined with a method that verifies the data attributes with regard to correctness of its values;

- to identify transfer journeys (linked trips) from the EFC dataset. The original dataset consists of unlinked trips only. Unlinked trips are individual boardings without considering journeys where a passenger had to transfer onto another bus to get to the desired final destination. However, it is important for the OD extraction algorithm to know which records are unlinked and which trips are transfer journeys (linked trips). An algorithm needs to be developed that uses several assumptions in order to classify the records into two categories: linked and unlinked trips. This algorithm needs to be tested and validated;

- to analyse the transfer journey results set. This objective focuses on the analyses of transfer journey with the aim to more fully understand existing patterns, transfer nodes and waiting times;

- to provide a novel method to infer destinations of trip level passenger boarding records of a urban bus EFC system. This method needs to be justified, tested and validated. Furthermore it is aimed to focus on the precision of the algorithm rather than the amount of destinations that could theoretically be identified;

- to demonstrate the application of the results of the OD inference method. This will show how the results could be used to calculate performance measures that are valuable to operators and authorities.

1.4 Dublin Bus

Dublin Bus is Ireland’s largest public transport agency. In consideration of the fact that the data that are available are from 1999, the following section is mainly based on the annual report of Dublin Bus (DB, 1999) and the Scott Wilson report which was published in June 2000 (Wilson, 2000). The company ran a fleet of 1,054 buses, which provided according to Dublin Bus ‘... a comprehensive bus service for the city and county of Dublin and its hinterland...’ (DB, 2005). The total number of boardings in 1999 were 193 million.
increases with regard to population growth of the area served to a large extent by Dublin Bus (see Table 1.1). This includes an increase in population of the GDA of 17% from 1.4 million in 1996 to 1.65 million in 2016. The metropolitan area is expected to have an increase of 12.5% (450,000 in 1996 to 660,000 in 2011). Due to increasing house prices the population growth in the hinterland is expected to rise by 36%.

Figure 1.1 shows a spatial map of the GDA.

![Figure 1.1: Map of the Greater Dublin Area (DTO, 2001)](image)

Table 1.1: Demographic Projections of the GDA

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Population (million)</td>
<td>1.35</td>
<td>1.41</td>
<td>1.46</td>
<td>1.75</td>
</tr>
<tr>
<td>Households (’000)</td>
<td>402</td>
<td>446</td>
<td>521</td>
<td>675</td>
</tr>
<tr>
<td>Employment (’000)</td>
<td>452</td>
<td>549</td>
<td>681</td>
<td>878</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>16%</td>
<td>12%</td>
<td>6%</td>
<td>5%</td>
</tr>
<tr>
<td>Car Ownership (per 1000 population)</td>
<td>247</td>
<td>292</td>
<td>342</td>
<td>480</td>
</tr>
<tr>
<td>% Growth in GDP since 1991</td>
<td>-</td>
<td>42%</td>
<td>79%</td>
<td>260%</td>
</tr>
</tbody>
</table>

According to Dublin Bus (DB, 2005)
The principal objective of the urban public transport agency [is] to provide a passenger service by road for the city and county of Dublin and contiguous areas and to provide ancillary services, within the state and between the state and places outside the state; conferred by the board by the Transport Act, 1950.

The agency provides an extensive network of bus routes - radial, cross-city, peripheral, DART feeder, Airlink, Nightlink, and sightseeing tours. Dublin Bus currently employs approximately 3000 people. This network can be seen in Figure 1.2.

Apart from the headquarters office located in the city centre there are seven other locations, which are called depots, in the GDA. Each depot is responsible for a certain region within the GDA and is managed by the depot manager. The depot manager has the responsibility to schedule and set routes, make sure all transport units are operable, and that human resources are used most appropriately. The managers also have the authority to change routes or to increase or decrease frequency of the buses.

The Dublin Bus fleet consists of Minibuses (MB), Single Decks (SD) and Double Decks (DD), which are distributed to different depots around Dublin City. The numbers shown in Table 1.2 are taken from the website of Dublin Bus (DB, 2005). The bus fleet is broken down
Dublin Bus is embracing Intelligent Transportation Systems (ITS), and have had magnetic strip cards installed on buses for over a decade. Dublin Bus is also in the planning stages of introducing a system of integrated ticketing with the use of contact-less smart cards. Dublin Bus uses EFC boxes to validate either pre-purchased magnetic strip cards or to issue paper tickets for mostly single fares. The magnetic strip cards, which cannot be purchased on board, are available in various options (1 Day, 3 Day, 5 Day, and 7 Day, monthly, and annual). No change is given for purchasing single fares on board, but the remainder can be reclaimed at the head office in the city centre. For tickets bought on board, the bus operator chooses the fare category and the stage of the journey. The boarding of Old Aged Pensioners (OAP), who have free travel, are recorded by the EFC system. The OAP's make up about 13-14 percent of all passenger journeys. Dublin Bus is also in the process of testing a system of Real Time Passenger Information (RTPI) called Q-time on the Lucan, Clondalkin and Ballyfermot Quality Bus Corridors (QBC's). This system is currently in a trial stage and it is anticipated that it will be rolled out across the Dublin Bus over the next couple of years.

The fares and ticketing strategy of the DTO stated that "integrated fares and ticketing will be introduced that allows all public transport users to complete a full journey with only one ticket ...making most journeys involving interchange cheaper" (DTO, 2003). This will further increase the potential of EFC data as multi-operator data can potentially be combined to yield more advanced results.

The EFC system implemented on the entire fleet of Dublin Bus is provided by the UK company Wayfarer. The following components are part of the system:

- Bus mounted control units (keypad, clock, control software, and ticket issuing machine);
1.5. ELECTRONIC FARE COLLECTION SYSTEMS

- Bus mounted magnetic strip card validation units. These units are connected to the main control unit so that the boarding record is recorded;
- Each driver has a data cartridge which is responsible for storing the information produced by the main control unit and the magnetic strip card validation unit;
- Cartridge readers that copy the data of each individual cartridge. Each bus depot has one computer that is responsible for gathering all the recorded information;
- Central computer which accumulates the data collected by all depot computers;
- ‘Gold cartridges’ are programming cartridges that have the ability to re-program the bus mounted control units. This is necessary to add new ticket types or additional bus routes;
- In addition there are a number of software programmes in operation that allow the re-programming of the ‘gold cartridges’ and to create basic, mostly revenue based reports based on the recorded data.

Although the technological equipment that is mounted in the buses originally served mainly as a revenue control tool, the advances in IT allow a more thorough analysis of the data that is stored on a daily basis. This aspect will be explored throughout the thesis.

1.5 Electronic Fare Collection Systems

Data produced by electronic fareboxes is primarily used to ensure correct revenue collection. Public transport operators often neglect the potential of such a large dataset. Decisions with regard to operation or schedule planning are often made without consulting the past records stored in data files (Furth, 2000) even though the extraction of that information stored in the files can be straightforward if the data are stored in a database. Running queries relating to e.g. headway, frequency, bunching, etc. may be done effortlessly and cost effectively provided the database is set up correctly.

Other industries use more advanced data analysis techniques such as data mining to comprehend information about the organisation that is stored in the database. Telecommunications, insurance and retail sectors apply Online Analytical Programming (OLAP) or data mining techniques (Inmon, 2002) to improve and optimise their operation management and strategic planning. However, it is clear that advanced data analysis of EFC data is still in its infancy throughout the public transport sector (Furth, 2000). Furth (2000) further says that it is very common among public transport agencies to store the passenger journey data only for revenue analysis reasons. However, this data can be used to provide Quality of Service (QOS) indicators or to
assist in planning and scheduling tasks.

In a data-poor environment, planners and schedulers must depend on ad hoc feedback from vehicle operators and passengers to justify adjusting routings, running times, and service frequencies (Strathman et al., 2002). Information-rich data on the other hand may provide detailed information on bus running times and passenger loads (Furth, 2000) that may be used to improve operations and scheduling. Furth (2000) and Strathman et al. (2002) agree that several performance indicators can be used together to evaluate operating conditions. For example, headway maintenance and passenger load data can be jointly analysed to determine if heavy load patterns are a consequence of insufficient frequencies or headway maintenance problems. Such analyses can be carried out once the EFC data are stored in a structured format.

Apart from solving any technological problems related to the implementation of large data structures it is important to identify areas that would benefit from various EFC analyses. An understanding of the transport sector is crucial in order to produce results that are relevant and useful for public transport agencies. The examples shown in Table 1.3 show some of the business questions and scenarios that could be addressed using EFC data.

<table>
<thead>
<tr>
<th>Department</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marketing</td>
<td>Where do passengers, who use a weekly/monthly ticket, live?</td>
</tr>
<tr>
<td>Finance</td>
<td>Forecasting and calculating annual growth and the consequent impact studies of that growth. Route profitability forecasts.</td>
</tr>
<tr>
<td>Operations and Scheduling</td>
<td>Reports about various performance measures. E.g., what are the boarding patterns of all stages on route XYZ? On-time performance analysis.</td>
</tr>
<tr>
<td>Strategic Planning</td>
<td>How does a newly developed housing estate impact on passenger volume of a route? How many buses will the agency need to provide in 1 year, 3 years, or 5 years?</td>
</tr>
<tr>
<td>Business Development</td>
<td>What is the impact of rolling out newly structured weekly/monthly tickets and how will it influence passenger loads? What would be the most beneficial pricing structure?</td>
</tr>
</tbody>
</table>

Section 2.3 will further detail the concept of EFC systems and also explore the potential of its automatically collected data.

### 1.6 Research Problems

Research problems were identified in the following areas:

- As the project is mainly data driven it was anticipated that problems would arise due to the vast amount of data. The data needed to be checked and then migrated into a database without losing information. Furthermore, the data used for this project are not
the entire population as the records of cash paying customers are not available. This caused problems when the final results needed to be interpreted;

- Validation of the results turned out to be critical but also problematic. This was mainly due to the lack of a representative and validated sample result of destinations. Other forms of validation were therefore used;

- The runtime of all alterations, extensions and program executions took a considerable amount of time. It was therefore important that all functions and processes were designed and developed with the runtime aspect in mind.

- Dublin Bus did not want to contribute to this project due to sensitivity of information which made it difficult at stages to deduct missing bits of information or interpreting anomalies within the data.

1.7 Research Context (Scope and Audience)

This thesis may be of interest to researchers or practitioners who focus on the analysis of large datasets, public transport performance measure analysis, level of service analysis, transfer journey analysis and Origin/Destination destination and analysis. The developed algorithms and methods may be applied to datasets that contain the main attributes required for the procedures. This may also be applicable for a network that consists of different modes (e.g. bus, light/heavy rail) and/or multiple operators. The key lies in the data and whether these can be formatted to such an extent that the algorithm can apply the chosen assumptions.

1.8 Structure of the Thesis

Chapter 2 provides an overview with regard to Automatic Data Collection (ADC) systems. This includes EFC, automatic vehicle location, and automatic passenger counting systems. All these systems store large amounts of data that may serve as foundation for basic and advanced analysis. The chapter further explores the potential of such datasets and compares these to more traditional methods of data collection such as surveys.

Chapter 3 provides a literature review of the research area. The main focus lies on the estimation of OD matrices using historical data recorded by EFC systems. The estimation of OD matrices for public transport networks using historical data is only at its infancy and available publications are limited. A brief overview of alternative OD estimation techniques is also provided.
Chapter 4 focuses on the main data and how those data are migrated into a database and then how various extensions facilitate further analyses. A newly developed universal 4-Phase data migration framework is introduced. This chapter further elaborates on the design and development of an iterative classification algorithm. The algorithm identified transfer journeys (also known as linked trips) which play an important role throughout the rest of the thesis. The results of this algorithm are thoroughly validated. The final section of the chapter analyses the impact of human involvement throughout the data recording stages.

Chapter 5 focuses on the analysis of the newly identified transfer journey data that were generated using the iterative classification algorithm. Such an analysis emphasises the importance of understanding passenger behaviour and operational performance measures for decision support and decision-making. A large part of the chapter investigates the network with regard to its symmetry. A new method to calculate a Degree of Symmetry is proposed. This analysis is carried out on both single and transfer journeys.

The OD estimation algorithm and all its components are described in Chapter 6. This includes the introduction of a novel method to infer substitutional routes based on the public transport passenger boarding patterns recorded. Classes and rules are created and justified to build the rule base which forms the core of knowledge that the OD extraction algorithm accesses during the inference process. A second iteration of the OD algorithm analyses and evaluates emerging patterns that occurred throughout the entire ticket life span which is then used to further identify OD pairs. This is followed by a thorough validation process and a discussion of the limitations and boundaries of both data and algorithm.

Chapter 7 initially shows some aggregate figures with regard to the performance of the algorithm itself such as results grouped by ticket type, OD pairs and repeated OD pairs. The second part of this chapter focuses on the estimation of in-vehicle time and its variability for some selected sample routes with the aim to demonstrate some applications of the extracted OD information.

Chapter 8 summarises the research and comments on the research objectives that were stated at the beginning of the project. This is followed by a brief summary and discussion on the methodology applied. The main contributions of the body of knowledge are listed. The chapter also discusses the main findings and possible future implications. Suggestions for further research are provided, followed by the final remarks.
Chapter 2

Background

2.1 Introduction

Urban public transport networks have often millions of customers who travel regularly using the services offered. The electronic systems that are implemented on the public transport fleet serve primarily to improve service and operations of the urban public transport provider. These systems also produce large datasets as a by-product which in turn can be used by the operators to improve decision making. This chapter provides a brief overview of the main Automatic Data Collection (ADC) systems that are widely used. It then explores the Electronic Fare Collection (EFC) systems in greater detail.

Rule Based Reasoning is explained in detail as this concept was used to estimate passenger’s destinations. Furthermore, the rule base and various advantages and disadvantages of this method are outlined.

A study carried out at the University of Westminster under the supervision of Prof. Peter White focused on the utilisation of EFC data. In particular the information needs of public transport operators, county councils and Passenger Transport Executives are presented. These information needs are revisited in Section 8.2 to determine whether the proposed methods of this research contributed in providing required information.

2.2 Automatic Data Collection Systems

Using innovative analysis approaches on large datasets has been one of the fastest growing and changing business areas over the last decade. It is often known as data mining, business intelligence, or knowledge discovery. All terms related to the analysis of large datasets use techniques such as rule based reasoning, neural networks, clustering, or classification. There is
also a more recent trend where public transport operators use their automatically recorded data to make more informed decisions.

Data originating from ADC systems have a number of advantages which can be utilised by the urban transport operator. The first advantage is that the required technologies to record the data are already in place. It can therefore be considerably cheaper to analyse ADC data than to plan and execute a survey. Furthermore, surveys are expensive and are therefore rarely carried out on a network wide level whereas ADC data are generally available for the entire network.

The second advantage is the amount of data that can be utilised while analysing the datasets. Whereas surveys generally rely on samples and only cover a small area of the network, ADC systems facilitate the data collection of the entire network at all times. For example, it is therefore easier to focus on trend analyses with regard to changing parameters such as supply and demand, passenger behaviour or operational focus. Depending on the analysis that a study requires it is also likely that the ADC data eliminate bias as the data recorded are more exact than people’s perceptions or estimations. For example, in-vehicle time analyses often rely on customer perception whereas ADC data can determine a more exact value. Due to the spatial and temporal coverage ADC systems can offer not only the replacement of some surveys but also facilitate the analyses of previously unknown scenarios. However, it is noteworthy to mention that ADC data also lacks in information which may have been possible to capture using surveys. This could include the trip purpose as well as detailed demographics.

It is further possible to focus on special events or weather conditions (Zhao, 2004; Hofmann and O’Mahony, 2005a).

Table 2.1 compares various aspects of manual and automatic data collection techniques (adapted from Wilson et al. (2005)):

<table>
<thead>
<tr>
<th>Manual</th>
<th>Automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>low capital cost</td>
<td>high capital cost</td>
</tr>
<tr>
<td>high marginal cost</td>
<td>low marginal cost</td>
</tr>
<tr>
<td>small sample sizes</td>
<td>large sample sizes</td>
</tr>
<tr>
<td>aggregate</td>
<td>more detailed, disaggregate</td>
</tr>
<tr>
<td>unreliable</td>
<td>errors and biases can be estimated and corrected</td>
</tr>
<tr>
<td>limited spatially and temporally</td>
<td>ubiquitous</td>
</tr>
<tr>
<td>more flexible data collection (e.g. trip purpose)</td>
<td>restricted to ADC system data attributes</td>
</tr>
</tbody>
</table>

Most of the differences in Table 2.1 are self explanatory. The high capital cost of automatic data collection cannot be solely assigned to the analysis aspects as the technology is required
whether analysis on the data is carried out or not. However, in the case where large datasets need to be analysed the costs of workstations and/or servers need to be covered as well as software cost and human resources.

The following main ADC systems will be introduced in greater detail: Electronic Fare Collection (EFC) Systems; Automatic Passenger Counters (APC); and Automatic Vehicle Location (AVL) Systems.

2.3 Electronic Fare Collection Systems

2.3.1 Definition of EFC

EFC systems are implemented to ensure the correct validation of tickets and to issue correct tickets to public transport passengers. EFC systems are also known as Electronic Fareboxes, Automatic Fare Collection, or On-board Ticket Processors. Initially the data was mainly collected for revenue purposes. Nowadays, innovative projects emerge that use the vast amount of data to improve decision making. This could include decisions with regard to

- improved fare policies (Fleishman et al., 1996);
- fare change forecasting (Fleishman et al., 1996);
- understanding ridership patterns (Fleishman et al., 1996) by trip, stage, route segment, route and network level using different periods of time such as peak, off-peak, weekday, weekend, etc.;
- estimating passenger miles using average trip length (Furth, 2000);
- running time and running time variability (Furth, 2000);
- route economic performance comparisons (Furth, 2000);
- on-time performance (Furth, 2000);
- system level schedule adherence (Furth, 2000);
- trend analysis to identify fare collection policy problems (Stern, 1997);
- ridership/capacity ratio (Parkinson and Fisher, 1996);
- headway (Parkinson and Fisher, 1996);
- schedule adherence (Parkinson and Fisher, 1996; Furth, 2000);
- trip summaries (Furth, 2000);
- Origin-Destination inference at trip level (Barry et al., 2002; Hofmann and O'Mahony, 2005c; Zhao, 2004; Trepanier et al., 2007a).
2.3.2 Types of EFC Systems

There are two main differences when implementing EFC systems. One system is an entry-only system which means that the ticket is only checked/validated when the public transport passenger boards the vehicle. This type of EFC system is most commonly implemented. Because the ticket is only checked at the boarding, no data is recorded when and where the passenger alights. Dublin Bus has an entry-only system installed.

The entry-exit validation systems require a valid ticket for entry and exit of the vehicle. There are only a few examples where exit-entry systems are implemented. For example, in the USA only Washington D.C.’s Metro and the BART network in California successfully use entry-exit validation. In the UK, London uses an entry-exit system for their underground services. Singapore uses contact-less smart cards to record entry-exit validation on their bus network (May, 2004). From a data analysis point of view such systems would be ideal as the exit of a passenger is recorded and linked to the entry using a unique ticket ID. Knowing the travel path of all passengers would lead to an environment where the passenger movements of an entire transport network is transparent. Although exit validation is possible for metro and light rail stations it might not be feasible for bus networks unless the technology advances to such extend that the ticket is validated without any delay while exiting the bus. Initial tests were carried out by Siemens in Switzerland where tickets were entirely contact-less (known as vicinity cards) meaning that the public transport vehicle automatically detected and validated the tickets of all passengers (Siemens, 2001). Such validation occurred on entry and then on exit. The smart card was then charged depending on the exact distance travelled. However, such systems have not been implemented on a commercial basis.

2.3.3 Methods of Payment

More recent systems have introduced magnetic strip cards and smart cards as methods of payment on public transport vehicles. Some EFC systems also accept on-board cash payments. With respect to the aspect that is researched throughout this thesis it is important to realise that EFC data records of cash paying customers are not traceable. This means that cash paying customers are entirely anonymous. However, passengers that validate a ticket when boarding the vehicle are not entirely anonymous as each ticket has a unique ticket ID (for most EFC systems). This ticket ID facilitates analyses such as the one that is presented in this thesis. There are significant differences between smart cards and magnetic strip cards. For example, smart
2.3. ELECTRONIC FARE COLLECTION SYSTEMS

cards can often be re-charged for as long as 5 years (Bagchi and White, 2005). The main difference, however, is that smart cards can process and store information. This allows the operator to load period travel passes, single and return travel tickets, stored value and concessionary entitlements onto the smart card (Bagchi, 2003a).

There is also a difference with regard to data that are recorded when using smart cards and magnetic strip cards. For example, smart cards may have cardholder information and purchase information stored either on a centralised system or on the card itself. Although some of the cardholder information is protected by privacy laws, other information can be used without violating any of these laws. Such exclusions from the privacy laws could be area of residence, age, sex, employment, etc..

Magnetic strip cards were initially introduced in London and the Long Island Rail Road (LIRR) in 1964, which led to the first stored value application in 1966 (Fleishman et al., 1996). Nowadays most network operators offer magnetic strip cards or smart cards to customers. The general trend to integrated ticketing also pushes the trend to use smart cards as a payment method.

As this thesis does not focus on the method of payment, smart cards and magnetic strip cards are not further explored.

2.3.4 Measurement Errors

Measurement errors can occur either before, during or after the validating or issuing process of the ticket. Correct policies and procedures can minimise such errors (Trepanier et al., 2007a). Furth (2000) lists the following errors:

- The bus driver does not indicate the correct stop or presses a wrong key on the control panel (further see Section 4.6.1);
- The cassette has not been signed on correctly and therefore fails to record the boardings or the dataset lacks completeness due to missing data attributes (e.g. route identifier);
- Once the bus has returned to the depot the staff do not extract the data from the cassette which may lead to loss of the data;
- Fare categories are either ambiguous or not complete (applies to cash fares only);
- Problems with the equipment due to the manufacturer or lack of maintenance.

The following items can be added to this list:
• New bus stops have not been added to the software that programs the fare media reader. This may lead to wrong assignments of passenger boardings;
• Magnetic strip cards are not as reliable as smart cards and tend to cause reading errors more frequently;
• Passengers with Old Aged Pensioner (OAP) passes or annual passes often do not validate the ticket but only show it to the bus driver. This leads to an incomplete data source;
• Complexity of fares structure may affect extent to which stop-level records are updated (e.g. flat versus graduated).

Some of these errors and their impact on the subsequent data analysis will be addressed throughout this thesis.

2.3.5 Summary of EFC Systems
Data that were extracted from an EFC system build the core data source for this project. They are also the most versatile with regard to data analysis. Although it is not possible to identify alightings from the dataset (unless for entry-exit systems) or real time location information, EFC data form the most suitable base for historical data analysis.

2.4 Automatic Passenger Counters
2.4.1 Definition of APC
APC systems are a significant component for estimating ridership volumes with regard to service planning, scheduling and forecasting (Boyle, 1998). These systems record passenger boardings and alightings using various different technologies. They were first used in the 1970s (Schiavone, 1999). Some APC systems use mechanical sensors that record a change in pressure when a passenger's foot is set onto the device. Other systems use infrared devices which record the alighting and boarding of buses depending which of the beams is broken first. Others again can use optical imaging (Furth, 2000). Generally, a limited proportion of the fleet is equipped with APC systems which are then dispatched on all routes over a certain period of time. The information can be downloaded in different ways including radio frequency, cable or extractable storage devices (Schiavone, 1999). The passenger count needs to be recorded in combination with the location of the vehicle. This can be done using odometer readings or AVL information (Furth, 2000). Although this was not found in any literature the location information could also be extracted using EFC systems where the common link of the two datasets could be the time of the recordings.
2.4. AUTOMATIC PASSENGER COUNTERS

APC systems may be used for the following applications (Casey et al., 1996):

- Provide detail to the dispatcher to support informed decision making on immediate cor­rective action (e.g., short-turn the empty bus, not the full one);
- Provide information to RTPI systems;
- In the case where there is a requirement to submit data to a national transit database, passenger trips and passenger miles can be calculated using the system;
- Future scheduling;
- Planning and positioning new bus shelters for passengers. The information could be used with other demographic information in a GIS application;
- Improve fleet planning, scheduling and dispatching.

One of the main benefits of APC systems is the decreased data collection cost and the increased type and range of data that becomes available. Some of the cost is reduced as almost no time needs to be invested into the pre-processing stage of the data. The data are stored electronically without the need to enter them into a computer system as would be the case for many manual counts. The automatically recorded data may contribute to improved overall operating efficiency.

2.4.2 Measurement Errors

As with EFC systems, APC systems can also have measurement errors that may bias the results. Furth (2000) lists the following reasons for measurement errors:

- Hardware malfunctioning which leads to incorrect reading results;
- Miscounting passengers;
- Failing to identify the correct bus stop;
- Incorrect data segmentation;
- Incorrect sign-on.

It is critical to the overall results to verify the recorded data. For example, to compare the on-counts with the off-counts for the period of measurement. Other literature focuses on validation, sampling and statistical inference of APC data (Furth and McCollom, 1987; Strathman and Hopper, 1991; Kimpel et al., 2003).
2.5 Automatic Vehicle Location Systems

2.5.1 Definition of AVL

AVL systems mainly serve as additional system components for tasks such as measuring system performance (e.g. ridership volume, on-time performance), improving real time passenger information (RTPI) displays, announcing next stop information for on-board passengers, and displaying vehicles on electronic maps for both passengers and operational managers (Okunieff, 1997). The primary task of AVL systems is to identify the location of the vehicle in real-time. The improvements today in the area of global positioning systems (GPS) allows the extraction of exact spatial locations even in city centre areas. Chicago Transit Authority (CTA) for example uses the AVL system in combination with an automatic stop announcement system (AVAS). However, this AVL system solely serves the AVAS system and location data is not sent to the control centre in real-time.

AVL systems were firstly implemented in 1969 (Okunieff, 1997). However, many urban public transport operators still do not operate AVL systems for the entire fleet as the installation and maintenance often outweigh the benefits. Many operators only equip vehicles with AVL when an RTPI system supplies passengers with estimated arrival times based on real-time data feeds.

AVL systems sometimes also store vehicle location data on on-board computers which can then be used to analyse trip times. For the purpose of control centres’ data requirements it is enough to receive location data in 90 second intervals (Furth, 2000). However, this is sometimes not often enough or exact enough to use the data as trip time data source. The data that are recorded may contribute to more widespread and detailed planning information at a much lower cost and higher accuracy than manual collection could achieve.

There are many benefits that can be attributed to AVL systems. This may include improvements of on-time performance, opportunity to react to disruptions and bus operators can overview schedule adherence (Casey et al., 1996). This results in an improved level of service which in turn may increase attractiveness and thus ridership. It further increases safety and security as the location of the vehicle is known. Another application of AVL systems is a traffic signal preferential treatment system (Casey et al., 1996).

2.5.2 Measurement Errors

Measurement errors in AVL systems are often linked to the technology used to implement the system. The following is a list showing some of the errors that can lead to data problems:
• Depending on the technology applied, transmission errors may occur (Fallon, 2000);
• Lack of update in infrastructural changes may lead to cause errors (Higgens et al., 2000);
• Data recording may rely on passing a signpost in case the odometer and signpost or wayside method is implemented. This method also lacks in providing continuous location information (Jones et al., 2003);
• Ground-based radio may result in transmission errors or failures due to tall buildings which would result in missing data (Jones et al., 2003);
• When differential GPS is implemented then differential correction must be altered and updated frequently. If this is not maintained wrong data may be recorded (Jones et al., 2003);
• AVL systems are often only installed for some routes which means that there is never a time when data are available for the entire network.

2.6 Manual Data Collection

Manual data collection is still part of most public transport operators. This may include data collection techniques such as operator trip cards, traffic checkers (Furth, 2000) and various different types of surveys. By nature, manual data collection techniques are mostly applied to a small sample of the population over a limited period of time due to the related costs. Surveys that particularly focus on the estimation of OD matrices will be introduced in greater detail throughout this thesis.

2.7 Potential of ADC Data

ADC technology offers great potential for future data analysis. Integrated systems could be developed which complement each other. This could lead to a data structure that facilitates more complex approaches with regard to various analyses. Furthermore, each transport network can potentially record the same data attributes. The introduction of standards with regard to data recording, data storage and data extensions would greatly benefit future progress in this research area. Having such standards would allow software to be developed that can serve many operators and therefore decrease development cost and increase usability and functionality of such programmes. In addition, regional or national databases can be created that offer comparisons of different networks, calculation of performance measures of similar networks, and the option to learn from successful network changes or other implementations.
Another possibility of such an ADC database would be the possibility to add other non-transport related datasets. Such an application was shown in Hofmann and O’Mahony (2005a) where EFC data was combined with hourly weather records to see what impact adverse weather has on urban bus service delivery. Another potential area where an ADC database could be used is to analyse the impact of obstructions such as construction sites or traffic accidents on urban bus service delivery. Knowing about these impacts would improve decision making with regard to proactive and informed dynamic changes to ensure service delivery.

Unfortunately, most systems that are currently in place were implemented for other reasons but data collection. AVL systems, for example, were implemented in Dublin to facilitate correct operations of the RTPI system. The data are only stored temporarily and are not kept for further analysis. The AVL system is not used in combination with the EFC system. The EFC systems themselves still serve mostly as a revenue collection tool without any considerations of data needs.

Nevertheless, EFC and other ADC data are potentially very useful data sources. As shown in this thesis, a normal dataset allows one to infer transfer nodes, passenger destinations, in-vehicle time, waiting time, and its variability. There are many more measures that can be extracted from EFC data.

### 2.8 Rule Based Reasoning

Rule Based Reasoning (RBR) is a technique that uses a reasoning process to connect data to conclusions. It is mostly used in Expert Systems that are concerned with problem solving approaches (Ralston et al., 2000). It is the formal definition of the thinking process when the aim is to extract patterns by using production rules. These rules are the most common method to represent knowledge (Ralston et al., 2000). Rule based reasoning therefore finds its application in most disciplines reaching from linguistics to medicine to transportation.

There are two main approaches to knowledge representation (Tan et al., 2006):

1. Procedural representation is when the knowledge and the manipulation of the knowledge is linked, as in most computer code;
2. Declarative representation separates knowledge and the algorithm to manipulate knowledge. This method allows changing the knowledge without interfering with the algorithm
itself and therefore is preferable when developing reusable inference procedures. This is the method used for this algorithm.

2.8.1 Rules in RBR
A rule is composed of a condition and an action - more commonly known as IF and THEN parts. The IF part consists of a logical combination of conditions whereas the THEN part states the action which has to be taken in the case where the conditions are fulfilled. For example IF (certain conditions apply) THEN (take appropriate action). This type of knowledge representation is known as action-oriented (Ralston et al., 2000).

Two main types of IF/THEN rules exist:

- **Propositional statements** are variable free and describe a particular instance. The following example illustrates why the propositional statement is not suited to represent general relations with a large number of cases:

  \[
  \text{IF Female(Mary) and Divorced(Mary) THEN WasMarried(Mary)}
  \]

  The example emphasises the shortcomings of this method. Although it is known that Mary is female and divorced and therefore was married the rule is fixed to the (logical) constant Mary. A method that allows for a more flexible approach is required to represent the knowledge that is known for this project;

- **First-order statements** overcome this shortcoming by introducing variables in the rules (Duda et al., 2001):

  \[
  \text{IF Route(a,b) and DifDirection(a,b) THEN PotentialOD(a,b)}
  \]

  This statements says that if the routes of boarding records \( a \) and \( b \) are equal and are going in different directions they are part of a potential OD pair. This example demonstrates the increasing complexity of the rule which now can be applied to the entire dataset. A first-order logic rule can also be applied to constants (propositional rules). It is further possible to include functions in the IF/THEN rule such as

  \[
  \text{IF PotentialOD(a,b) and (IsRepeated(a,b) < 1) THEN OD_case(a,b,x)}
  \]

  where \( \text{IsRepeated}(a,b) \) is a function returning the number of repeats of this combination as a numerical value. The expression \( (\text{IsRepeated}(a,b) < 1) \) returns one of the Boolean variables TRUE or FALSE which is used in the rule. The complexity of the rules governs
the granularity of conditions and conclusions. By changing the granularity rules can become more specific or more general (Pedrycz, 2002).

Rules in RBR are highly readable and have the advantage of being modular and very modifiable. This means that rules can be added and deleted in a flexible fashion (Pedrycz, 2002). Rules can be used in combination with database applications and also have the advantage that they can be easily interpreted (Duda et al., 2001). The values of the attributes can contain numerical and non-numerical data types which is a very important predicate for this project as various different data types are contained in the dataset. RBR is particularly suitable for problems where general relationships among entities can be categorised into classes. The relationships can then be defined by rules applying IF/THEN statements (Duda et al., 2001).

The disadvantage of a rule is often seen as the missing parameter or notation of probability. When data have a high level of noise and a large Bayes error then rules can often be difficult to implement (Duda et al., 2001). This however does not apply to the dataset used for this project and can therefore be disregarded.

Each system has an inference engine which is made up of a problem solving procedure or method. This engine relies on the knowledge collected and stored in the knowledge base. For this project the knowledge base stores the rules which will be identified as the decision rules (see Section 6.5). The problem solving procedure uses IF/THEN statements and chains these together to create larger more complex rules. This is known as forward chaining because the method begins with a set of conditions and aims to make a conclusion (Ralston et al., 2000). Forward chaining is a data-driven process in which the rules are used to generate or infer new knowledge from an initial dataset. This knowledge is inferred by applying the rules to the dataset. Backward chaining on the other hand is goal-directed inference. Whereas forward chaining uses the rules to produce new information, backward chaining uses the rules to answer whether the goal clause is true or not. The purpose of RBR in this project is to decide whether a set of boarding records belong to a return journey or not. This is done by executing each part of the rule to see whether the boarding records are in fact return journeys. If all conditions set in a rule were true then the OD information could be inferred. The methodological approach therefore had to be based on backward chaining.

2.8.2 The Rule Base
The rule base (also known as knowledge base) is the location where the identified rules are stored. In this project the rules are stored in two separate text files; one for single journeys and
one for transfer journeys. The rules will be identified in Section 6.5 where decision tables and expert knowledge is used to define all possible and logical scenarios.

2.8.3 Advantages and Disadvantages of RBR

The following advantages are associated with RBR systems:

- The rules are easily comprehensible;
- The rules can easily be defined;
- New knowledge can easily be incorporated into existing rule bases (Kolodner, 1993);
- Existing knowledge can be easily changed by modifying rules that are already defined in the rule base (Heymans et al., 2005);
- Each rule represents a standalone portion of knowledge or a unit of knowledge (Duda et al., 2001).

Systems using RBR as their method to infer knowledge are often faced with the following problems:

- The number of rules can sometimes grow to a large amount and the rules may lose their effectiveness from a readability point of view. This however is not the case in this project as the 36 rules may change over time and maybe a few more rules will be added. However, it is unlikely that the number of rules will exceed an amount that interferes with the readability advantage;
- Rules set in the rule base are not fundamental or universal rules. This means that sometimes set rules that were true, or at least believed to be true, need to be changed to false. The consequences of this is that some of the ‘facts’ may need to be ‘taken back’. Changes could be necessary if the ‘world’ (the transport network in this case) that is represented by the facts changes to such an extent that the rules do not reflect the changes anymore (Patterson, 1990);
- In larger more complex systems it can become expensive to formulate all rules. Large expert systems of companies could have several thousand rules that describe the business environment which is represented by the rules. This does not apply to this project as the problem to be solved is in a specific contained environment. In other words the defined 36 rules were sufficient to represent the boarding records and the corresponding rules required to infer OD information.
2.9 Use of EFC Data for Decision Making by Urban Bus Operators

An interesting study with regard to EFC data utilisation was carried out by Bagchi (2003a) at the University of Westminster. The study focused on the current data collection practices of transport service providers. This was carried out using surveys and face-to-face interviews of 15 transport service providers including seven Passenger Transport Executives (PTE), two county councils and six bus operators. For confidentiality purposes the results were encoded so that the type of the transport service providers was known but not their names or locations. The following paragraphs will summarise the results which were obtained by this study.

OD survey data was carried out by 67% of the survey participants. The 33% that did not carry out such surveys were exclusively urban bus operators. In fact only one of the six operators carried out OD related surveys. Household surveys were carried out by 60% of the participants. Three bus operators and three PTEs did not carry out household surveys that consisted of panel/non-panel travel diaries. All but one participant received EFC data. Two bus operators mentioned that they use EFC data solely to generate aggregate summary data in a predefined format. As the PTEs only regulate transport but do not run their own services they do not collect EFC data directly. However, they have access to the EFC data on request.

One aspect of the survey focused on the main purpose of EFC data. The result was that the primary use of EFC data for PTEs is related to revenue allocation of pre-paid tickets and revenue allocation for concessionary fare reimbursements. Some bus operators use the data for more detailed analyses. Two operators use the EFC data to monitor service reliability. Five participants (3 operators and 2 PTEs) use the data for fare take up monitoring. Only four participants focus on the measurement of patronage and only one of these four was a bus operator. The use of EFC data as an evaluation method for marketing initiatives is only used by three participants.

Another section of the comprehensive survey confronted the participants with nine different statements which were as follows (Bagchi, 2003a):

1. Having a continuous record of journeys undertaken by individuals using our services and being able to link the records to those individuals;
2. Identifying more of our customers by name and address not just the travel card holders;
3. Recording the origin and destination of all journeys (boarding and alighting points to/from vehicle) undertaken on our buses;
4. Having the same journey information available for a larger and more representative proportion of our customers than is possible to obtain through existing surveys;

5. Having a continuous record of journeys undertaken by identified customers using our services so that we can see how people vary their travel behaviour day to day, week to week, etc.;

6. Being able to identify the pattern of interchange with other buses as part of a customer’s journey from one activity to another (e.g. home to work);

7. Monitor and measure demand at different time periods and at different places (e.g. interchanges);

8. Measure more accurately than presently possible how many new customers have started using our services over a period of time;

9. Measure more accurately than presently possible how many customers have stopped using our services over a period of time.

The statements were then rated with regard to their importance. The following Table 2.2 shows the results with regard to the categories each participant placed the statements in:

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
<th>S9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very important</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Fairly important</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Neither important nor unimportant</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Fairly unimportant</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Not important</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Unfortunately the study did not reveal a breakdown by county council, bus operator and PTEs. Nevertheless, the rating of the statements is positive with regard to the information requirements. In particular the ranking by the participants of statements 3, 6 and 7 are significant for this thesis as the proposed algorithms attempt to provide such information. These three statements are also the ones that are ranked as most important with an average of fairly important to very important. This will be analysed at the end of this thesis (see Section 8.2) in order to address which statements can successfully be addressed using EFC data.

### 2.10 Summary

ADC data utilisation technologies have not advanced significantly over the last decades. Most systems are still independent without integration with other ADC components. The data collection and analyses aspect of ADC systems is still very much in the background. As transport
networks become more complex and mature to multi-modal and multi-operator networks, the availability of data will become more important in order to support the decision making process. The data collection capabilities of ADC systems are already provided; the next step is to fully utilise their potential, providing urban public transport planners with a database that can be used to optimise operational and strategic decisions.

Rule Based Reasoning is a valid method to estimate OD information as the rules are capable of capturing domain expertise. Furthermore, the rule base can easily be extended without changing the algorithm itself. Due to the small rule base the general disadvantages of rule based reasoning are not applicable.

The presentation of the current data collection practices of various public transport stakeholders was mapped to their information needs. This will be followed through and in Chapter 8 the results of this thesis address the information needs again, identifying which ones can be provided using historical EFC data.

The following chapter will outline the literature that is available with regard to Origin/Destination (OD) estimation, the use of OD data for public transport planning and the role of transfer journeys for network optimisation.
Chapter 3

Literature Review

3.1 Introduction

Origin/Destination (OD) data are used in many transport models as one of the main parameters. This chapter will outline existing methods to estimate OD matrices. Initially a historical overview will be described followed by a brief section on OD estimation using count data. Then research efforts based on OD estimation on trip level, on which this project is based, will be outlined in greater detail. These trip level models are critically compared and evaluated at the end of this chapter.

The most common method of generating OD datasets is through surveys which are often expensive and time consuming (DFT-UK, 2004). Identifying possible sources of existing movement data can cause duplication of effort and further increases the cost of transport modelling (DFT-UK, 2001). Consultants are particularly interested in timely sources of existing movement data due to the need of assessing alternative options when developing transport models (DFT-UK, 2001, 2004).

For example, an OD survey was carried out by the New York Metropolitan Transportation Council (NYMTC) in 1990 which cost $1.5 million. The council distributed 2 million forms out of which 450,000 were valid responses. NYMTC also states that there were considerable biases with regard to over reporting from the central business district (CBD) and under reporting of trips made by students. Furthermore, return tips were also under reported. NYMTC has decided not to repeat the survey due to the potential of EFC data. However, corridor-level surveys are still carried out to supplement EFC data. This example shows that automated extraction of destinations based on historical EFC data is a possibility and can eliminate some of the surveys. The will further be outlined throughout this chapter.

Kadiyali (1984) states several needs for OD surveys. This includes the need to determine
the exact origin and destination as well as the zone of trips, time of the journey, and trip purpose. Although the proposed method in this thesis does not allow one to define trip purpose or household characteristics it is possible to infer some of the other parameters such as location of origin and destination and the time aspect of trips.

3.2 Historical Overview of OD Estimation

Prior to more modern approaches the first large-scale cordon count which resulted in a OD matrix was implemented in Chicago in 1916 (Easa, 1993). Between this period and the late 1960s statistical surveys such as travel diaries and interviews, roadside surveys or license plate surveys were mostly used to infer OD matrices (Cremer and H.Keller, 1987). Such methods are now often called conventional analysis (Peng, 1998). Due to the changing characteristics of transportation it became more and more difficult and too expensive to infer such information using some surveys types and interviews (Sun and Porwal, 2000). Furthermore, frequent changes in land-use made inferred matrices often quickly outdated. Gravity, opportunity and Fratar models are often used to infer OD matrices based on conventional analysis methods (see Easa (1993) or Ortuzar and Willumsen (2001)). In more recent years the focus is on models that use traffic counts and sampled OD data (Cremer and H.Keller, 1987). Of particular interest to this project are the recent advances in trip level OD estimation for public transport networks such as Barry et al. (2002); Zhao et al. (2007); Trepanier et al. (2007a).

3.3 Aggregate OD estimation

The area of OD matrix estimation receives considerable attention in the literature due to its fundamental importance in transport modelling and planning. Numerous methods exist of which this section will briefly describe the key concepts. Noteworthy is that the following methods estimate aggregate OD matrices and are not able to infer 'real' trip level OD data per passenger. This is of importance as this research aims to produce trip level OD data meaning that the destination of each individual passenger is attempted to be inferred based on real recordings of that same passenger rather than the aggregate OD volume of pre-defined nodes and their volume counts.

There are many models that estimate OD matrices. The main differences often depend on the frequency of count data measurements, the treatment of congestion, the mathematical formulation and the network configuration. The following section will describe each of these categories.
3.3. AGGREGATE OD ESTIMATION

Static models only consider time independent count data and are therefore more concerned with estimating an OD matrix over a long time interval. Many different approaches exist including entropy based models (Zuylen and Willumsen, 1980; Bell, 1984), least squares (Cascetta, 1984; Ashok and Ben-Akiva, 1993; Bierlaire and Toint, 1995), maximum likelihood (Spiess, 1987), and others. Dynamic OD estimation focuses on the estimation of time dependent link flows over a shorter time interval (Bierlaire, 2002). Dynamic OD estimation can be solved analytically (Cascetta and Cantarella, 1990; Friesz et al., 1993) or by using simulation tools (Mahmassani et al., 1993; Ben-Akiva et al., 2002).

The treatment of congestion can be split into proportional assignment which is frequently used for less or not congested networks as it assumes that link costs and link volumes are independent (Kim, 2006). Equilibrium based models are more suited for congested networks as it assumes that link cost depend on link flows. This also means that it is assumed that network users will always find the shortest path with the least resistance. The equilibrium is reached when no traveller can improve his/her path by choosing a different one. The equilibrium assignment can be solved using mathematical programmes (Merchant and Nemhauser, 1978; Janson, 1991; Kuwahara and Akamatsu, 1997; Jayakrishnan et al., 1995; Lin and Lo, 2000), or it can be seen as an optimal control problem (Friesz et al., 1989; Wie et al., 1994; Lam and Huang, 1995). Another method to solve equilibrium problems is by applying variational inequality as shown by Ran and Boyce (1996), Chen and Hsueh (1998) and Heydecker and Verlander (1999).

When classifying OD estimation approaches into mathematical formulation the following two categories emerge: Traffic modelling approach and statistical inference approach (Kim, 2006). The traffic model approach mostly consists of the minimum information or entropy maximisation model using an older OD matrix that is changed to meet traffic counts (Maher, 1983). The statistical inference approach can be further split into several categories including: maximum likelihood (Spiess, 1987; Watling, 1994; Lo et al., 1996; Hazelton, 2000), Bayesian approach (Maher, 1983; Lo et al., 1996; Hazelton, 2008), Kalman filtering (Bhattacharjee et al., 2001; Lin and Chang, 2007; Zhou and Mahmassani, 2007), and generalised least square (Bell, 1991; Hazelton, 2000; Nie et al., 2005).

Many methods combine some of the approaches listed above to propose a new model. As the proposed research does not use any of the above approaches we ask the reader to refer to the cited material for more information. The following section reviews the existing literature available for trip level OD estimation for public transport passengers.
3.4 Trip Level OD Estimation

Trip level OD estimation for public transport networks based on historical EFC data is a rather new research area. Most other methods (as outlined above) aim to create aggregate OD matrices in which information and their characteristics of individual journeys are not retrievable. However, in particular for public transport networks, such information could be used to calculate performance measures such as in-vehicle time and its variability, waiting time at transfer nodes, passenger miles travelled per route segment, route or network level, etc. Therefore, a method that successfully infers OD information at trip level could potentially be very useful to public transport operators, authorities and consultancies. The following will be an outline of existing algorithms, techniques and methods that focus on the retrieval of OD data at trip level using historical EFC data.

New York City Transit were the first operator who carried out a project that used EFC data to obtain OD matrices for their metro network. The algorithm used two assumptions. The first assumption states that a high percentage of riders return to the destination station of the previous journey to begin the following trip. The second assumption states that a high percentage of passengers end their final trip of the day at the metro station where they started the first trip of the day.

The assumptions were verified by randomly selecting 100 New York residents from the NYMTC survey who made two journeys per day. The sample of 200 journeys was then verified against both assumptions. This analysis showed that the assumptions can be correctly applied to 90% of all trips. The second analysis took a sample of 150 individuals with more than two journeys per day. The sample data also originated from the NYMTC survey. The 150 randomly selected passengers made 595 trips in total. The assumptions were then verified against the 595 trips. This analysis also showed that the assumptions are valid for 90% of all trips. The MetroCard is used as the method of payment for 80% of all subway trips. Three per cent of all validations only record one trip per card and no OD estimation is therefore possible. Barry et al. (2001, 2002) state in their paper that the share of MetroCard data to inferring passengers’ destinations lies at 78%.

After obtaining the trip level destinations results three methods were used to compare the inferred OD pairs to ridership counts. The first method, focused on the comparison between estimated passengers at a destination station and the total exit counts recorded by the EFC system. Barry et al. (2002) conclude that the estimated results correspond closely to the exit
3.4. TRIP LEVEL OD ESTIMATION

Counts. The other two verification methods focused on comparisons of passenger numbers on trains during peak load points and the inferred results. The second method calculates the total passenger counts on each train using the OD pairs. The third method uses a trip assignment and network model with a capacity constraint shortest path routine in TransCAD. The results were again compared to the totals of the inferred OD pairs. The results of this approach and the inferred OD estimation matched to 97.5%.

This method seem to have worked well for the metro network in New York. However, it cannot be assumed that the passenger always returns to the destination where he/she alighted to board the next journey. It will be shown later in this thesis that this method of trip chaining would not have worked very well for the Dublin Bus network. Although using the method would create more OD pairs, the precision and correctness of results is not exact enough as too many false negatives would be inferred. Furthermore, the assumption that the boarding in the morning was the destination in the evening may not be valid for other networks. This thesis will elaborate on this issue when investigating the validity of the assumption for the Dublin Bus network (see Section 6.2.2).

A similar approach was used for a project that was carried out by the Chicago Transit Authority (CTA) in collaboration with the Massachusetts Institute of Technology (MIT). Again, the aim of the project was to infer trip level OD information for Chicago's rail network using EFC data. The main assumptions of this project were the same as stated by Barry et al. (2002). However, due to the less complex rail network the aim was to improve the algorithm with mode choice assumptions. The EFC system of the CTA records the passengers' entry to the nearest second whereas the New York City Transit Authority's (NYCTA) EFC system only records validations to the nearest 6 minutes. CTA's EFC system does not have many stations where direction-separated fare collection is available. The CTA network has seven rail lines and a total of 145 rail stations with 178 entry points. The EFC system records approximately 525,000 weekday boardings.

Three existing datasets were used to verify the results of the OD estimation algorithm: traffic checks of passenger volumes at peak load points, the annual OD survey known as Cross Platform Transfer Survey of over 10,000 records and ride-check samples. None of these data sources have been specifically collected to verify the proposed algorithm and all of them are comparisons on an aggregate level. The initial verification method compared the inferred OD results with passenger counts. The comparison resulted in accurate results. The second verification method compared manually recorded load profiles with the estimated load profile and
again, the inferred information matches approximately the other dataset. When analysing the correctness of direction of travel the algorithm assigned 24% of all trips incorrectly. As not all stations lack directional attributes this only affected 96 out of a total of 1,629 station entries.

Rahbee and Czerwinski (2002) also address the issue of records for which the inference of a destination was not possible. Various matrix balancing techniques or optimisation techniques are described in Spiess (1990); Furth and Navick (1992); Kikuchi and Perincherry (1992).

It is suggested that EFC data from urban bus services need to be integrated to refine rail destination estimates. This is mainly due to rail-bus transfers as the location of alighting can only be inferred when the boarding location of the bus service is known.

The most recent contribution to Rahbee and Czerwinski (2002) was carried out at MIT in collaboration with CTA (Zhao, 2004; Zhao et al., 2007). This approach used the assumptions that were already used by Barry et al. (2002); Rahbee and Czerwinski (2002); Rahbee (2003). The contribution of this research project was in the integration of bus EFC and AVL data to refine the location of rail destinations. However, it did not estimate any destinations of bus passengers.

The EFC system on CTA's fleet is an entry only system. A flat fare is charged to each passenger regardless of how far the passenger travels. A record is stored for each validation of the fare card including a time stamp and an identifier whether this journey is part of a transfer journey or not. The bus ID is stored indirectly. The location information of each boarding is only represented by the route identifier. This is different to the Wayfarer system which is installed on the Dublin Bus fleet where the bus stage is also recorded.

CTA's AVL system is part of the Automated Voice Annunciation System (AVAS), which is a requirement set by the Americans Disability Act (ADA). AVAS provides information about the route and announces the next stop using AVL technology. The following data attributes are stored for each event: location with timestamp; the location details; bus route number; bus number; and status information (i.e. door is open/closed). The location details contain longitude and latitude, run pattern ID and stop sequence numbers. The bus stop is not recorded directly but via a stop sequence. Zhao (2004) uses a migration technique whereby bus EFC and AVL data are combined. The timestamp of both datasets, the route ID and the bus ID are used to extract the bus stop ID where the passenger actually boarded.

Zhao (2004) introduces the following assumptions to infer OD information:
3.4. TRIP LEVEL OD ESTIMATION

- Public transport passengers do not use private transportation (car, motorbike, bicycle, etc.) between trip segments;
- No passenger walks a long distance in order to board at a different station to the one he/she alighted. The typical accepted walking distance is 1,320 feet or five minutes walk at a walking speed of 3 miles per hour;
- The last trip of the day ends where the first trip of the day was recorded.

The study then analyses sequences of trip segments with regard to the mode used. Only train to bus and train to train sequences are analysed as it is the main aim of the study to infer the destination of rail passengers. Therefore, if it is a rail to bus trip sequence and bus EFC, GIS and AVL data are available it is possible to estimate the destination of the rail station where the passenger alighted. Where only GIS or AVL is available the estimation of the final destination can often not be exactly determined.

The study analysed boarding records over a one week period which included a total of 5,948,454 transaction records. Two and more records per day were recorded for 83.6%. The modal split of these boardings was 52.3% rail and 47.7% bus. Throughout the week, just over 1 million unique farecards were validated resulting in an average of 5.5 boardings per card. The main rail to rail algorithm, train to bus algorithm, trip chain symmetry algorithm and multiple swipe cases inferred destination of 50.1%, 11%, 2.4%, and 1.4% respectively. The algorithm infers destinations for a total of 65.5% of all rail boardings.

A recent publication by Trepanier et al. (2007a) showed a model that uses Transportation Object Oriented Modelling (TOOM), classifying data into user, trip, route and route-stop objects. This model compares boarding record details with vanishing stops and routes of the remaining records. It further uses geographic coordinates to infer the location of alightings as it is not likely that passengers travel long distances without using public transport (Trepanier et al., 2007a). If any of the possible (vanishing) alighting stops are within 2km distance of the following boarding then a location can be assigned. Trepanier et al. (2007a) further revisits boardings for which no destination could be found and compares these to existing inferred destinations aiming to find a similar record. The assumption that the destination of the last journey of the day is the location of boarding of the following day is extended by also analysing the previous day's boarding location. A total of 66% of records could be inferred with a destination. However, 20% were inferred using the 'last journey - first journey' assumption which may not suit all networks as outlined later in this thesis.
Navick and Furth (2002) say that relevant measurements such as passenger loads, passenger miles and OD patterns require at least an estimate of the location where the passenger alighted. The paper by Navick and Furth (2002) analyses the potential of farebox data to carry out more advanced data analyses. The authors consider farebox data as a valuable source of information on passenger travel patterns. Many relevant measures of interest require the number (or at least an estimate) of passenger’s alightings by stop. The assumption is made that the pattern of passenger alightings in one direction mirrors the daily boardings pattern in the opposite direction. The study has been carried out using a full day’s set of on-off counts on five Los Angeles area routes. In order to generate the number of passenger alightings it requires labour intensive ride checks or APC’s. However, both methods are very expensive and are generally not carried out for the entire network (Furth, 2000). Navick and Furth (2002) developed an algorithm to generate location data of bus passengers using EFC data. The authors assume that the transactional data records are equipped with location stamps. These location stamps provide information with regard to what stage the boarding took place. This data is then fed through the algorithm and results in an empirical distribution. Origin-Destination matrices can be useful for analysing routing and scheduling options such as short-tuning, expressing, and route splitting. Navick and Furth (2002) introduce a propensity function that incorporates behavioural characteristics of urban bus travel. After obtaining the alighting patterns it is possible to generate other valuable travel pattern measures including passenger miles, load profiles, and origin-destination matrices.

The model applied a bi-proportional method to the initial or seed matrix (also known as iterative proportional fit - see Ortuzar and Willumsen (2001)). The aim was to match the row and column totals by repeatedly scale their values. This method, which is further described in Ben-Akiva et al. (1985), is similar to a doubly constrained gravity model for trip distribution.

A propensity function is introduced by Navick and Furth (2002), customised to the urban bus behavioural characteristics. The new propensity function is based on Navick (1997), which was originally a distance-based propensity function. Short trips have the tendency to be made on foot and the authors decided that travel propensity increases with the distance $d_{ij}$ between origin $i$ and destination $j$ of the trip $u$. The new propensity function is the product of two exponential functions:

$$s_{ij} = [1 - \exp(-\alpha * d_{ij})] * \exp(-\beta * d_{ij})$$

(3.4.1)

where $\alpha$ is represents the resistance to walking longer distance and $\beta$ resembles the resistance
3.5. LITERATURE OF TRANSFER JOURNEYS

Origins and destinations of passenger trips generally become more dispersed (Allen et al., 2003). This led to more complex networks including several different modes which makes it more difficult for public transport customers to travel from origin to destination. Improving
to riding the bus for longer distances.

The tests indicate that estimates of historic OD data may successfully replace manual or APC counts. Again, this method does not allow one to identify the destination of individual passengers but focuses more on aggregate OD matrices over a period of one day or one week.

The OD estimation results of the New York city project were used for several other project including schedule design of summer schedule planning during reconstruction projects. It was further used to forecast weekday morning peak travel demand, input data to calibrate models, supplement census journey-to-work data, and station-to-station MetroCard trip table in combination with the transit trip assignment package in TransCAD (Barry et al., 2002). Zhao (2004) show that the results of the study can be applied to a path choice model for parts of the rail network. Trepanier et al. (2007a) mentions that trip purpose could be derived from users' travel habits. There is however no evidence that this was further analysed.

Barry et al. (2002) consider travel behaviour analysis over several days as a possibility to enhance the inference algorithm. They further said that the extension of the model could also includes bus services data to infer journeys that were carried out using several public transport modes. Zhao et al. (2007) suggested that EFC data from urban bus services need to be more fully integrated to refine rail destination estimates. A study by Bryan (2007a,b) also outlines the importance and usefulness of historical EFC data and its potential to infer OD data at trip level. Various applications of smartcard data are outlined in Bryan (2007b) including detection of corruption, turnover rate, and interchange information.

This section outlined the existing literature that indicates the need for a model that focuses on the extraction of OD information at trip level for public transport networks. The existing models are all optimistic models where the assumptions are broad and, although the recall rate might be good, the precision would not be very high when applying these to the Dublin Bus EFC data. This is mainly due to the different characteristics of the public transport network. The proposed model in this thesis takes a rather pessimistic approach and only assigns a destination based on more reasonable assumptions and rules. This will be further outlined in Chapter 6.

3.5 Literature of Transfer Journeys

Origins and destinations of passenger trips generally become more dispersed (Allen et al., 2003). This led to more complex networks including several different modes which makes it more difficult for public transport customers to travel from origin to destination. Improving
the transfer process can therefore contribute to attracting new and retaining current passengers (Allen et al., 2003).

Transfer journeys (also known as linked trips) are important factors in public transport assignment as some passengers prefer to minimise the number of transfers while others favour the minimisation of travel time (Nielsen, 2000). Synchronising schedules may reduce transfer times of public transport networks (Chowdhury and Chien, 2001). However, effective synchronisation of schedules requires information of transfer journeys. This may include the main transfer nodes, the volume at these nodes and the temporal distribution of these volumes.

Transfer time is one of the most important levels of service indicators to evaluate intermodal public transport networks (TCQSM, 1999; Chowdhury and Chien, 2001). Although increased frequency would decrease travel time, it is often not cost effective due to variation of demand over space and time. Schedule optimisation is therefore an alternative to improve transfer services and ultimately customer satisfaction.

EFC data can be used to identify such information about transfer journeys. A study carried out at Westminster University focused on the utilisation of EFC data to improve decision making. The study (Bagchi and White, 2003) focused on the improvement of understanding travel behaviour using EFC data. The aim was to analyse the travel behaviour of elderly concessionary travellers. The cities Bradford and Southport, both in the UK, were analysed. It was found that in Bradford 7.6% of 8,086 boardings and in Southport 3.5% of 90,062 were transfer journeys. The paper further states that 98% of all identified transfer journeys consisted of two boardings (one transfer). The remainder consisted of three or more boardings. This research was based on Bagchi (2003a) which primarily focused on the use of smart card data for travel behaviour analysis and its implications for marketing. Inferring bus-to-bus interchange was achieved using a rule-based process which is similar to the approach used in this thesis. The users were identified using unique identifiers that were stored on each smart card. The transfer time limit was set to 30 minutes meaning that the second journey boarding time had to be recorded 30 minutes or less after the first boarding. This analysis resulted in 24% linked trips when analysing period travel cards.

Similar research with regard to transfer journey analysis for operators’ marketing, urban city planning and passenger travel behaviour was carried out by Okamura et al. (2003). This study included a section that focused on the extraction of transfer passenger volumes and waiting times at four major transfer points. These four points make up over 80% of all transfer
passengers in the city. The ratio of transfer journeys over total trips is approximately 20% at these four nodes. Figure 3.1 shows the calculated waiting time based on the EFC records. The exact method how this information is obtained is not described in the paper.

Figure 3.1: Distribution of Waiting Time during Peak and Off-peak time (Okamura et al., 2003)

The paper also uses decision trees to classify passenger transfer behaviour with regard to time of travel. This way it was possible to identify which times most passengers transferred at each of the major transfer nodes (Okamura et al., 2004).

3.6 Summary and Implications of this Research

Origin/Destination (OD) data is used in many transport models as one of the main parameters. The most common method of generating OD datasets is through surveys which are often expensive and time consuming (DFT-UK, 2004) and count data. Identifying possible sources of existing movement data can cause duplication of effort and further increases the cost of transport modelling (DFT-UK, 2001). The concept of inferring OD matrices by analysing EFC data was first introduced by Barry et al. (2001). This research effort was then continued by a project carried out in Chicago. Both projects aimed on the estimation of OD matrices using a set of assumptions that needed to be applied to the metro EFC data.

The projects carried out in New York and Chicago indicate the need for OD matrices and performance measures that improve the decision making process. Both projects identified the need to define a model that can be applied to infer OD data from historical EFC data originating from an urban bus network. Trepanier et al. (2007a) proposes a model that that focuses on the estimation of trip level OD data from an urban bus network. The model produces promising results.

There is a difference between OD surveys/estimation that provide data for traffic modelling
and those that target public transport optimisation. For example, it is much more important to have exact time values for public transport optimisation. Furthermore, when generating performance measures such as on-time performance, in-vehicle time, frequency, bunching, etc. it is critical to base the analyses on factual data. In short, surveys can be used to obtain more aggregate information about passenger demands, satisfaction and needs (i.e. trip distribution and assignment models (Kadiyali, 1984)). For analyses that require exact information with regard to the operational services of a provider surveys are not necessarily most suited although operators often still rely on surveys as a data source.

The following chapter focuses on the migration of the EFC data. Furthermore various extensions to the dataset are introduced including a method to identify transfer journeys. A brief section focuses on the quality of the recorded data.
Chapter 4

Database Design, Data Storage and Data Extension of EFC Data

4.1 Introduction

Due to several million boardings per day or per week (depending on the transport network) the storage of EFC data can become a task that requires careful consideration. The technology used for facilitating such data storage is vital to the success of any project that is based upon EFC data. A technology is needed that supports data validation, reduction of data redundancy, ensures data integrity, updateability, extendibility and facilitates an interface for other applications to update, insert and retrieve data from and to the database.

This chapter will outline how over 160 million records were stored and structured to allow for further querying and analysis of the data. A well established and mature technology called relational databases was used to provide an infrastructure for the data storage. It further outlines what steps were involved in order to migrate the data that were initially stored in semi-structured and semi-encrypted format into the relational database structure. A 4-Phase framework was developed to ensure data consistency, data integrity and data validation. Such a framework can be applied by other parties who wish to transform and store their EFC data in a relational databases management system.

It further describes what data are available and what steps were necessary to extend the initial EFC data so that further analysis was possible. This includes the integration of two spatial identifiers so that each boarding location was not just an encoded number but a location with semantics. One aggregated identifier divided Dublin and the Greater Dublin Area (GDA) into 21 different zones (known as the DTO coarse zones) whereas the other divided Dublin into 131 urban and suburban areas.

It was also necessary to identify which boarding records were part of a transfer journey
(also known as linked trip) as the initial dataset did not contain such information consisting only of unlinked boarding records. An iterative classification algorithm is introduced that uses a set of assumptions which then conclude whether boarding records were single journeys or belonged to a transfer journey. Various validation methods including a comprehensive Monte Carlo simulation are also described.

The final section of this chapter focuses on how ‘good’ the obtained Dublin Bus EFC data are. There were initial concerns about the data as its recording is not fully automated. The interaction of bus drivers is required to identify the correct location of the bus. A number of analyses were carried out to show the strengths and weaknesses of the dataset.

4.2 General Background of Relational Databases

It was a paper in 1970 written by an IBM employee who provided a solution to data storage for the future. E.F. Codd introduced in his paper called ‘A Relational Model of Data for Large Shared Data Banks’ (Codd, 1970) a new and improved method to store data. Hierarchical databases and network databases were the predecessors. However both these systems had major problems with regard to their flexibility, data integrity and redundancy. Codd, a mathematician and chemist, based his model mainly on mathematical theory known as relational algebra. The main concept of relational databases is a two-dimensional table (known as relation) containing rows (also known as tuples or records) and columns (also known as attributes or fields). But in reality there are many such tables in which some column(s) of one table contain a reference (relationship) to another column in a different table. For example, one table could contain all details of bus routes whereas another table could contain data of individual passengers boarding a bus service. The referencing columns could be the bus route ID’s of each table.

Relational databases are used in everyday life. Whether products are scanned at the cash register or when withdrawing money at an automatic teller machine, relational databases are always involved to support the transaction. Well known relational databases include, among many others, Oracle, MS Access, MySQL, DB2 and Informix. These systems are all known as Relational Database Management Systems (RDBMSs). More fundamental information about relational databases can be found in Date (2003) and Ritchie (2002).

4.2.1 Why Relational Databases?

Relational databases are the de facto standard technology for storing large data sets. Hierarchical, network based or file-system databases are legacy technologies that could not cope with
the amount of data or the required flexibility. The two-dimensional tables that build the objects
contained in a RDBMS create an environment to store data that are suited to the data structure
which is present in EFC data. Relational databases have matured over the last 37 years. Initial
problems such as cost of multiple joins have in more recent RDBMS version been improved
tremendously (D'Souza and Wills, 1998). The access language is called Structured Query Lan­
guage (SQL) which is an ISO standard (ISO / IEC 9075-1, 2003), easy to apply, and can be
embedded in most advanced programming languages. Additions to SQL such as Procedural
SQL (PL-SQL) allow procedural functions to be applied to the database which facilitate a more
complex structure compared to the declarative query structure of SQL. Nowadays, information
is treated as an asset. It is therefore important to store all data and information that are produced
by the company’s transactions and work processes. Currently, relational databases can easily
handle data sets that are as big as several terabytes (1 Terabyte = 1,000 Gigabyte) (D’Souza and
Wills, 1998). Depending on the structure of the database it is possible to retrieve specific or
aggregated information almost instantly from the RDBMS (Date, 2003).

Object Oriented Database Management Systems (OODBMS) were the only database mod­
els that were further considered when deciding on the storage technology for this project. How­
ever OODBMS are often considered to be more revolutionary than evolutionary (Cattell and
Skeen, 1992). Since its introduction OODBMS are mostly used in areas where navigation
through a complex data structure is required. For example, Computer Aided Design, GIS or
software engineering related projects. They lack the maturity that the RDBMS has gained over
the years and generally do not fully support declarative query access. The Object Data Manage­
ment Group (ODMG) standard is an attempt to remedy this disadvantage (D’Souza and Wills,
1998; Devarakonda, 2001). The current version of the standard is ODMG 2.0.

The immaturity of the OODBMS and the fact that there was no drawback in using relational
databases for this project led to the decision that an RDBMS was considered as being more
suited to store the EFC data of this project and was, in addition, freely available for research
purposes.

The following section will outline how these data files were migrated into a relational
database structure.
4.3 Data Migration Procedure – A Multi-Step Framework

This section describes the procedure that has been developed to import the EFC data obtained from Dublin Bus into a relational database. The multi-step import framework describes a 4-phase procedure (see Figure 4.1) to transfer data stored in flat files (text files with no definite structure) into the relational database. This data migration was necessary to facilitate all future analyses. The primary purpose of the final version of the database is to create a platform from where information can be extracted from the underlying data source. This information can then be used by decision makers and is called ‘actionable information’ (Yoder et al., 2002). Each of the four phases uses different technologies and/or tools to produce the deliverable for the next phase.

As shown in Figure 4.2 the data as they are received from each individual bus are in semi-structured form with regard to time, route and passenger specific details.

Pre-formatting of the Wayfarer data files is therefore necessary to retrieve the information of the record types and data attributes. The goal is to import the data into a relational database so that querying, extracting and further formatting can be carried out with ease. Since relational databases can be manipulated and accessed with SQL and PL/SQL it was decided to format the data in such a way that the results can be uploaded directly into the predefined Oracle database allowing restructuring of the data within the database. The following phases which were outlined in Figure 4.1 will be introduced in greater detail:

- Phase 1 – Pre-formatting the raw data;
4.3. DATA MIGRATION PROCEDURE – A MULTI-STEP FRAMEWORK

810D0B0K010704000222 --> Control Record
82148505040000000000 --> Duty Record
8311 101310030105100 --> Journey Record
85006610030000000000 --> Stage Record
8B264100020000000000 --> Validation Record
85006710160000000000 --> Stage Record
85006810180000000000 --> Stage Record
OA7C0F5D000000000000 --> Boarding Record
85006910200000000000 --> Stage Record
OA7C0F62000000000000 --> Boarding Record

Figure 4.2: Raw Data File Records

• Phase 2 – Importing data with SQL Loader;
• Phase 3 – Initial data model population with PL/SQL;
• Phase 4 – Cleaning and extension process.

4.3.1 PHASE 1 – Pre-Formatting the Raw Data

The initial task has been defined as pre-formatting the data in such a way that it can be uploaded into a single table stored within the RDBMS. This table, called ‘Rawdata’, consists of the following four attributes including their data type and maximum length in brackets (e.g., the data type of the ID is Number with a maximum length of 10 digits):

• ID (Number, 10) – This is an identifier which was created to uniquely identify each record within the record type. This parameter is entirely for internal use;
• String (Varchar, 128) – This string stores the actual information as created by the EFC system. Some strings needed to be decrypted by using a bit-shifting technique to extract the captured information;
• Date (Date) – This defines the date when the record was produced by the EFC system;
• Row Number (Number, 10) – The row number is a further internal parameter so that each row has a unique identifier (primary key).

Wayfarer used a bit shifting technique to store more data in the 20-character hexadecimal string. A decoding procedure was developed to deal with overlapping bits into previous or next bytes. The data migration algorithm also created summary records for each imported file.

The C++ program carried out these steps to pre-format the data in such a way that it complied with the requirements set by the pre-defined database structure and an Oracle tool called
SQL Loader, which will be described in greater detail in Phase 2 of the framework. Various error detecting parameters have been introduced in order to determine the quality of the data and to detect any abnormalities.

4.3.2 PHASE 2 – Importing Data with SQL Loader

The aim of this phase is to import the pre-formatted data produced in Phase 1 into a pre-defined database table. SQL Loader version 9.2.0.1.0 is the Oracle utility program used to upload data from flat files (.dat files) into a predefined database structure (Rich, 2002). The tool transferred nearly 20,000 records per second from the text files into the database table, which stored 160 million records in a table called 'RAWDATA'. Before importing the data were stored in 730 different files.

4.3.3 PHASE 3 – Initial Data Model Population using PL/SQL

Phase 3 details the restructuring of the database and transferring the records into the pre-defined database structure. PL/SQL code is used to identify the record type and correctly transferring the record values to the correct table. Important is that primary key and foreign key referencing is adhered for all records created so that joining tables will result in correct datasets.

After finalising Phase 3 of the 4-Phase framework all data are stored in the appropriate tables which are at least in second normal form.

4.3.4 PHASE 4 – Cleaning and Extension Process

The purpose of this stage is to clean the data and to extend tables by attributes that have been considered as useful for future analysis. This last phase will also ensure that the database is in its most optimised form and complies with all relational database rules. This is mainly necessary to improve performance and results of future analyses (Fayyad et al., 1996).

4.3.5 Updated Database Model

Figure 4.3 shows an Entity Relationship Diagram (ERD) for the updated database. An ERD is a data modelling technique that creates a graphical representation of the database. The diagram shows the entities and the relationships between entities.

The three unconnected tables at the bottom of Figure 4.3 store information that is not directly part of the project database but provide aggregate information about the data or the importing process.
4.3.6 Data Warehousing

A data warehouse is a collection of computerised data which is organised in such a manner that analysis and reporting are optimised. One could argue that the created database is a data warehouse when the data are rearranged. However at this stage no formal data warehouse schema is applied to the relational database.

4.3.7 Summary of the Migration Process

The original files (one per day) have been imported into the project database using a 4-Phase framework (see Figure 4.1). The result of this framework is a relational database storing EFC data of a two year period (1998 and 1999)(see Figure 4.3).

The ERD diagram shown in Figure 4.3 graphically represents the structure of the project database. This is the database that facilitates all analyses carried out for the project. One of the main advantages of relational databases is that it is easy to extend and link other databases to the existing database. This is described in Section 4.4 where bus stop information is added and linked to the existing data.
A more detailed description of this section was published at the World Congress on Intelligent Transport Systems (Hofmann et al., 2003).

The entire process of developing the framework and migrating the data took approximately 3-4 months. However, once familiar with the data and the algorithms used to migrate and format the data it is straightforward to upload the data into a database. It is also straightforward to add data to existing datasets as the structure of the database is already present. The system allowed 50 million records to be pre-formatted and migrated in less than 12 hours (depending on the database and the IT equipment).

4.4 Populating the Database with Information about Bus Stages

The EFC data obtained from Dublin Bus do not incorporate or provide geographic coordinates or any similar spatial identifier. For many analyses spatial information is needed to understand and interpret the obtained results. The Dublin Bus dataset contains data such as route number, bus stage information and direction of travel. The bus stage is recorded as two digit number (from 0 – 99) whereas the direction attribute is of Boolean format that can consist of 0 or 1 referring to outbound or inbound direction respectively. A bus stage may contain a number of bus stops depending on the length of the bus route as the maximum number of bus stops per route is 99. The terms bus stop and bus stage are used interchangeably within this thesis unless otherwise stated. However, the more bus stops are assigned to one bus stage the less is the accuracy of the results. The three attributes (route, stage number and direction) are the unique identifier for each location and were therefore used to enhance the database with spatial information.

A complete list of locations of all bus stages was not available and it therefore had to be generated manually. The database consisted of over 8,200 different route/stage/direction combinations. The data required for each bus stage included the bus route indicator, the stage number, the route direction, the stage description, the coarse zone and a suburban area description. The coarse zone is a code that has been introduced by the Dublin Transport Office (DTO) and divides the GDA into 21 different zones (see Appendices A and B). The size of the coarse zones differs as indicated on the map. Most bus services travel through at least 2 zones. The area description consists of 131 different city centre and suburban area names.
4.5 Extension of the Project Database – Transfer Journey Identifier

Due to various differences between single journeys and transfer journeys (also known as linked trips) it is necessary to identify the two different types of journeys so that they can be incorporated into future analyses. The EFC system does not record or link passenger boardings which leads to a gap in knowledge when analysing the records stored in the database. This section focuses on the identification of transfer journeys.

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. 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. The identification of transfer journeys is also important to determine the true amount of travel by mode (Hensher, 1999). In particular in true multi-modal networks the identification of transfer journeys is vital for the optimisation and interpretation of the public transport network. Without the use of a GIS it is not possible to determine whether the trip was part of a forced or voluntary travel choice.

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 (as shown in Bagchi and White (2003, 2004)) and Hofmann and O’Mahony (2004). This section describes an iterative classification algorithm that classifies individual passenger boardings into two categories; single journey and transfer journey.

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 (DETR, 2000). For the purpose of this project 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 Figure 4.4 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 node. A1 is the journey origin and B2 is the journey destination.

Figure 4.4: Transfer Journey Definition

One of the main aims of transfer journey analysis is to identify whether the bus services are meeting the needs of the passengers (Bagchi and White, 2003). 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) (TCQSM, 1999) can be partially created by extracting information from an EFC database. Chapter 5 describes some of the analyses that are possible after identifying all transfer journeys. These analyses may contribute to decision and policymaking.

The enriched dataset facilitates analysis of the data stored from the first boarding (A1), data from the second boarding (B1) and from the two boardings together. Data items of the two alightings (A2 and B2) can only be estimated to a certain degree because passengers are not required to insert their ticket when leaving the vehicle (exit validation). Figure 4.5 shows what data attributes are available for this analysis. As indicated in Figure 4.5 there is no direct information of the alightings A2 and B2 available apart from the data that applies to the entire transfer journey. However, the following assumptions with regard to A2 and B2 can be made:

- The location of alighting (A2) from journey A is relatively close to the location of boarding of journey B at B1. This means that A2 and B1 are more than likely in the same coarse zone and within walking distance;
- The time of alighting (A2) from journey A is likely to be close to the boarding time of journey B at B1;
- The alighting of journey B is in the direction (i.e. outbound or inbound) as indicated in ‘B1 – Direction of Travel’ and on route as indicated in ‘B1 – Boarding Route’.

Information that is specific to the individual passenger such as ticket type, ticket id, ticket name and validity period of ticket remains the same throughout the transfer journey. This also
4.5. EXTENSION OF THE PROJECT DATABASE - TRANSFER JOURNEY IDENTIFIER

Transfer Journey with one transfer at A2/B1

Transfer Stages

A1 - Boarding Route
A1 - Boarding Stage
A1 - Direction of Travel
A1 - Exact Time of Boarding
A1 - Coarse Zone of Boarding
A1 - Area description of Boarding

Estimated area of alighting
Estimated time of alighting
Coarse zone of alighting

B1 - Boarding Route
B1 - Boarding Stage
B1 - Direction of Travel
B1 - Exact Time of Boarding
B1 - Coarse Zone of Boarding
B1 - Area description of Boarding

Estimated journey destination direction of journey

Figure 4.5: Transfer Journey Data Diagram

applies to the date of the transfer journey as transfer journeys that include two different dates have been ignored by the iterative classification algorithm which will be further explained in Section 4.5.2.

Extending a database with additional data is often necessary in order to extrapolate some of the information or knowledge that is hidden in the dataset. The original EFC data did not reveal whether the passenger boarding was part of a transfer journey. However, this information is hidden in the dataset and has to be extracted by applying the iterative classification algorithm that is described in the following section.

4.5.1 Alternative Solutions to Identify Transfer Journeys

Various methods exist to identify transfer journey details. The following section will explore some research efforts that focus on this problem. Also refer to Section 3.5.

One possibility to identify transfer journeys is through public transport surveys that include
questions about the passenger’s trip. Travel diaries often include start time of the entire journey (e.g. National Travel Survey (NTS) UK) and intermediate modes used, although not necessarily timing of interchanges (London Area Transport Survey (LATS)). The NTS, for example, further collects data about boardings and stages. However, the ratios between volumes of different data collection methods are considerably different. For example, in London the data collected from operators and the NTS differed between 15% and 65% (White, 2006).

Survey methods are however difficult to carry out on a network wide level due to general time and financial restrictions. Furthermore the survey can only be carried out over a short period of time leading to problems when transfer trends or changes need to be identified (White, 2006). An analysis on a trip level is often not possible and all analyses have to be carried out on route level. In summary, for those type of transfer journey identification the project would lead to high marginal costs, small sample sizes and are limited spatially and temporally.

Another method to identify transfer journeys would be to base the analysis on census data. No literature was found that used this approach. This analysis could only be carried out on an aggregate level as the census does not record what route a passenger takes and also does not account for journeys that were made as part of a leisure activity. However, it would be possible to assign routes based on the home address and the work address to identify whether the passenger needed to transfer to get to his/her final destination. The census is however only carried out every couple of years and follow-up studies on a network wide level would realistically be impossible. Furthermore, the analysis could only identify the transfer journey demand by area but not by route or even trip level.

Household travel surveys probably offer the most information but can only be implemented on a small sample of passengers mainly due to marginal cost constraints. Using these type of surveys it is then possible to analyse trip level behaviour of transfer journeys.

Using EFC data as the source for transfer journey analysis certainly offers the cheapest method of continuous observation. The question is whether the linking algorithm and ultimately the results can be validated. The only method to validate the algorithm in its entirety would be to have a dataset that stores all the EFC attribute and also holds passenger information about the trips. This is however not possible for this study as such information is unobtainable. Many research efforts focused on the use of EFC data to obtain linked trips that were carried out on a network. Unfortunately, none of these studies focus on the validation of the method. This project will look at some validation methods later on in this chapter. The main difference of this
study to most others is that the aim is not to generate aggregate results but to actually identify transfer journey by passenger and their individual trips. It is then possible to derive aggregate statistical or other measures but it is also possible to analyse passenger transfer patterns on a micro level. Furthermore the transfer journey identification will be a cornerstone for inferring Origin/Destination data which will be introduced in Chapter 6. The proposed algorithm therefore extracts data that will facilitate trip level analyses as well as route or network level analysis of transfer journeys.

4.5.2 Definition of the Iterative Classification Algorithm

The original project dataset includes data that have 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 (at least two boardings). In order to generate transfer passenger analysis, it has to be inferred which individual passenger boardings were linked together. The solution to this problem 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 following section describes the iterative classification algorithm in greater detail.

The concept of the proposed algorithm is to solve a complex problem by dividing it into simpler smaller problems. The solutions of the sub-problems can then be combined to solve the 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 (as shown in Ishwar and Yoo (1997)).

The algorithm is based on the selection and comparison of individual passenger boardings and verifies whether certain data attributes match or variables apply. The fourth generation programming language C++ was used to implement the algorithm on the dataset. 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 SQL statements, which feed the newly extracted information back into the relational database thus facilitating more comprehensive transfer journey analyses.

The format of the extracted dataset of all passenger boardings is shown in Figure 4.1. The semantics of each extracted record is as follows (examples in brackets are taken from the first record): Unique Record ID (e.g. 0017144751), Ticket ID (e.g. 011378) – Ticket Type ID (e.g. 0662) – Boarding Time (e.g. 1358) – Date (e.g. 04251999), Route ID (e.g. 19A) and Internal Number (e.g. 0059503460).
Algorithm Components

Definition 1: As defined previously a transfer journey consists of two different individual passenger boardings that were recorded following these rules:

- 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 individual passenger boarding records;
- Both passenger boardings had to be recorded on the same day;
- Second passenger boarding occurred less than 90 minutes after first ticket validation;
- Route ID had to be different for the two individual passenger boardings.

Let \( g_i \) be the ticket ID and \( h_i \) the ticket type ID of journey record \( i \), where the records are ordered by ticket type ID, ticket ID and then time \( t_i \). The route is defined as \( r_i \). We define records \( i \) and \( i + 1 \) to be a transfer journey iff \( g_i = g_{i+1}, h_i = h_{i+1}, t_{i+1} - t \leq m \) (where \( m \) is a variable expressed in minutes) and \( r_i \neq r_{i+1} \). Note: because of the ordering in the labels \( i \) of the records, any potential transfer journey will be stored in adjacent records \( i \) and \( i + 1 \).

Setting the time limit could be seen as arbitrary. Another study carried out by Westminster University applied a similar algorithm with a considerable shorter period of time. The project which was under supervision of Prof. P. White identified 30 minutes to be, under the particular circumstances, the appropriate constraint focusing on the analysis of trips carried out by elderly passengers. The paper by Bagchi and White (2004) clearly outlines that this 30 minute constraint has been chosen given the shorter distances travelled by elderly passengers and further states that this might be increased for other user groups or networks. It is unclear whether potential transfer journey pairs were eliminated when the same bus route was used for both boarding records. This additional rule would account for and eliminate false identification of journeys where passengers travelled to their final destination with their first boarding record and return to their origin with the second boarding record. Furthermore, the study took place in Bradford, UK and Southport, UK which are considerably smaller than Dublin and therefore
4.5. EXTENSION OF THE PROJECT DATABASE – TRANSFER JOURNEY IDENTIFIER

Total travel time is less. Another project which focused on the city of Hiroshima, Japan used a transfer constraint of 60 minutes (Okamura et al., 2003). This is equivalent of the life span of a single ticket meaning that once a ticket was validated (first boarding) it was valid for 60 minutes.

Dublin Bus shows estimated journey length per route on their website. Table 4.2 shows some descriptive statistics of these journey lengths. In total there were 142 routes of which the estimated journey length was published on the website. These journey lengths reach from 20 to 110 minutes with an arithmetic mean of 61.14 minutes and a standard deviation of 18.28 minutes. It is noteworthy to mention that the extracted times are total journey length of the services and only a few passengers will actually ride on the bus for the entire distance. As the estimated journey times are from 2006 it can probably be argued that the traffic in Dublin has increased slightly and that the average journey lengths were slightly lower in 1999. On the other side there were no Quality Bus Corridors (QBC) in place which in turn would result in longer journey times. The quartiles lie at 50, 60, 76.25 minutes for the 25, 50 and 75 percentiles respectively, meaning that 50 percent of all total journey lengths take 60 minutes or longer. This is another indicator that the 90 minute constraint can be justified.

Table 4.2: Descriptive Statistics of Total Route Journey Time

<table>
<thead>
<tr>
<th>N</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>142</td>
<td>90</td>
<td>20</td>
<td>110</td>
<td>61.14</td>
<td>18.284</td>
<td>-.126</td>
<td>-.602</td>
</tr>
</tbody>
</table>

The 90-minute constraint for this project was further 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 most transfer journeys. 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. In addition Dublin Bus more recently introduced a ticket which is valid for 90 minutes. Starting from the initial validation of the ticket the passenger can board as many buses as needed as long as the validation occurs within the 90 minute constraint.

As the time difference between A1 and B1 is a derivable attribute, future analysis can be restricted to transfer journeys that had e.g. a time difference of 60 minutes if one wished to run another analysis using different time constraints.

In an ideal scenario where geographic coordinates of the bus stops are available the dataset
could be enhanced with estimated journey times and waiting times on an individual basis. This would lead to a variable transfer journey time constraint which may lead to a more precise result. However, as this type of data was not available for this project this method cannot be implemented.

**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. The algorithm uses the following procedures (see also Figure 4.6):

Assume data are held in files, each file \( j \) corresponding to one day’s records.

For each file (day) \( j \) in database:

For each record \( i \) in file \( j \):

if \((g_i = g_{i+1} \land h_i = h_{i+1} \land r_i \neq r_{i+1} \land t_{i+1} - t_i \leq m)\)

then

\((i, i + 1)\) is a transfer journey

end

end

**Step 1:** All variables are declared and initialised - A data structure is used to store the various data attribute values.

**Step 2:** Select file \( j \), read boarding record \( i \), compare it with \( i + 1 \) and verify that the conditions listed in steps 4, 5, 6 and 7 are true.

**Step 4:** Compare ticket ID numbers \((g_i \) and \( g_{i+1}\)) of the two records to check if the transfer journey was carried out by the same passenger.

**Step 5:** Compare matching ticket type ID \((h_i \) and \( h_{i+1}\)) as the same Ticket ID could have been issued for more than one ticket type but only once for the same ticket type.
4.5. EXTENSION OF THE PROJECT DATABASE – TRANSFER JOURNEY IDENTIFIER

**Step 6:** Compare the boarding time of each boarding and ensure that they occurred less than 90 minutes apart from each other ($t_{i+1} - t_i \leq m$ minutes).

**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 $m$ 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.

**Step 8:** 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 $i = i + 1$ to $i + 1$ starting at step 4.

This analysis focused on transfer journeys with one transfer point. A study on a small sample (data of one day) indicated that less than 1% of all transfer journeys have potentially two or more transfer points (passenger boards three or more buses to reach his/her destination). It therefore was decided to ignore these transfers.

The average runtime for 122 files was approximately 6 hours on a 2.4 MHz Pentium 4 with 1024 MB RAM. This PC was used for all calculations.

4.5.3 Possible Errors

It is obvious that many circumstances could lead to errors when running the transfer journey algorithm. There are two types of errors: Firstly the EFC data can contain errors that might have been caused by faulty equipment or during the data transferral process (see 2.3.4). Secondly, the algorithm may fail to address a transfer journey or links two boardings to form a transfer journey although they were in fact unlinked trips. The following section will address errors that fall into either of these two categories.

Equipment or media malfunction can lead to data that were either recorded wrongly or not at all. If a faulty ticket could not be read no record exists which may lead to a false positive transfer journey identification. Furthermore, wrongly recorded data may also lead to either false positive or false negative identification of linked trips. This is more a back-office problem and cannot be addressed by the algorithm which is focused on trip level results. It would be possible to identify the data error rate in an independent analysis and then incorporate this error into the final aggregated results.

Another error that could arise is through ticket swapping, meaning that more than one person uses the same ticket. This would lead to a bias when analysing individual passenger travel paths. This however is not a data or algorithm problem and was therefore not further researched.

OAP's often board the bus and only show their free travel pass leading to missing data. Again, this can not be incorporated into the trip level results but could at best be added as an error term to the aggregated results.

The last error type focuses on false positive and false negative identification by the proposed algorithm. False positive identification could occur when passengers actually had two main purposes throughout their journey of which one purpose was addressed at the transfer point. As already mentioned, the 90 minute constraint would allow the passengers to go shopping
or carry out other tasks before boarding the second bus. At present this error could only be addressed by introducing a variable time constraint between boardings that is based on distance travelled or by incorporating the time it took to travel the distance from the first boarding to the transfer point. Furthermore it would be possible to analyse the arrival time of the bus service at the transfer stage which also allows to assign a variable time constraint. Both methods can only be implemented when the EFC data are enriched with geographic coordinates. Another false positive case occurs when the passenger uses an alternative route to travel back to their origin. This was identified in the following section where a simple random sample analysed the correctness of the identified transfer journeys. False negative transfer journey pairs are believed to be not as common in particular due to the 90 minute constraint. In theory it might be possible that the 90 minute constraint is not long enough to incorporate journey times and waiting times. Such journeys, however, would need to take place during peak time from the very south to the very north of the city (or vice versa) including a boarding onto a very infrequent second route.

The following sections will address the validation of the transfer journey results.

4.5.4 Validation of the Results Using a Simple Random Sample

Ireland does not have the equivalent of the NTS in the UK. Other validation methods therefore needed to be explored to validate the classification algorithm. The following three sections show various methods.

A simple random sample was extracted from the results set in order to analyse how the transfer journey algorithm performed. This model is one method to analyse the shortcomings of the algorithm and possibly identifies potential improvements for the algorithm. The population size of all identified transfer journeys was 1,433,120. The sample size was calculated using the following formula (Dillman, 2000):

$$
Ns = \frac{(N_p)(p)(1 - p)}{(N_p - 1)(B/C)^2 + (p)(1 - p)}
$$

(4.5.1)

Where:

- $N_s$ sample size required for the desired level of precision
- $N_p$ size of population
- $p$ Proportion of the population expected to choose one of the two response categories. In this case $p$ will be 0.5 (50/50 split) in order to ensure a maximum variation in the sample.
- $B$ Acceptable amount of sample error
- $C$ Z statistic associated with the confidence level
The sample size for a sample with a 95% confidence interval and a sampling error of 5% is therefore

\[ N_S = \frac{(1,433,120)(0.5)(1 - 0.5)}{(1,433,120 - 1)(0.05/1.96)^2 + (0.5)(1 - 0.5)} \]

\[ = 384 \text{ cases} \quad (4.5.2) \]

The simple random sample of 384 individual transfer journey records was identified by retrieving and randomly ordering the recordset from ORACLE. Each of the 384 transfer journey records was then manually compared to maps and schedules. The aim was to categorise each observation into possible and not possible categories. The category possible contained observations of which it was believed that the transfer journeys could have occurred in relation to time between boardings, location of the boardings and time differences between boardings. Table 4.4 shows the results that were obtained from this manual verification. The study identified 95.6% of the simple random sample instances as correct or possible. Only 17 of the 384 randomly selected transfer records were believed to be incorrect (false positive). This was mainly due to the fact that the passengers used the second leg of the journey to return to their origin resulting in a violation of the assumption. Incorporating a list of alternative routes when comparing the boarded route may eliminate this problem. It would further be possible to improve the algorithm by introducing a dynamic adjustable variable for the time difference between the two potential boarding records. For example, if the two potential records of a transfer journey both lie in the same zone then the maximum allowed time difference between these two boardings needs to be reduced. Ideally, geographic coordinates identify the distance between the two zones and then calculate the maximum time it would take to travel from one zone to the next. This time could then be used to determine the maximum time delay between boarding one and boarding two. As geographic coordinates were not available and the result of the manual validation using the simple random sample was satisfactory this method was not further pursued. False negative cases are much harder to find and simulate. For example, the last of a 10-journey ticket is used to validate boarding one and a new 10-journey ticket is used to validate the second leg. Successfully identifying such a scenario would be next to impossible without knowing the IDs of both 10 journey tickets.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not possible</td>
<td>17</td>
<td>4.4</td>
</tr>
<tr>
<td>Possible</td>
<td>367</td>
<td>95.6</td>
</tr>
<tr>
<td>Total</td>
<td>384</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 4.4: Validation Results of the Simple Random Sample of Transfer Journey Records
4.5.5 Validation of the Results using Survey Data

The core concept of the transfer journey states that the journey has to have a single main purpose and therefore excludes trips where passengers made a detour. The Dublin Transportation Office (DTO) rolled out a survey in 2002 that focused on journeys that passengers carried out to and from work. The DTO received 64,456 valid responses of which 9,231 passengers used a bus to carry out both journeys to and from work. A total of 11,118 passengers used a bus to work and 11,925 used the bus as return mode. One question of the survey focused on whether the participating passenger made a de-tour or not. For the journey to work over 95.5% made no detour. This percentage is slightly lower for the return journey where only 85.9% travelled directly to their final destination. Although it is unknown how many passengers needed to transfer to a different service for reaching their desired destination one could derive from such a result that public transport passengers that use a bus to and from work mostly travel direct to and from their origin and destination. This in turn consolidates the part of the assumption that states that passengers need to travel directly without making any stops for different purposes. As this survey focused on work journeys the validation of the transfer journey algorithm can only be applied to peak time journeys. Due to missing detail of information it was not possible to use this survey for any other form of validation.

Ideally data gathered by a National Travel Survey (NTS) should be available for validation purposes. However, such a survey does not exist in Ireland (White, 2006).

4.5.6 Validation of the Results using a Monte Carlo Simulation

A Monte Carlo simulation was applied to a sample of the results to identify how the algorithm would perform if an error is randomly assigned to various attributes. Such an analysis aims to determine the robustness of the algorithm by testing its performance after randomly introducing a certain error percentage in the data files. This introduction of error could easily resemble the most common EFC errors. The simulation focused on the results of one day (Wednesday in April 1999). A total of 91,080 boarding records were stored for this particular day. We look at how the results of the transfer journey analysis change as we add error to the dataset.

The aim was to create datasets where 5%, 10%, 15%, 20%, and 25% of the value of one attribute were in error. The error was introduced by assigning a random valid value to that attribute. A C++ program randomly assigned the errors to the source file which was later used by the transfer journey identification algorithm. Three attributes that the transfer journey algorithm uses were looked at: *Time of Boarding*, *Ticket ID* and *Ticket Type ID*. The setup created 100 files for each attribute for each of the given error percentages. For each attribute and percentage error, 100 data sets were simulated and the transfer journey algorithm run. The difference in the result with that of the original file is computed as percentage. This resulted in 500
files for each test. Although a larger number of iterations would strengthen this study due to the runtime of the transfer journey algorithm it was decided that 100 iterations is sufficient to analyse the robustness of the algorithm with regard to the most common data errors. The following paragraphs present the findings of the Monte Carlo simulation.

For each simulated file we measure the percentage of records that were classified differently (as transfer journey of non transfer journey) from the original file. We call this the classification error. This can also be expressed as

\[
\varepsilon = \frac{100}{n} \left( |T_J^{\text{true}} \cap (\neg T_J^*)| + |(\neg T_J^{\text{true}} \cap T_J^*)| \right) \tag{4.5.3}
\]

where \( n \) is the number of records, \( T_J \) are identified transfer journeys in the original dataset and \( T_J^* \) are transfer journeys in the simulated file.

After the original data file was randomly changed and the transfer journey algorithm was executed it was necessary to analyse the results. The aim was to identify how many transfer journey pairs were matching when comparing the original result with the error induced result. Therefore false negative and false positive results in the Monte Carlo files were not included as correct transfer journey pairs. The false positive and false negative pairs could have occurred by randomly assigning a value to an attribute which would then lead by chance to a transfer journey pair.

**Time Attribute:** The Time attribute stored the time of boarding of the passenger. The value of this parameter was changed by randomly selecting 5%, 10%, 15%, 20%, and 25% of the records and then a random assignment of a value that was within the range of possible values. The range of possible values therefore reached from 00:00 to 23:59. Thus, time could only be randomly changed to another time but not to any random number. The purpose of changing the value of this variable was to simulate some of the most common EFC errors such as incorrect sign on, wrong time due to lack of maintenance and read-write errors.

Table 4.5 shows the descriptive statistics of the Monte Carlo simulation for the various induced error percentages of the Time attribute. The error in identifying transfer journey runs almost linear in relation to the randomly introduced error (see Figure 4.7(a)).

<table>
<thead>
<tr>
<th>Simulation Study</th>
<th>N</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>5% Error</td>
<td>100</td>
<td>.86</td>
<td>4.40</td>
<td>5.26</td>
<td>4.78</td>
<td>.1798</td>
<td>.0323</td>
</tr>
<tr>
<td>10% Error</td>
<td>100</td>
<td>1.57</td>
<td>8.80</td>
<td>10.37</td>
<td>9.48</td>
<td>.2651</td>
<td>.0703</td>
</tr>
<tr>
<td>15% Error</td>
<td>100</td>
<td>1.44</td>
<td>13.25</td>
<td>14.69</td>
<td>13.93</td>
<td>.3089</td>
<td>.0954</td>
</tr>
<tr>
<td>20% Error</td>
<td>100</td>
<td>2.00</td>
<td>17.49</td>
<td>19.49</td>
<td>18.38</td>
<td>.3968</td>
<td>.1575</td>
</tr>
<tr>
<td>25% Error</td>
<td>100</td>
<td>2.78</td>
<td>21.91</td>
<td>24.69</td>
<td>22.82</td>
<td>.4014</td>
<td>.1611</td>
</tr>
</tbody>
</table>
This analysis lead to a more specific Monte Carlo simulation in the sense that the range of possible values was decreased. Two more simulations were defined: the first simulation included data files which time values were randomly changed to values in a range of -15 to 15
minutes; the second simulation used a range of ±60 minutes.

Table 4.6 and Figure 4.7(b) show the results of the ±60 minute simulation. This Monte Carlo simulation shows that the algorithm performs very well when small errors with regard to the time attribute are introduced. Even a 25% error rate only results in an average 0.9% error of the transfer classification algorithm.

Table 4.6: Summary of Classification Error (as %) for Time Attribute (60 Minute Range) Simulation Study

<table>
<thead>
<tr>
<th>N</th>
<th>Range Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>5% Error</td>
<td>100</td>
<td>.28</td>
<td>.30</td>
<td>.16</td>
<td>.0463</td>
</tr>
<tr>
<td>10% Error</td>
<td>100</td>
<td>.25</td>
<td>.45</td>
<td>.33</td>
<td>.0573</td>
</tr>
<tr>
<td>15% Error</td>
<td>100</td>
<td>.33</td>
<td>.67</td>
<td>.51</td>
<td>.0674</td>
</tr>
<tr>
<td>20% Error</td>
<td>100</td>
<td>.55</td>
<td>.87</td>
<td>.70</td>
<td>.0684</td>
</tr>
<tr>
<td>25% Error</td>
<td>100</td>
<td>.66</td>
<td>1.05</td>
<td>.90</td>
<td>.0657</td>
</tr>
</tbody>
</table>

Table 4.7 and Figure 4.7(c) show the results of the Monte Carlo simulation of the time attribute with introduced errors of ±15 minutes. The average caused transfer identification error of the algorithm is at 25% error only 0.4%. It can therefore be stated that small EFC errors with regard to the time attribute can be neglected.

Table 4.7: Summary of Classification Error (as %) for Time Attribute (15 Minute Range) Simulation Study

<table>
<thead>
<tr>
<th>N</th>
<th>Range Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>5% Error</td>
<td>100</td>
<td>.25</td>
<td>.25</td>
<td>.07</td>
<td>.0554</td>
</tr>
<tr>
<td>10% Error</td>
<td>100</td>
<td>.43</td>
<td>.43</td>
<td>.13</td>
<td>.0950</td>
</tr>
<tr>
<td>15% Error</td>
<td>100</td>
<td>.45</td>
<td>.45</td>
<td>.19</td>
<td>.1087</td>
</tr>
<tr>
<td>20% Error</td>
<td>100</td>
<td>.59</td>
<td>.59</td>
<td>.30</td>
<td>.1118</td>
</tr>
<tr>
<td>25% Error</td>
<td>100</td>
<td>.60</td>
<td>.72</td>
<td>.40</td>
<td>.1299</td>
</tr>
</tbody>
</table>

In summary, we believe that small errors for the time attribute have no significant impact on the results produced by the transfer journey algorithm. This is mainly due to the selection of \( m = 90 \) minutes. False negatives and false positives can occur when the error changes the time in such a manner that the time difference of the two boardings is around 90 minutes. The Monte Carlo simulation clearly showed evidence that supports this statement.

**Ticket ID:** This attribute stored the Ticket ID which is in combination with the ticket type a unique number which can be used to identify individual passengers over a certain period of time. The value of this parameter was changed by randomly selecting 5%, 10%, 15%, 20%, and 25% of the records and then a random assignment of a value that was within the range of possible values.
Table 4.8 shows the descriptive statistics of the Monte Carlo simulation for the various induced error percentages of the Ticket ID attribute. The graphical representation of this simulation can be seen in Figure 4.7(d). Randomly introduced errors cause a high transfer journey identification error. This had to be expected as changing Ticket ID assigns the boarding record to a 'new' passenger.

<table>
<thead>
<tr>
<th>Error Percentage</th>
<th>N</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>100</td>
<td>.95</td>
<td>7.68</td>
<td>8.63</td>
<td>8.20</td>
<td>.179</td>
<td>.0321</td>
</tr>
<tr>
<td>10%</td>
<td>100</td>
<td>1.08</td>
<td>15.70</td>
<td>16.78</td>
<td>16.22</td>
<td>.227</td>
<td>.0516</td>
</tr>
<tr>
<td>15%</td>
<td>100</td>
<td>1.46</td>
<td>23.46</td>
<td>24.92</td>
<td>24.11</td>
<td>.272</td>
<td>.0741</td>
</tr>
<tr>
<td>20%</td>
<td>100</td>
<td>2.12</td>
<td>30.81</td>
<td>32.93</td>
<td>31.82</td>
<td>.338</td>
<td>.1143</td>
</tr>
<tr>
<td>25%</td>
<td>100</td>
<td>1.53</td>
<td>38.93</td>
<td>40.46</td>
<td>39.54</td>
<td>.261</td>
<td>.0682</td>
</tr>
</tbody>
</table>

Ticket Type ID: This attribute stored the Ticket Type ID which is in combination with the Ticket ID a unique number which can be used to identify individual passengers over a certain period of time. The value of this parameter was changed by randomly selecting 5%, 10%, 15%, 20%, and 25% of the records and then a random assignment of a value that was within the range of possible values.

Table 4.9 shows the descriptive statistics of the Monte Carlo simulation for the various induced error percentages of the Ticket Type ID attribute. The graphical representation of this simulation can be seen in Figure 4.7(e). Randomly introduced errors cause a high transfer journey identification error. This had to be expected as changing Ticket Type ID assigns the boarding record to a 'new' passenger that may or may not already exist.

<table>
<thead>
<tr>
<th>Error Percentage</th>
<th>N</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>100</td>
<td>.79</td>
<td>9.04</td>
<td>9.82</td>
<td>9.40</td>
<td>.161</td>
<td>.0261</td>
</tr>
<tr>
<td>10%</td>
<td>100</td>
<td>1.12</td>
<td>17.93</td>
<td>19.05</td>
<td>18.56</td>
<td>.211</td>
<td>.0445</td>
</tr>
<tr>
<td>15%</td>
<td>100</td>
<td>1.68</td>
<td>26.75</td>
<td>28.43</td>
<td>27.52</td>
<td>.266</td>
<td>.0705</td>
</tr>
<tr>
<td>20%</td>
<td>100</td>
<td>2.38</td>
<td>35.57</td>
<td>37.95</td>
<td>36.28</td>
<td>.322</td>
<td>.1037</td>
</tr>
<tr>
<td>25%</td>
<td>100</td>
<td>2.82</td>
<td>44.61</td>
<td>47.43</td>
<td>45.02</td>
<td>.316</td>
<td>.1001</td>
</tr>
</tbody>
</table>

The various Monte Carlo simulations showed that recorded error in the attributes Ticket ID, Ticket Type ID and Time result in high transfer journey identification errors. However, smaller changes in time result in error rates that can be almost neglected. Even when 25% of all time values were changed by either up to 15 or 60 minutes the maximum transfer journey
identification error was only 1.05%. The algorithm therefore performs very well with regard to small errors of the time attribute which is also one of the most likely incorrect attributes as system time is not set centrally. Errors in Ticket Type ID or Ticket ID on the other hand result in a larger error because the boarding record is attributed to a different passenger when either of these values change.

### 4.5.7 Summary of the Iterative Classification Algorithm

The iterative classification algorithm extends the original database with information about transfer journeys. The newly added database table shows a new record for each transfer journey (see Table 4.10).

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Width</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Numeric</td>
<td>10</td>
<td>Unique ID for each transfer journey record</td>
<td>01045447</td>
</tr>
<tr>
<td>ID A</td>
<td>Numeric</td>
<td>10</td>
<td>Link to individual passenger boarding A</td>
<td>15716867</td>
</tr>
<tr>
<td>ID B</td>
<td>Numeric</td>
<td>10</td>
<td>Link to individual passenger boarding B</td>
<td>15733816</td>
</tr>
<tr>
<td>A Route ID</td>
<td>String</td>
<td>4</td>
<td>Route ID of individual passenger boarding A</td>
<td>R24</td>
</tr>
<tr>
<td>B Route ID</td>
<td>String</td>
<td>4</td>
<td>Route ID of individual passenger boarding B</td>
<td>R7</td>
</tr>
<tr>
<td>A Stage ID</td>
<td>Numeric</td>
<td>2</td>
<td>Stage ID of individual passenger boarding A</td>
<td>25</td>
</tr>
<tr>
<td>B Stage ID</td>
<td>Numeric</td>
<td>2</td>
<td>Stage ID of individual passenger boarding B</td>
<td>75</td>
</tr>
<tr>
<td>A Direction</td>
<td>Numeric</td>
<td>1</td>
<td>Direction of individual passenger boarding A</td>
<td>0</td>
</tr>
<tr>
<td>B Direction</td>
<td>Numeric</td>
<td>1</td>
<td>Direction of individual passenger boarding B</td>
<td>0</td>
</tr>
<tr>
<td>A Coarse Zone</td>
<td>Numeric</td>
<td>2</td>
<td>Coarse Zone of individual passenger boarding A</td>
<td>Z</td>
</tr>
<tr>
<td>B Coarse Zone</td>
<td>Numeric</td>
<td>2</td>
<td>Coarse Zone of individual passenger boarding B</td>
<td>Y</td>
</tr>
<tr>
<td>A Area</td>
<td>String</td>
<td>20</td>
<td>Area description of individual passenger boarding A</td>
<td>Area 1</td>
</tr>
<tr>
<td>B Area</td>
<td>String</td>
<td>20</td>
<td>Area description of individual passenger boarding B</td>
<td>Area 7</td>
</tr>
<tr>
<td>Time Difference A/B</td>
<td>Numeric</td>
<td>2</td>
<td>Time difference between boarding A and B</td>
<td>37</td>
</tr>
<tr>
<td>A Stage Time</td>
<td>Date</td>
<td>20</td>
<td>Boarding time of passenger boarding A</td>
<td>6:29</td>
</tr>
<tr>
<td>B Stage Time</td>
<td>Date</td>
<td>20</td>
<td>Boarding time of passenger boarding B</td>
<td>7:06</td>
</tr>
<tr>
<td>A Start Hour</td>
<td>Numeric</td>
<td>4</td>
<td>Boarding Hour of passenger boarding A</td>
<td>6</td>
</tr>
<tr>
<td>B Start Hour</td>
<td>Numeric</td>
<td>4</td>
<td>Boarding Hour of passenger boarding B</td>
<td>7</td>
</tr>
<tr>
<td>Time Period</td>
<td>Numeric</td>
<td>1</td>
<td>Morning Peak, Evening Peak or Off Peak Period</td>
<td>Off Peak</td>
</tr>
<tr>
<td>Date</td>
<td>Date</td>
<td>20</td>
<td>Date of transfer journey</td>
<td>01/04/1999</td>
</tr>
<tr>
<td>Ticket ID</td>
<td>Numeric</td>
<td>7</td>
<td>Unique ticket ID of transfer journey</td>
<td>7052</td>
</tr>
<tr>
<td>Ticket Type</td>
<td>Numeric</td>
<td>6</td>
<td>Ticket Type used</td>
<td>671</td>
</tr>
<tr>
<td>Ticket Name</td>
<td>String</td>
<td>36</td>
<td>Name of the ticket used</td>
<td>Weekly Adult</td>
</tr>
<tr>
<td>Ticket Category</td>
<td>Numeric</td>
<td>4</td>
<td>Ticket Category such as OAP, Adult, Student, Child, Family</td>
<td>Adult</td>
</tr>
<tr>
<td>Ticket Duration</td>
<td>Numeric</td>
<td>4</td>
<td>Period the ticket is valid (1 month, 1 weeks, 3 days, 1 day)</td>
<td>1 Week</td>
</tr>
</tbody>
</table>

Each of these records is based on data of two individual passenger boardings (A and B). The new table facilitates much more detailed analysis about transfer journeys than the original database offered. Now it is possible to generate analysis that will contribute to more fully understand transfer passenger behaviour. Further analysis can be applied to improve decision support on an operational or policymaking level. A more detailed analysis and discussion can
be found in Chapter 5 where the identified transfer journey data set will be explored.

A more detailed description of the transfer journey identification section was published by Hofmann and O'Mahony (2004, 2005b).

Some of the weaknesses of this algorithm are elaborated in greater detail in Section 6.12.3.

The following section describes an analysis that was carried out to explore the level of data quality with regard to the recordings that required human interaction.

4.6 How 'Good' is Electronic Fare Collection Data?

EFC data are used more and more frequently in many academic and commercial public transport projects. Various data attributes are automatically recorded when passengers board a public transport vehicle. The main focus of this section is the quality of the data. This is a common concern among the research community, particularly as often the bus driver has a certain influence on the recorded data. In many transport networks it is the bus driver's responsibility to provide the system with information such as which bus stop the public transport vehicle is currently serving and therefore is also in charge of identifying the correct location of passenger boardings. This section shows three different analyses that provide results with regard to the reliability of the bus driver and his/her input. It further discusses the real impact on the measured data quality.

The first approach analyses the frequency of recorded bus stops. The second approach analyses odd boarding distributions on the premise that more people board at busy bus stops. The last approach focuses on the arrival time of the vehicle at the bus stop on the premise that the arrival times on bus stops should not be the same for a number of bus stops.

4.6.1 What Defines Data to be 'Good'

Four common categories of data quality exist (Strong et al., 1997):

- Contextual Data Quality - relevancy, value added, timeliness, and completeness;
- Intrinsic Data Quality - accuracy, objectivity, believability and reputation;
- Accessibility of Data - accessibility and access security;
- Representation of Data - interpretability, ease of understanding, concise representation and consistent representation.
The main relevant dimensions for this analysis are completeness and accuracy. For example, when a study focuses on passenger boardings then all passenger boardings have to be represented by the data source (completeness) and all records have to reflect the correct data values (accuracy). The following analyses will demonstrate how the data were tested for completeness and accuracy. Perfect datasets are rare due to human or machine errors, however, knowing to what extent a dataset is incomplete still allows it to be used for analysis.

4.6.2 Analysis Methodology

Three different approaches were used to test the integrity and quality of the data. The aim was to investigate the bus driver’s consistency in recording the current location. The following three approaches were used:

- Frequency Tables;
- Odd Boarding/Alighting;
- Time Series.

Four different types of routes will be subject to the analysis. The different characteristics of each route can be seen in Table 4.11 and Figure 4.8. It further shows the number of bus journeys carried out on each route for one particular day split into inbound and outbound journeys. The variability of changing numbers of bus journeys per day is very small and can therefore be disregarded.

<table>
<thead>
<tr>
<th>Route</th>
<th>Number of Busses per Day</th>
<th>Type of Route</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Outbound</td>
<td>Inbound</td>
</tr>
<tr>
<td>18</td>
<td>64</td>
<td>70</td>
</tr>
<tr>
<td>13</td>
<td>59</td>
<td>51</td>
</tr>
<tr>
<td>123</td>
<td>163</td>
<td>180</td>
</tr>
<tr>
<td>3</td>
<td>95</td>
<td>87</td>
</tr>
</tbody>
</table>

The following sections will introduce the methodologies that were applied for this analysis.
exact this task has been carried out. The initial assumption of the approach was that each bus stop on a particular route should be recorded as often as any of the remaining bus stops. However, the location is only recorded in case the bus actually stops and passengers board. It is therefore expected that a certain pattern occurs which can be repeatedly produced analysing different bus routes with different route characteristics and different bus drivers. A smaller dataset only including data from a Wednesday in October 1999 was extracted for the analysis. A cross-tabulation was based on the dataset with the aim to obtain frequency numbers for each route’s bus stops of that particular day. The data were then grouped into bus routes and direction of the journey. This format favoured the generation of bar charts which were then compared. A random selection of charts will be displayed and discussed.

4.6.4 Odd Boarding/Alighting

This approach focuses mainly on the boarding patterns of passengers. It is assumed that passengers are more likely to board at main bus stops such as shopping malls, main streets or multi modal transfer nodes. A randomly chosen set of bus stops are identified and explored with the main aim to find odd boarding patterns. Odd boarding patterns could be considered as patterns where passengers boarded unexpectedly which could lead to the assumption that the bus driver keyed in the wrong bus stop or forgot to indicate the correct bus stop. For example, if bus stop 5 on route X is the closest bus stop to the pedestrian zone then it is expected that more people
board than on a bus stop outside the city centre (at least for an outbound journey). Main bus stops will be selected of which the researcher knows that a boarding pattern above the average is to be expected. This approach further focuses on bus stops where the route intersects with another mode of transport such as light rail or metro. These transfer nodes are expected to have a larger passenger boarding pattern than the previous or following bus stop. Figure 4.9 shows this scenario when the light rail intersects with the bus route at Stop 3 and Station 1. It is therefore expected to have a higher boarding pattern at bus stops 3 and 5 than at bus stops 4 and 6.

![Figure 4.9: Sample Layout of a Transport Network Section](image)

### 4.6.5 Time Series

The time series approach is the final test that is applied in assessing the quality of EFC data with regard to the influence of the bus drivers on data quality. The assumption is that the difference of bus stop arrival times of two bus stops cannot be less than a pre-defined period of time. This time frame depends on the network and also whether the bus stops were recorded in the city centre where stops are closer together than in suburban areas. Comparing the bus stop arrival times and calculating their differences can therefore provide an indicator of the data quality. A variability study strengthens this analysis.

### 4.6.6 Results

**Frequency Tables**

The frequency tables were created for each bus route.

The dataset consists of a grouped count of bus stops. The bus stops are in spatial sequential order for inbound and outbound journeys. As shown in Figure 4.10 the bus stop ID’s (shown on
the x-axis) are unique throughout each route hence the difference in bus stop numbers for the inbound and outbound journeys.

For example bus stop ID '80' is the corresponding bus stop of ID '20' in the opposite direction. The y-axis shows how often each particular bus stop was recorded throughout the day. The recording of the first bus stop is simultaneously the total number of journeys that were dispatched on the particular route and its direction (throughout one day).

Figure 4.10 shows the distributions of recorded bus stops for the four routes described in Table 4.11.

As expected each chart shows a reoccurring pattern for each route. The recording of the first bus stop is mandatory for the bus driver to start the recording of the journey. After recording the highest number of bus stops at the beginning of the journey the number of recorded bus stops generally drops slightly before levelling off for a few stops before consistently decreasing over the last few stops. The data show that the bus driver records less bus stops as the route approaches the final stops. This may reflect the fact that fewer passengers board during the last few stops as there is no need to board for only two or three bus stops. However, it was expected that more people exit the bus when approaching city centre which would imply that the bus had to stop more often than in suburban areas. This further implies that the bus driver only
identifies the location of the bus when passengers are boarding. This may become a problem when calculating arrival times which will be further elaborated in Section 7.5 because of missing stage records towards the end of each journey as they are not as frequent as they were at the beginning.

It is also interesting that the bar charts consist of a certain amount of symmetry with regard to inbound and outbound journeys. This further assumes consistency of the bus driver's location recording pattern. Furthermore it seems that the number of outbound journeys marginally exceeds the number of inbound journeys.

It is noteworthy that this pattern is not dependent on the time of the day. Various analyses have shown that peak time and off-peak time show the same characteristics with regard to the patterns shown in Figure 4.11. It could therefore be argued that the drop of recorded bus stops in the last section of the journey is not related to fatigue of the bus driver.

Some graphs show a second peak during the journey recordings. These are popular bus stops or suburban areas where more passengers are expected to board.

There are three ways to remedy that not all bus stages are recorded. The first method is
to automate the recording using an AVL system that is integrated with the EFC system. The second option would be to make the bus drivers to press the button when passing the bus stop regardless whether passengers board or alight. The last option would require a statistical model that may be able to infer the arrival times of buses at bus stops based on historical data.

Odd Boarding

A frequency table was created showing the number of passengers that boarded on one particular day (Wednesday in October 1999). A series of bar charts then visualise the boarding patterns along a certain route. Table 4.12 shows the different routes which were subject to the analysis and their total recorded passengers throughout the day. The table differentiates between boardings on outbound and inbound journeys. The variability between the day used in this study and other days is minimal and can therefore be disregarded. The approaching symmetry between total numbers of boardings of outbound and inbound journeys led to the decision that the direction of the journey with regard to passenger boarding numbers can also be disregarded.

Table 4.12: Inbound - Outbound - Total Passenger Numbers

<table>
<thead>
<tr>
<th>Route ID</th>
<th>Outbound</th>
<th>Inbound</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>649</td>
<td>583</td>
</tr>
<tr>
<td>13</td>
<td>594</td>
<td>542</td>
</tr>
<tr>
<td>123</td>
<td>939</td>
<td>1,051</td>
</tr>
<tr>
<td>18</td>
<td>809</td>
<td>816</td>
</tr>
</tbody>
</table>

Figure 4.12 shows the results of the analysis. Each bar chart presents the boarding pattern of one route. The x-axis shows the identifier of the bus stop in sequential spatial order and also labels the stops where an increased boarding pattern occurred.

The charts show an increase in passenger boarding at demographic landmarks such as shopping streets, churches, multi modal transfer nodes, hospitals, suburban centres and schools. Considering the results it can be argued that the bus driver identified the correct bus stop when passengers boarded at these landmarks.

Time Series

Unfortunately, the time recorded with each boarding only includes hours and minutes in a 24 hour format but does not contain seconds. This prohibits an exact analysis because in theory the bus could serve two bus stops within the same time stamp (time recorded to nearest minute), particularly in the city centre where the distances between bus stops are short or along dedicated
bus corridors where the speed of the bus would be increased. It was decided to show a graphical representation of arrival times of several buses at each bus stop. Different peak/off peak times and directions (inbound/outbound) were chosen.

Figure 4.13 show the different scenarios. The horizontal axis shows the sequential bus stop IDs in a non-spatial manner. The vertical axis shows the time of arrival. Each line within the line charts indicates a bus that serves a number of bus stops. Horizontal bars connecting two nodes indicate that the time recorded for the two bus stops was identical.

Figure 4.13(a) shows the arrival times of several buses serving the Route 123 in outbound direction throughout the morning peak period. The arrival times of buses at the various bus stops are different most of the time. However, it was noticed that the time recorded towards the end of the route often remains the same which would indicate that the bus driver did not identify the correct bus stop. It could be argued that only passengers with a magnetic strip card boarded that close to the final stop and the identification of the correct bus stop was therefore not necessary.

Figure 4.13(b) shows the arrival times of several buses serving Route 13 in outbound direction throughout the afternoon off peak period. The same explanation as given above applies to
4.6. HOW 'GOOD' IS ELECTRONIC FARE COLLECTION DATA?  

Figure 4.13: Arrival Time at Bus Stops

this graph. Apart from a few times throughout the route the recorded time difference over the period of the service was different, indicating that the bus driver keyed in the correct bus stops. This graph also shows that there is no problem with regard to bunching. The frequency of the service is regular and the headway is spread evenly.

Figure 4.13(c) shows the arrival times of several buses serving Route 18 in inbound direction throughout the evening peak period. Again, only a few bus stops were recorded with the identical time to the previous bus stop. Although this is not related to data quality it has to be pointed out that the graph favours the presentation of bunching and headway.

Figure 4.13(d) shows the arrival times of several buses serving the Route 3 in inbound direction throughout the morning peak period. For some reason some of the bus services did not collect data for the last three to five bus stops. Although this pattern was noticed to a certain extent throughout the other routes it was not as drastic as identified with route 3. There is no explanation for this pattern. It could have been that some of the routes were scheduled not to provide a service after bus stage ‘75’ (O’Connell Street – Main City Centre Street).

The following points could be identified when analysing the sequential time differences of
4.6.7 Discussion & Summary

All three approaches presented in this section show that the bus driver interaction has a consistent pattern with regard to recording the bus location and thus the boarding time. There are error margins which have to be considered when interpreting the results obtained from EFC data analysis. The four routes that served as subjects of this analysis represent average radial and orbital routes.

In summary, the interaction of the bus driver as location indicator has certainly an impact on the data quality. However, the main recordings are correct and it is debateable whether the error margin would bias the results considerably. Even if an actual in-vehicle time of 35 minutes is inferred as 32 or 38 minutes it is still better than having no identification of such a performance measure at all. Taking the data quality measurements introduced by Strong et al. (1997) this study can conclude that the accuracy of the data is, within a margin of error, acceptable. After having carried out the analysis it was concluded that the data quality of the recorded records is representative. However, it was further found that the records that were not recorded (e.g. in case no one boarded) but could have been recorded have a great impact when focusing on stage level analyses. Therefore the data quality measurement 'completeness' introduced by Strong et al. (1997) is not entirely fulfilled. This especially applies to the bus stage records as demonstrated above. In theory such missing records could be inferred. For example, missing stage records could be inferred using a Geographic Information System (GIS) and the data of the recorded stages in order to estimate the arrival time of the missing stages by considering time of previous and subsequent record, average speed and distance travelled. However, this was not further pursued as no GIS of the bus network was available.
4.7 Summary

This chapter mainly focused on the methods used to extend the database with three main attributes; the spatial identifiers, coarse zone and area description and the transfer journey identifier. The first two were necessary to add semantics to the dataset with regard to the spatial location of bus stages whereas the second was needed to identify transfer journeys (linked trips).

The last section then explored the data quality of the dataset.

The following are the main conclusions that can be drawn from this chapter:

- Relational databases suit the structured format of EFC data better than any other database model;
- There is a need for a formal data migration framework as the data produced by the Wayfarer EFC system are in semi structured and semi encrypted text file format. Error checks are further important to identify any illogical data values. The dataset received from Dublin Bus has almost no illogical errors. A combination of C++, PL/SQL and SQL was used to decrypt, structure, migrate and validate the data. The structure of the database was designed following the rules of the relational database model;
- Assigning each bus stage to a spatial zone was considered to be important. Unfortunately there were no geographic coordinates of bus stages available for the Dublin Bus network. In a commercial project this should be remedied as knowing the exact locations of bus stages would certainly improve the potential usage of the dataset as well as improve the representational capabilities of the results;
- The extension of the database with an identifier that categorises boarding records into single and transfer journeys was necessary for a transfer journey analysis (see Chapter 5) but more importantly for the OD estimation which will be introduced in Chapter 6. It was further decided that the 90 minute cut-off point was the correct choice although this could be reduced to 75 minutes resulting in the loss of only 10% of originally declared transfer journeys. Denormalisation of the transfer journey table was implemented to ensure faster data retrieval speeds thus shortening the time required for the transfer journey analysis. Depending on the system, the process of identifying transfer journeys could become unnecessary when smart cards are in operation. Nevertheless, the identification of transfer journeys is important for the immediate analysis of transfer journeys and for future analyses or extensions where linked trips have to be considered separately;
- Various validation performance test methods were applied to the algorithm to check its
validity, robustness and its performance. A simple random sample, secondary survey and Monte Carlo simulations were used. All validation methods showed promising results.

- The data quality section was considered important as the quality of the generated results can only be as good as the quality of the data itself. Although the analysis concluded that the recorded data is accurate (within a margin of error) it was also discovered that the dataset is not as complete as it could have been. This is mainly with regard to stages where no passengers boarded which especially occurs towards the final stages of each route. However, this also means that passengers alight towards the final stage of the route and therefore the bus had to stop at the bus stage. There are two possible explanations: First, the bus driver did not indicate the stage as no passenger alighted or second, the bus driver did indicate the stage but the Wayfarer system did not record it as no person boarded. In summary, although there is frequent human interaction the system where the bus driver has to indicate the location of the stop does work and the recorded data seem to be accurate. With the growing integration of more complex technology most systems will not have to face this problem for much longer.

The following chapter provides some analyses of the data including the new data attribute that was created using the iterative classification algorithm. Furthermore the network and route symmetry of single and transfer journeys are explored.
Chapter 5

Transfer Journey Analysis

5.1 Introduction

This chapter focuses on the exploratory data analysis of transfer journeys. The previous chapter described the process of linking two formerly independent passenger boardings together by comparing the various parameters of each of the individual boarding records. Some of the two independent boardings built a record of a transfer journey which provides the opportunity for further analysis of the EFC data (Bagchi and White, 2003; Hofmann and O’Mahony, 2004). The previously identified variables (see Table 4.10) will be used throughout this section to create numerous transfer journey related statistics and analyses which can be used to more fully understand public transport operations and passengers’ behaviour. The results of this section have to be interpreted with the fact in mind that only magnetic fare card data were used for generating the dataset.

Furthermore a comprehensive analysis will be described focusing on a widely accepted travel symmetry assumption. The obtained transfer journey pairs as well as single journey records will be used to analyse symmetry with regard to transfer patterns in Dublin’s public transport network.

5.2 General Descriptive Statistics of Transfer Journeys

The iterative classification algorithm (see Section 4.5.2) enriched the database with an identifier that shows whether the individual passenger boarding was part of a transfer journey. This newly added data attribute facilitates the analysis documented in this chapter. 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 analyses are based on these four months unless stated differently.
Any comparison between transfer journeys and single journeys only considers the difference between total boardings minus twice the number of transfer journeys (two boarding records for each transfer journey). Table 5.1 shows a breakdown and summary of the identified transfer journeys in relation to the total number of magnetic strip card transactions of the four-month period.

<table>
<thead>
<tr>
<th>Month</th>
<th>Transfer Journeys</th>
<th>Single Journeys</th>
<th>Total Boardings</th>
<th>Percentage of Transfer Journeys</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr-99</td>
<td>322,869</td>
<td>1,179,911</td>
<td>1,825,649</td>
<td>35.4%</td>
</tr>
<tr>
<td>May-99</td>
<td>342,854</td>
<td>1,226,389</td>
<td>1,912,097</td>
<td>35.9%</td>
</tr>
<tr>
<td>Sep-99</td>
<td>353,828</td>
<td>1,303,030</td>
<td>2,010,686</td>
<td>35.2%</td>
</tr>
<tr>
<td>Oct-99</td>
<td>413,569</td>
<td>1,565,427</td>
<td>2,392,565</td>
<td>34.6%</td>
</tr>
<tr>
<td>Total</td>
<td>1,433,120</td>
<td>5,274,757</td>
<td>8,140,997</td>
<td>35.2%</td>
</tr>
</tbody>
</table>

### 5.3 Analysis of Time Difference between Boardings

The time difference between A1 and B1 is one of the deciding parameters whether a passenger boarding is part of a transfer journey. The assumption stated that the maximum accepted time difference between boarding at A1 and boarding at B1 is less than 90 minutes as defined in Section 4.5.2. As this is a derived data attribute (boarding time at B1 minus boarding time at A1) this parameter can be redefined when carrying out various analyses on the final data. 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.

Table 5.2 shows the main descriptive statistics of the time difference attribute considering all identified transfer journeys of the four month period. The arithmetic mean of the time difference is 37.75 minutes. Quartile 1, Quartile 2 and Quartile 3 are 20 minutes, 34 minutes and 53 minutes respectively. The distribution of time differences is positively skewed which is shown in Figure 5.1.

<table>
<thead>
<tr>
<th>N</th>
<th>1,433,120</th>
<th>Quartile 1</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>37.75</td>
<td>Quartile 2</td>
<td>34</td>
</tr>
<tr>
<td>Mode</td>
<td>21</td>
<td>Quartile 3</td>
<td>53</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>22.25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.3. ANALYSIS OF TIME DIFFERENCE BETWEEN BOARDINGS

The maximum percentage of transfer time difference is reached at the 20-25 minute interval. Almost 10% of transfer passengers transfer (boarding the second bus) within 10 minutes after boarding the first bus. Approximately 50% of all transfer customers transfer within 34 minutes. Only 10% of all transfer passengers transfer between 72 and 89 minutes after boarding the first bus. It would therefore be possible to decrease the 90 minute cut-off point to a lower value without losing too many occurrences. The problem lies with the ratio of false negatives and false positives. Cutting off the transfer journeys that in fact took longer than the cut-off point will result in a bias of the analysis especially when trying to use the study to reduce the journey time of longer trips.

5.3.1 Relationship between Time Difference of A1/B1 and Boarding Hour at A1

The sunflower scatterplot (see Figure 5.2) shows the relationships of the time difference between the boardings A1/B1 and the hour of boarding for the week from the 12/04/99 to 16/04/99 on one particular route. The diagram is a method of identifying clusters of passengers’ boarding times and time differences of A1 and B1. The presentation of the data does not favour the extraction of exact values but focuses on clusters that indicate a certain pattern of passenger behaviour. The sunflowers represent one or more cases that have similar data values and therefore
are plotted closely together. The more defined the sunflowers the more cases were identified at that position. Four obvious clusters of boarding time at A1 and time differences between A1 and B1 can be recognised when analysing Figure 5.2. A cluster is a representation of a pattern within the graph and is mostly present where the graph peaks.

![Figure 5.2: Sunflower Scatter Diagram of Time Difference and Hour of Boarding](image)

Cluster 1 shows a concentration of transfer journeys during 7:00 and 7:59 and a time difference of A1 and B1 between 55 and 75 minutes. The interpretation of this pattern could be that passengers who know that they have a long journey time leave earlier in order to reach their destination on time. Cluster 2 covers all departure times between 7:00 and 9:59. The time differences between boarding A1 and B1 lies between 12 and 32 minutes. This cluster indicates that a relatively large amount of passengers transfer in the morning peak time 12 to 32 minutes after boarding the first bus. Cluster 3 shows a high concentration of transfers at 11.30. The time difference between boarding A1 and boarding B1 lies between 30 and 50 minutes for this cluster. This may represent passengers whose work commences at 12 pm. Cluster 4 shows a concentration of transfers in the evening peak time. Again, passengers with longer journeys seem to commence their journey slightly earlier.

5.4 Transfer Matrices

This section focuses on the generation of various matrices based on the transfer journey data. Matrices may serve as input for statistical analyses or transport modelling. The analyst can...
generally define the aggregation level as well as the dimension of matrices. This means that parameters such as peak/off-peak time, ticket type or journey date can be used to create a more detailed and focused matrix or one that includes a wider range of records.

This section introduces four different types of two-dimensional matrices including a Coarse Zone Matrix, Transfer Node Matrix, Area Description Matrix and Route Matrix.

Only the coarse zone matrix can be displayed fully within this thesis as it has only 21 possible columns and rows (21 coarse zones). The remaining matrices cannot be displayed due to their extensive size.

### 5.4.1 Transfer by Coarse Zones

The coarse zone matrix is a way of representing trip level data (passenger boarding records) on an aggregate level. The DTO divided the GDA into 21 zones called coarse zones. The coarse zones have mainly been introduced for transport modelling purposes. A description of each of the 21 zones as well as a map can be found in Appendix A and Appendix B respectively.

The 3-dimensional bar chart (see Figure 5.3) shows the number of transfers from the first coarse zone (A1) to the second coarse zone (B1). As determined before most transfers occur within the city centre zone 1. This graph also shows that transferring within the same zone is quite common (1-1, 4-4, 6-6, 12-12 and 14-14). This can be explained by the fact 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. It also shows that a considerable amount of passengers transfer in suburban areas which is indicated by the smaller clusters. The label ‘0’ on the y-axis and x-axis represent boardings for which the location of the boarding zones was unknown.

Most of the transfer passengers transfer from A1 to B1 in zone 1 (City Centre) (863,832 passengers, 60.28%). The least number of transfer passengers transfer in zones 2 (1,149), 9 (18), 16 (33) and 17 (1,168). Zone 2 is close to city centre but very small whereas zones 9, 16 and 17 are further away from Dublin city centre and are not as frequently served by the bus operator.

The spatial graph (see Figures 5.4) shows a map of Dublin and its respective coarse zones. The spatial map visualises the numbers of transfer journeys originating at the start zone of A1. Almost 77% of all transfer journeys started outside the city centre zone 1 (see Figure 5.3. Each of the zones bordering the city centre zone 1 (apart from zone 2) is responsible for 4.27% to
10.70% of transfer journeys. Furthermore coarse zones 7, 8, 11, 12 and 14 contribute to the total number of transfer journeys with 4.27%, 4.39%, 6.57%, 10.12% and 8.26% respectively. This shows that many passengers who live further away from the city centre have to transfer in order to reach their final destination.
5.4. TRANSFER MATRICES

Table 5.3: Total Number of Transfers by Zones at A1 and B1

<table>
<thead>
<tr>
<th>Zone</th>
<th>Total</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>856,642</td>
<td>64.28%</td>
</tr>
<tr>
<td>2</td>
<td>1,138</td>
<td>0.09%</td>
</tr>
<tr>
<td>3</td>
<td>58,631</td>
<td>4.40%</td>
</tr>
<tr>
<td>4</td>
<td>77,301</td>
<td>5.80%</td>
</tr>
<tr>
<td>5</td>
<td>49,331</td>
<td>3.70%</td>
</tr>
<tr>
<td>6</td>
<td>87,150</td>
<td>6.54%</td>
</tr>
<tr>
<td>7</td>
<td>11,773</td>
<td>0.88%</td>
</tr>
<tr>
<td>8</td>
<td>16,797</td>
<td>1.26%</td>
</tr>
<tr>
<td>9</td>
<td>17</td>
<td>0.00%</td>
</tr>
<tr>
<td>10</td>
<td>30,293</td>
<td>2.27%</td>
</tr>
<tr>
<td>11</td>
<td>75,118</td>
<td>5.64%</td>
</tr>
<tr>
<td>12</td>
<td>2,513</td>
<td>0.19%</td>
</tr>
<tr>
<td>13</td>
<td>49,478</td>
<td>3.71%</td>
</tr>
<tr>
<td>14</td>
<td>4,743</td>
<td>0.36%</td>
</tr>
<tr>
<td>15</td>
<td>32</td>
<td>0.00%</td>
</tr>
<tr>
<td>16</td>
<td>971</td>
<td>0.07%</td>
</tr>
<tr>
<td>17</td>
<td>10,821</td>
<td>0.81%</td>
</tr>
</tbody>
</table>

(a) Total Number of Transfers at B1 (b) Total Number of Transfers at A1

5.4.2 Transfer Node Matrix

A transfer node identification matrix shows the volumes of transfers at each transfer node. The nodes are defined by route ID and stage 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/A2 is a transfer node. The significance of this transfer node depends on the daily volume of transfer passengers. The size of the matrix is the square of all bus stops. 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 on a micro level. For example, is there a need for orbital or cross-city routes where currently only arterial routes exist? It can further be used to identify transfer nodes and their corresponding passenger volumes. The main purpose of such a matrix is for bus planning and policy making.

Table 5.4 shows the most frequent transfer nodes. The table shows Route A/Stage A which is the spatial identifier of the first boarding and Route B/StageB which is the spatial identifier of the second boarding. The transfer column represents the total number of transfers recorded for the particular transfer nodes. Stage 25 in the Stage B column frequently occurs.

5.4.3 Transfers by Area Description

The area description matrix is a way of representing trip level data with regard to the location where the passenger boarded the first time (A1) and the location where the person boarded the second time (B1). The city was divided into 131 different suburban/city centre area descriptions.
Table 5.4: Most Frequent Transfer Nodes by Boarding Locations

<table>
<thead>
<tr>
<th>Route A</th>
<th>Stage A</th>
<th>Route B</th>
<th>Stage B</th>
<th>Transfers</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>65</td>
<td>39</td>
<td>25</td>
<td>1,349</td>
</tr>
<tr>
<td>40</td>
<td>15</td>
<td>10</td>
<td>25</td>
<td>912</td>
</tr>
<tr>
<td>77</td>
<td>54</td>
<td>75</td>
<td>22</td>
<td>720</td>
</tr>
<tr>
<td>10</td>
<td>65</td>
<td>130</td>
<td>75</td>
<td>671</td>
</tr>
<tr>
<td>40</td>
<td>15</td>
<td>77</td>
<td>25</td>
<td>643</td>
</tr>
<tr>
<td>77</td>
<td>54</td>
<td>123</td>
<td>66</td>
<td>634</td>
</tr>
<tr>
<td>10</td>
<td>65</td>
<td>25A</td>
<td>25</td>
<td>629</td>
</tr>
<tr>
<td>10</td>
<td>65</td>
<td>37</td>
<td>25</td>
<td>611</td>
</tr>
<tr>
<td>10</td>
<td>65</td>
<td>51B</td>
<td>25</td>
<td>606</td>
</tr>
<tr>
<td>10</td>
<td>65</td>
<td>66</td>
<td>25</td>
<td>587</td>
</tr>
<tr>
<td>10</td>
<td>65</td>
<td>77</td>
<td>25</td>
<td>581</td>
</tr>
<tr>
<td>39</td>
<td>56</td>
<td>10</td>
<td>25</td>
<td>577</td>
</tr>
<tr>
<td>90</td>
<td>75</td>
<td>10</td>
<td>25</td>
<td>535</td>
</tr>
<tr>
<td>10</td>
<td>65</td>
<td>42</td>
<td>75</td>
<td>529</td>
</tr>
<tr>
<td>40</td>
<td>15</td>
<td>7</td>
<td>25</td>
<td>505</td>
</tr>
<tr>
<td>39</td>
<td>56</td>
<td>7</td>
<td>25</td>
<td>489</td>
</tr>
<tr>
<td>40</td>
<td>15</td>
<td>51B</td>
<td>25</td>
<td>484</td>
</tr>
</tbody>
</table>

Transfers are examined in this section in relation to where they were made and in the next section in relation to the number of transfers made between routes. It is possible to create a matrix that shows the location of boardings of A1 and B1. 131 different city areas were identified throughout the analysis. The size of the matrix is M(131 x 131). The results of this analysis indicate that almost all transfer journeys are taking place on radial arterial routes. The question as to whether some of the transfer journeys could be replaced by an orbital route arises. This will be analysed in greater detail in Chapter 7 where it is described how the analysis of O/D matrices can be used to justify (or not) new routes.

The matrix includes all possible combinations of areas which may explain some of the small numbers or no numbers of transfer journeys. The matrix was populated using transfer journey data over the four month period. Although this matrix includes the data of the entire four month period it could also be created focusing on a smaller period of time (month, week, day and hour). The matrix shows the area of the first boarding which is mapped to the location of the second boarding. Therefore it facilitates the extraction of transfer journey numbers for each particular area combination pair.

5.4.4 Transfer Route Matrix

The transfer route matrix is another type of matrix that can be developed to show the transfer volume of passengers that occurred between all combinations of routes. The sparse matrix shows Route ID of the first boarding and the Route ID of the second boarding (Matrix(First
Boarding Route ID’ x ‘Second Boarding Route ID’)). This information can be used for further statistical analysis or for transport modelling purposes. 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.

Table 5.5 shows a small subset of the matrix as the entire matrix cannot be graphically displayed in this thesis due to its size.

<table>
<thead>
<tr>
<th>First Boarding Route (A)</th>
<th>1</th>
<th>10</th>
<th>101</th>
<th>102</th>
<th>103</th>
<th>105</th>
<th>11</th>
<th>111</th>
<th>114</th>
<th>115</th>
<th>116</th>
<th>11A</th>
<th>11B</th>
<th>Etc…</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td></td>
<td>2</td>
<td>161</td>
<td>10</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>101</td>
<td></td>
<td>19</td>
<td>9</td>
<td></td>
<td>13</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>102</td>
<td></td>
<td>131</td>
<td>2</td>
<td></td>
<td></td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>103</td>
<td></td>
<td>8</td>
<td>5</td>
<td>69</td>
<td></td>
<td></td>
<td>17</td>
<td>3</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>105</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1,382</td>
<td>31</td>
<td>24</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>111</td>
<td></td>
<td>13</td>
<td></td>
<td>14</td>
<td>1</td>
<td></td>
<td></td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>114</td>
<td></td>
<td>51</td>
<td>35</td>
<td>6</td>
<td></td>
<td>16</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>115</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>116</td>
<td>4,538</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>11A</td>
<td>212</td>
<td>5</td>
<td>6</td>
<td></td>
<td></td>
<td>280</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>11B</td>
<td>765</td>
<td>13</td>
<td>3</td>
<td></td>
<td></td>
<td>782</td>
<td>18</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>…</td>
</tr>
<tr>
<td>Etc…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

The route number (i) of the first boarding (A) is arranged vertically whereas the horizontal route numbers (j) define the second boarding (B). As mentioned earlier it was assumed that a transfer was made between 2 distinct routes and that a passenger could not make a transfer from a particular route to the same route. Therefore the diagonals on this matrix are zero. The size of the entire matrix is 187x187 due to 187 different bus routes identified in the database.

For the purpose of this study it has been decided to extract all route interchanges (defined by $A_i$ and $B_j$) where the number of boardings is over 1000. This resulted in 161 different route interchanges. As shown in Table 5.6 the most frequent route interchange has been recorded on journeys involving route 39 as route A and route 10 as route B. 4,538 passengers transferred from 39 to 10 while 4,113 passengers transferred from 10 to 39. This and other combination pairs suggest that there is a pattern emerging between transfer combinations in one direction and transfer journeys in the opposite direction with regard to similar number of journeys undertaken. This will be more fully analysed in Section 5.10 where the symmetry of the network will be explored.
5. TRANSFER JOURNEY ANALYSIS

Table 5.6: Most Frequent Route Combinations of Transfer Passengers

<table>
<thead>
<tr>
<th>1st Boarding (A)</th>
<th>2nd Boarding (B)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>10</td>
<td>4,538</td>
</tr>
<tr>
<td>10</td>
<td>39</td>
<td>4,113</td>
</tr>
<tr>
<td>10</td>
<td>46A</td>
<td>3,347</td>
</tr>
<tr>
<td>46A</td>
<td>10</td>
<td>3,219</td>
</tr>
<tr>
<td>18</td>
<td>77</td>
<td>3,036</td>
</tr>
<tr>
<td>77</td>
<td>123</td>
<td>2,832</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>2,794</td>
</tr>
<tr>
<td>78A</td>
<td>18</td>
<td>2,609</td>
</tr>
<tr>
<td>77</td>
<td>18</td>
<td>2,591</td>
</tr>
<tr>
<td>17A</td>
<td>41</td>
<td>2,569</td>
</tr>
<tr>
<td>41</td>
<td>17A</td>
<td>2,559</td>
</tr>
<tr>
<td>77A</td>
<td>77</td>
<td>2,483</td>
</tr>
</tbody>
</table>

5.5 Analysis of Transfer Journey by Day of Week

This analysis explored the proportions of transfer and single journeys allotted to each day. There are similar percentages of transfers for the days Tuesday (18.0%), Wednesday (18.2%), Thursday (19.3%) and Friday (17.8%). Mondays have a slightly smaller patronage with regard to transfer journeys (15.0%). Sunday (4.1%) and Saturday (7.6%) recorded far fewer transfer journeys. Single journeys scored approximately the same ratios.

5.6 Transfer Analysis by Ticket Type

There were 69 different known types of tickets in use during the years 1998 and 1999. Each boarding record includes detailed information about the type of ticket. Some of the ticket ID’s found in the records do not correspond to the available description. These records were classified as 'Unknown'. This information allowed a more detailed analysis of the various ticket types used for single and transfer journeys.

Almost 54% of all transfer journeys are validated using a 'Weekly Adult Cityzone’ or 'Weekly Student Cityzone’ ticket. The same tickets are only responsible for 32.66% of single journey validation. The third most used ticket for transfer journeys was the '3-Day Bus’ ticket which was responsible for almost 16.90% of transfer journey validations compared to only 8.77% for single journey validations. The four most commonly used tickets for transfer journeys were responsible for 76% of all validations.

The ticket types that differ most between transfer and single journeys are

- Weekly Adult Cityzone - (Transfer: 35.09%, Single: 17.80%);
5.7. TICKET TYPE CATEGORY

- 3 Day Bus - (Transfer: 16.90%, Single: 8.77%);
- Adult 2-Journey (3 Stages) - (Transfer: 1.02%, Single: 5.03%);
- Schoolchild 2-Journey - (Transfer: 0.81%, Single: 8.90%);
- Adult 2-Journey (7 Stages) - (Transfer: 0.43%, Single: 6.23%).

Generally, one could assume that it is more economical to purchase a weekly adult ticket when transfer journeys have to be made on a regular basis. A transfer journey consists of at least two single journeys and is therefore more expensive when paid separately without having a daily, weekly or monthly ticket. The difference between passengers using the weekly adult ticket for transfer journeys (35.09%) compared to single journeys (17.80%) could therefore be explained by assuming that the passengers gain a financial advantage. The single fare of 10 short inner city journeys (e.g., two per working day of the week) would be less than the cost of a weekly adult ticket. The weekly ticket is therefore only economical when more than 10 single journeys or 5 transfer journeys are made per week or the length of the journey exceeds the cheaper inner city tariffs. School children seem to make less transfer journeys than single journeys when comparing the difference of transfer journeys and single journeys of school children (0.81% and 8.72% respectively). It could be assumed that school children live close enough to their school and therefore do not require a transfer to reach the destination. The adult 2-journey tickets with 3 and 7 stages were used more commonly for single journeys. Many transfer journeys exceed 7 stages and these two tickets would therefore not be sufficient to serve the purpose.

5.7 Ticket Type Category

For the purpose of this section all ticket types have been assigned to one of six passenger categories: Child, Student, Adult, Family, Pensioner and Unknown. The last category is for passenger boarding transaction records where the ticket type is unknown and therefore no category could be assigned. The analysis focuses on the months April, May, September and October (1999) and differentiates between single journeys and transfer journeys.

Comparing the percentages between single and transfer journeys the following could be determined: adults are responsible for 67.6% of all transfer journeys and for 56.6% of all single journeys. The numbers of students on the other hand do not support this trend. The numbers of transfer journeys (27.9%) and single journeys (27.3%) are almost identical. There is no difference between the adult and student single cash fare. However, students can purchase
reduced weekly or monthly tickets which make their use more economical than for adults. This may explain the similarities between single and transfer journeys of students and also the differences between the total numbers of ticket usage by students (almost identical) and adults (difference of 11.0%). The 'Child' category shows the largest difference in this comparison. Approximately 1% of all transfer journeys are made by children compared to 10.5% in the single journeys category. This suggests that most children do not have to transfer to reach their final destination.

5.8 Ticket Duration Period

A new data attribute called 'Duration' was added to the database which categorises the various ticket types into the length of the ticket validity. 1 Day, 1 Journey, 2 Journey, 10 Journey, 3 Day, 7 Day, 14 Day, 21 Day, 28 Day, 30 Day, 365 Day, and Unknown are the categories that determine the duration of the ticket validity. The category 'Unknown' replaced the validated boardings for which the ticket type and therefore the ticket validity period could not be determined. The purpose of this analysis is to investigate the use of the various ticket durations which may be useful for fare planning and policymaking decisions.

The most commonly used ticket validity period of transfer journeys is a 7 day ticket (59.2%). This is followed by a 3 Day ticket (16.9%), a 30 Day/monthly ticket (9.9%), 1 Day ticket (4.1%) and a 2 Journey ticket (2.6%). The remaining ticket durations are below 1.6% and therefore build the minority of ticket validation periods. Single journey passengers have a different pattern with regard to their ticket duration. This is mainly due to the 2-Journey ticket which was responsible for 26.8% of single journeys and only 2.6% for transfer journeys.

5.9 Time Analysis

This type of analysis aims to contribute to the understanding of the differences between transfer and single journeys. When attempting to optimise routes in such a manner that waiting times for transfer journey passengers are minimised one has to know the peaks and off/peaks of such journeys. Figure 5.5(a) displays the number of transfer journey boardings for each day of the week mapped to the hour of boarding. Figure 5.5(b) displays the number of single journeys of each day mapped to the hour of boarding. Both figures aim to display the peak and off-peak periods. The study was carried out over a one week period in October 1999. The analysis is based on boarding records of the entire network of that particular week. The single journey
chart does not include the individual journeys of the transfer journeys. There were 105,554 transfer journeys and 471,211 single journeys recorded throughout the week in October.

Although the general progression of boarding times creating the curves in Figure 5.5 are very similar, the small changes in the number of boardings are very significant. The morning peak of the transfer journeys is 7.00 to 8.00 whereas the single journey morning peak is from 8.00 to 9.00. However, Figure 5.5(b) clearly shows a high number of boardings between 7.00 and 8.00. The evening peak of the transfer journeys is from 16.00 to 17.00 while the peak time of single journeys lies between 17.00 and 18.00. The data recorded and displayed in the two diagrams suggest that the peak times are one hour earlier for transfer journeys than they are for single journeys. This may be related to longer total journey times of transfer journeys.

![Figure 5.5: Patterns of Boardings by Hour (Week 18/10/99 - 24/10/99)](image)

5.10 Symmetry Analysis of Transfer Journeys

5.10.1 Introduction and Definitions

The assumption that public transport networks have a symmetric behaviour is often used in models or algorithms that focus on public transport issues. This assumption is introduced to simplify models for ease of calculation. The assumption consists of the idea that each journey in one direction also has a corresponding journey in the opposite direction (return journey). The assumption is mostly used to justify the analysis in the first place or just to simplify the model or algorithm. That the assumption is not exactly true is widely accepted and has also been shown by Navick and Furth (2002). It is however still used and considered as a valuable assumption for estimating numbers of alightings on routes or for estimating passenger miles.
It can be argued that for regular commuting passengers those who carry out a transfer journey in one direction will also transfer for the return journey. If this could be assumed then the total number of transfer journeys with one combination of routes (e.g. route combination $R_1/R_2$) will be similar to the total number of transfer journeys with the opposite combination of routes (e.g. route combination $R_2/R_1$). The analysis also considers single journeys where the number of passengers per route in one direction is compared to the number of passengers in the opposite direction ($R_{1Inbound}$ vs. $R_{1Outbound}$). Various sparse matrices are analysed and the level of symmetry is determined.

Perfect symmetry is defined as the situation where both route combination pairs have exactly the same number of transfers (e.g. $R_1/R_2$ recorded 1,200 transfer journeys and $R_2/R_1$ recorded 1,200 journeys too). Realistically this will only be the case for a few route combination pairs from an entire network. This raises the question to what degree one might expect the journeys in a transport network to be symmetric. This section attempts to answer this question. A generic equation is proposed that quantifies the ‘Degree of Symmetry’ to a system, route or route segment level. A similar set of equations is introduced for single journeys.

We define a route pair $(i, j)$ to be symmetric if

$$F_{ij} = F_{ij}^T$$  \hspace{1cm} (5.10.1)

where

- $F$  
  Transfer route interchange matrix,
- $F^T$  
  Transpose of $F$,
- $i$  
  Route number of first boarding,
- $j$  
  Route number of second boarding.

$F_{ij}$ represents the transfer journey numbers for a particular route interchange combination (e.g. Routes 10 and 39) while $F_{ij}^T$ is the number of transfer journeys that were recorded in the opposite direction of the route interchange combination (e.g. Routes 39 and 10).

The aim is to provide a measure of symmetry when analysing or comparing routes or route segments. Using the symmetry assumption is one possibility to obtain Origin/Destination information of public transport passengers. However some routes do not have symmetry with regard to journey numbers as shown by this section and the assumption of symmetry would therefore bias or even misrepresent the results. The measures developed in this section should provide an
indication of the bias caused by the non-symmetric characteristic of travel on a public transport network.

As the future direction of this project is concerned with the extraction of Origin/Destination pairs the proposed method is one possible way to attempt this. Some research studies have used the assumption of network symmetry (Navick and Furth, 2002; Richardson, 2003). Both studies used the assumption of symmetry in their method to estimate performance measures based on electronic fare collection and automatic passenger counter (APC) data. Navick and Furth (2002) assumed that 'the boarding pattern for a route in one direction is equivalent to the alighting pattern in the opposite direction'. This assumption was tested with APCs that counted boardings and alightings at each stop. A Kolmogorov-Smirnov test was applied to the dataset. The test statistic was created by comparing the cumulative distribution of eastbound boardings with westbound alightings, with the westbound alighting allocated to the eastbound stops. The statistic was calculated for ten different routes. The test indicated an absence of perfect symmetry as 9 out of 10 routes did not provide statistical evidence to accept the hypotheses. The authors then simply compared the differences of boarding and alighting in percent. All but 2 routes showed less than 15% difference in boarding/alighting patterns. It is however noteworthy that the tests were carried out on single routes and not route combinations.

Other research (Richardson, 2003) applies the assumption to estimate average distance travelled. Over a period of 52 days each passenger boarding any of the 38 routes was asked for his/her destination. The collected data were then compared to the estimates set by the assumption: passenger numbers in one direction of a particular route are equal to the boarding numbers in the opposite direction. Both studies concluded that the assumption of symmetric travel patterns on most routes can be applied although details of routes and data should be checked for each route.

Horowitz and Patel (1999) presented trip tables for small urban areas. A method for quick response travel forecasting included the fact that the potential for travel has symmetry. Although this paper works with areas rather than routes their point is still valid for this section. It suggests that the trip opportunities from one area $i$ to area $j$ should be equal to the number of trip opportunities from area $j$ to area $i$. The trip opportunities in one direction are equal to the opportunities in the opposite direction in the urban bus public transport network presented in this paper.
5.10.2 Initial Analysis of Symmetry Patterns

In preparation, prior to the development of the degree of symmetry equation, it was necessary to test whether the transfer journey numbers for each route combination pair (e.g., 39/10) were similar to the corresponding numbers of the inverse direction of the same combination pair (e.g., 10/39). A subset of these results is shown in Figure 5.6. This figure shows that there is an emerging pattern between two route combination pairs by displaying and comparing the total number of transfers of each combination pair. For example, where routes $i = 39$ and $j = 10 \Rightarrow F^{(39, 10)} = 4,538$ and $F^{(10, 39)} = 4,113$. This means that 4,538 passengers transferred from route 39 to route 10 and 4,113 passengers transferred from 10 to 39. Although there is a difference of 425 transfers (9.3%) the two numbers are still relatively close. The closer the numbers are to each other the higher the degree of symmetry for that combination pair. E.g. 10/39 - 39/10 labels a pair of bars; the first one with 10/39 and the second bar with 39/10. The horizontal bar chart clearly indicates the tendency to symmetric behaviour of the route interchanges displayed in the graph. To establish this fact more quantitatively an analysis was needed to determine the degree of symmetry in a network including different parameters (e.g. transfer journey attributes such as date, time or specific routes).

![Figure 5.6: Relationship between Route Combination Pairs](image)

5.10.3 Development of the 'Degree of Symmetry' Equation

The aim of this section is to see if there is a degree of correlation between the total number of transfers in one direction and the total number of transfers in the opposite direction ($R_1/R_2$ and...
Evaluating each route combination pair individually showed that there is a relatively high degree of similarity in terms of passenger volume between most of the transfer route pairs. The next step was to show that this is true for all significant transfer route pairs of the entire transport network by calculating the 'Degree of Symmetry' of the entire transfer route matrix. A significant transfer route combination pair is a pair where one of the total numbers of transfer journeys lies above a predefined cut-off-point referred to as parameter \( \alpha \). The cut-off point \( \alpha \) was introduced to exclude combination pairs with a relatively low number of total journeys and to reduce the bias this may cause. The cut-off point is a variable and can be determined by the user. A small number of transfer journey boarding records on a particular route combination may not be significant with regard to the symmetry of the network but may bias the results considerably. For example, the route combination \( R_1/R_2 \) recorded 10 transfer journeys while the combination \( R_2/R_1 \) only recorded 2. In this case, the two numbers 2 and 10 are only fractions of the numbers from other transfer journey boarding records. This is also known as the small sample effect. The degree of symmetry equation focused on not favouring routes that run more frequently which may attract a larger number of passengers. The cut-off point of significance may change with the parameters the transfer matrix is based on. For example, should the cut-off point of a matrix that consists of one day’s transfer journeys be chosen more carefully than for a matrix that includes data of several months due to the differences in journey numbers? As it has been introduced as a parameter of the equation it can be changed at any stage throughout the analysis.

When looking at the symmetry of the entire transport network with regard to transfer journeys the total numbers of a transfer journey combination pair are interchangeable without changing the degree of symmetry as there is no further directional or temporal information attached. For example the combination '39/10' with a total number of 4,538 is interchangeable with the total number of 4,113 which is assigned to the transfer route combination '10/39' without interfering with the symmetry of the transfer route matrix. It was therefore not possible to apply a statistical technique such as a paired t-test or other methods to determine the degree of symmetry of the matrix. A one sampled t-test on the mean differences of each route combination pair showed that \( H_0(\mu_{\text{Diff}} = 0) \) has to be rejected on statistical evidence. The following equation defines the degree of symmetry \( S_{ij} \) for one route interchange \( i/j \) of matrix \( F \):

\[
S_{ij} = 1 - \frac{|(F_{ij} - F_{ij}^T)|}{(F_{ij} + F_{ij}^T)}
\]  

(5.10.2)
For example, the absolute of the transfer numbers of the combination pair 39/10 minus the transfer numbers of the combination pair 10/39 divided by the transfer numbers of both combination pairs and subtracted from one defines the symmetry $S_{39,10}$.

The largest possible value for $S$ can be 1 which therefore means there is perfect symmetry between the transfer journeys going in one direction and the transfer journeys in the opposite direction. As mentioned above, this may happen for some of the transfer route combination pairs but not for the entire transport network. The smallest number for $S$ can be 0 although this is not very likely. It can only occur when the absolute difference between the route combination in one direction (e.g. $R_1/R_2$) and the route combination in the opposite direction (e.g. $R_2/R_1$) is very large. A zero as a result of $S$ may indicate an error in the dataset or refers to routes that only serve one direction. The smaller the degree of symmetry $S$ the less symmetry is present among the transfer route combination pairs. It is worth mentioning that the equation produces a degree of symmetry of 0 if any of the two input values of transfer journeys ($F_{ij}$ and $F^T_{ij}$) are 0 independent of the non-zero term. It is therefore recommended to set the cut-off point $\alpha > 0$.

Equation 5.10.3 shows the calculation of the total symmetry of a number of routes or the entire network. The cut-off point $\alpha$ has to be chosen in advance to define the significance level of particular route interchanges. The formula states that as long as one value of the two transfer journey numbers ($F_{ij}$ or $F^T_{ij}$) is above the cut-off point $\alpha$ it has to be included as a route interchange. The equation states that the symmetry $S$ is the sum of all route combination pair symmetries divided by the total number of observations where $F_{ij} > \alpha$ or $FT_{ij} > \alpha$ has to be adhered.

$$ S = 1 - \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{|F_{ij} - F^T_{ij}|}{(F_{ij} + F^T_{ij})}}{n(n-1) \alpha} \quad F_{ij} > \alpha \vee FT_{ij} > \alpha \quad (5.10.3) $$

where

$S$ Degree of Symmetry,
$F$ Transfer route interchange matrix,
$F^T$ Transpose matrix of $F$,
$i$ Route number of first boarding,
$j$ Route number of second boarding,
$\alpha$ cut-off point - Total number of transfers at route interchange,
$n$ Number of routes,
5.10.4 Results of Degree of Symmetry Equation for Transfer Journeys

Three main matrices were generated in order to test Equation 5.10.3. The first matrix showed the total number of transfer journeys for the morning peak period (7.00 - 9.00), the second matrix showed the total number of journeys for the evening peak period (16.00 - 18.00) and the last matrix showed the total number of all transfer journeys. All three matrices were based on a dataset consisting of transfer journey records over a four month period. Due to the size of the matrices it was not possible to display them in this thesis. A subset of the matrix is however presented in Table 5.5. It was expected to get a relatively high degree of symmetry for matrix 3 as it showed all transfer journeys of the entire period of time. The proposed argument was that all transfer journeys in one direction will also occur in the opposite direction. The degree of symmetry of matrices 1 and 2 was expected to be very low because of the shift of morning and evening peak time journeys.

Table 5.7 shows the results of the degree of symmetry equation after it was applied to all three matrices. It further shows the disruptive effect of including small routes in the symmetry analysis (small sample effect).

Table 5.7: Results of Degree of Symmetry

<table>
<thead>
<tr>
<th>Cut off point ( c )</th>
<th>Number of Journeys</th>
<th>Route Interchanges</th>
<th>Degree of Symmetry</th>
<th>Number of Journeys</th>
<th>Route Interchanges</th>
<th>Degree of Symmetry</th>
<th>Number of Journeys</th>
<th>Route Interchanges</th>
<th>Degree of Symmetry</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>295,061</td>
<td>12,998</td>
<td>0.366</td>
<td>275,513</td>
<td>13,586</td>
<td>0.439</td>
<td>1,432,882</td>
<td>34,225</td>
<td>0.573</td>
</tr>
<tr>
<td>10</td>
<td>276,190</td>
<td>6,441</td>
<td>0.492</td>
<td>255,030</td>
<td>6,450</td>
<td>0.624</td>
<td>1,411,768</td>
<td>10,999</td>
<td>0.754</td>
</tr>
<tr>
<td>50</td>
<td>206,638</td>
<td>2,230</td>
<td>0.536</td>
<td>175,727</td>
<td>1,938</td>
<td>0.686</td>
<td>1,311,152</td>
<td>5,684</td>
<td>0.833</td>
</tr>
<tr>
<td>100</td>
<td>144,249</td>
<td>990</td>
<td>0.537</td>
<td>114,228</td>
<td>802</td>
<td>0.680</td>
<td>994,465</td>
<td>2,157</td>
<td>0.879</td>
</tr>
<tr>
<td>200</td>
<td>80,069</td>
<td>330</td>
<td>0.520</td>
<td>52,763</td>
<td>222</td>
<td>0.674</td>
<td>836,524</td>
<td>1,423</td>
<td>0.890</td>
</tr>
<tr>
<td>300</td>
<td>47,824</td>
<td>136</td>
<td>0.542</td>
<td>23,867</td>
<td>72</td>
<td>0.645</td>
<td>695,768</td>
<td>971</td>
<td>0.894</td>
</tr>
<tr>
<td>400</td>
<td>34,943</td>
<td>82</td>
<td>0.556</td>
<td>13,956</td>
<td>34</td>
<td>0.668</td>
<td>595,288</td>
<td>718</td>
<td>0.905</td>
</tr>
<tr>
<td>500</td>
<td>25,245</td>
<td>52</td>
<td>0.541</td>
<td>8,281</td>
<td>18</td>
<td>0.648</td>
<td>395,528</td>
<td>348</td>
<td>0.910</td>
</tr>
<tr>
<td>600</td>
<td>15,662</td>
<td>28</td>
<td>0.511</td>
<td>2,293</td>
<td>4</td>
<td>0.610</td>
<td>484,733</td>
<td>498</td>
<td>0.911</td>
</tr>
<tr>
<td>700</td>
<td>10,885</td>
<td>18</td>
<td>0.433</td>
<td>2,293</td>
<td>4</td>
<td>0.610</td>
<td>353,667</td>
<td>286</td>
<td>0.915</td>
</tr>
<tr>
<td>800</td>
<td>7,947</td>
<td>12</td>
<td>0.447</td>
<td>1,234</td>
<td>2</td>
<td>0.558</td>
<td>270,775</td>
<td>186</td>
<td>0.914</td>
</tr>
<tr>
<td>900</td>
<td>5,457</td>
<td>8</td>
<td>0.360</td>
<td></td>
<td></td>
<td></td>
<td>317,275</td>
<td>240</td>
<td>0.919</td>
</tr>
<tr>
<td>1,000</td>
<td>4,041</td>
<td>6</td>
<td>0.274</td>
<td></td>
<td></td>
<td></td>
<td>234,649</td>
<td>148</td>
<td>0.926</td>
</tr>
<tr>
<td>1,200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>219,687</td>
<td>134</td>
<td>0.925</td>
</tr>
<tr>
<td>1,300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>191,710</td>
<td>110</td>
<td>0.926</td>
</tr>
<tr>
<td>1,400</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>169,834</td>
<td>92</td>
<td>0.931</td>
</tr>
<tr>
<td>1,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>155,933</td>
<td>82</td>
<td>0.930</td>
</tr>
</tbody>
</table>
The table shows three main columns; one for each matrix symmetry analysis. Each main column is subdivided into three further columns showing the number of transfer journeys, the number of significant route interchanges and the calculated degree of symmetry. The symmetry equation was applied with changing cut-off point $\alpha$ as indicated in the left column of Table 5.7. Analysing the 'All' columns in Table 5.7 indicates a comparatively high symmetry when considering all transfer journeys. The degree of symmetry extends from $S = 0.573$ to $S = 0.930$ for the cut-off points 0 and 1,500 respectively. The cut-off point 0 may not be representative as too many insignificant interchanges bias the result. The degree of symmetry $S$ with regard to the morning peak and evening peak transfer journeys is as expected very low. This suggests that there is no symmetry at all due to the shift in transfer journeys.

Figure 5.7 shows a chart that is based on Table 5.7. It illustrates the degree of symmetry for different cut-off points. The degree of symmetry among all transfer journeys slowly increases as the cut-off point increases. This is mainly due to the exclusion of routes with less passenger boardings. It seems that the larger the numbers of transfers the more symmetric is the route. As expected, the symmetry $S$ of morning peak (7:00 - 9:00) and evening peak (16:00 - 18:00) is very low and leads to the conclusion that there is no symmetry present. Due to the smaller numbers of transfer journeys in morning and evening peak times (compared to all day figures) the number of transfer journeys are below 1,100 which explains the missing values of $S$ after the cut-off point surpasses 1,000.

![Figure 5.7: Degree of Symmetry Chart - Transfer Journeys](image-url)
5.10.5 Degree of Symmetry of Single Journeys

After analysing the symmetry of transfer journeys, it was decided to analyse the symmetry of all single journeys within the network. This study also concentrated on the months April 99, May 99, September 99, and October 99. A cross-tabulation table was generated showing the routes, the direction of travel (0 - outbound, 1 - inbound) and the relevant frequencies for each instance. The table was based on over 5 million single journeys that took place in the above-mentioned period (see Table 5.5). The route and the direction parameter (0 or 1) were used as combination pair (e.g., Route x in direction 0 or 1 was considered to be the combination pair $R_0$ and $R_1$).

Equation 5.10.4 was used to calculate the degree of symmetry for each individual route ($S_i$). The equation uses the absolute difference between the single journeys in one direction (e.g., outbound - 0) minus the single journeys in the opposite direction (e.g., inbound - 1) and divides this by the sum of both single journey numbers. The calculated value is then subtracted from 1 and results in the degree of symmetry for single journeys of a particular route.

$$S_i = 1 - \frac{|(R_{i0} - R_{i1})|}{(R_{i0} + R_{i1})} \quad (5.10.4)$$

where

- $S_i$: Symmetry of route $i$
- $R_{i0}$: Total number of single boardings on route $i$ in outbound direction (0)
- $R_{i1}$: Total number of single boardings on route $i$ in inbound direction (1)

Equation 5.10.5 calculates the degree of symmetry for the entire network or for a number of predefined routes. The calculation sums the obtained degree of symmetry of each route and divides this by the total number of observations where the cut-off point is adhered. This result is then subtracted from 1 which gives the total degree of symmetry.

$$S_{Total} = 1 - \frac{\sum_{i=1}^{n} \left( \frac{|(R_{i0} - R_{i1})|}{(R_{i0} + R_{i1})} \right)}{N} \quad R_0 > \alpha \text{ or } R_1 > \alpha \quad \text{for all } i \quad (5.10.5)$$

where

- $S_{Total}$: Symmetry of all routes considering both directions
- $i$: Route number
- $N$: Total number of observations where $R_0 > \alpha$ or $R_1 > \alpha$

The results after applying the equation to the single route/direction cross-tabulation table are shown in Table 5.8 and Figure 5.8. The analysis included data of the entire day, morning
peak and evening peak over a four month period. The different time parameters produced a result that supports the assumption that there cannot be symmetry during peak time periods as evidenced by a small value for the degree of symmetry ($S < 0.7$). The results for $S$ for all cut-off points ($\alpha$) during the morning and evening peak periods were between 0.429 and 0.600 which is equivalent to no symmetry. As observed throughout the symmetry analysis of transfer journeys the degree of symmetry increases as the cut-off point $\alpha$ increases. The single journey analysis almost produced a linear graph for the 'All Day' parameter starting at $S = 0.839$ ($\alpha = 0$) and almost linearly progressed to a degree of symmetry of $S = 0.916$ ($\alpha = 5,000$). Again underlining the previous made assumption that larger volume routes have higher symmetry. This can partly be explained by the small sample effect that sometimes occurs for smaller volume routes.

Table 5.8: Results of Degree of Symmetry - Single Journeys

<table>
<thead>
<tr>
<th>Cut off point $\alpha$</th>
<th>Number of Journeys</th>
<th>Significant route interchanges</th>
<th>Degree of Symmetry</th>
<th>Number of Journeys</th>
<th>Significant route interchanges</th>
<th>Degree of Symmetry</th>
<th>Number of Journeys</th>
<th>Significant route interchanges</th>
<th>Degree of Symmetry</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1,191,405</td>
<td>160</td>
<td>0.480</td>
<td>961,623</td>
<td>159</td>
<td>0.527</td>
<td>5,053,299</td>
<td>161</td>
<td>0.839</td>
</tr>
<tr>
<td>10</td>
<td>1,191,405</td>
<td>160</td>
<td>0.456</td>
<td>961,623</td>
<td>159</td>
<td>0.527</td>
<td>5,053,299</td>
<td>161</td>
<td>0.839</td>
</tr>
<tr>
<td>50</td>
<td>1,191,297</td>
<td>157</td>
<td>0.441</td>
<td>961,409</td>
<td>152</td>
<td>0.529</td>
<td>5,053,248</td>
<td>160</td>
<td>0.839</td>
</tr>
<tr>
<td>100</td>
<td>1,190,790</td>
<td>151</td>
<td>0.438</td>
<td>960,959</td>
<td>148</td>
<td>0.530</td>
<td>5,053,098</td>
<td>159</td>
<td>0.839</td>
</tr>
<tr>
<td>200</td>
<td>1,189,909</td>
<td>147</td>
<td>0.435</td>
<td>961,623</td>
<td>144</td>
<td>0.532</td>
<td>5,052,537</td>
<td>157</td>
<td>0.840</td>
</tr>
<tr>
<td>500</td>
<td>1,188,682</td>
<td>143</td>
<td>0.432</td>
<td>958,898</td>
<td>140</td>
<td>0.542</td>
<td>5,051,118</td>
<td>153</td>
<td>0.846</td>
</tr>
<tr>
<td>1,000</td>
<td>1,187,856</td>
<td>141</td>
<td>0.431</td>
<td>954,077</td>
<td>129</td>
<td>0.558</td>
<td>5,049,640</td>
<td>150</td>
<td>0.852</td>
</tr>
<tr>
<td>1,500</td>
<td>1,185,314</td>
<td>137</td>
<td>0.429</td>
<td>951,724</td>
<td>125</td>
<td>0.567</td>
<td>5,048,155</td>
<td>148</td>
<td>0.854</td>
</tr>
<tr>
<td>2,000</td>
<td>1,184,490</td>
<td>136</td>
<td>0.434</td>
<td>949,606</td>
<td>122</td>
<td>0.571</td>
<td>5,047,133</td>
<td>147</td>
<td>0.855</td>
</tr>
<tr>
<td>2,500</td>
<td>1,169,114</td>
<td>133</td>
<td>0.430</td>
<td>944,764</td>
<td>116</td>
<td>0.582</td>
<td>5,042,554</td>
<td>143</td>
<td>0.861</td>
</tr>
<tr>
<td>3,000</td>
<td>1,180,229</td>
<td>132</td>
<td>0.439</td>
<td>938,638</td>
<td>110</td>
<td>0.587</td>
<td>5,039,811</td>
<td>141</td>
<td>0.863</td>
</tr>
<tr>
<td>3,500</td>
<td>1,173,591</td>
<td>127</td>
<td>0.441</td>
<td>936,128</td>
<td>108</td>
<td>0.585</td>
<td>5,033,017</td>
<td>137</td>
<td>0.866</td>
</tr>
<tr>
<td>4,000</td>
<td>1,167,629</td>
<td>122</td>
<td>0.448</td>
<td>929,335</td>
<td>103</td>
<td>0.587</td>
<td>5,038,218</td>
<td>140</td>
<td>0.864</td>
</tr>
<tr>
<td>4,500</td>
<td>1,165,550</td>
<td>121</td>
<td>0.449</td>
<td>926,553</td>
<td>101</td>
<td>0.589</td>
<td>5,033,017</td>
<td>137</td>
<td>0.866</td>
</tr>
<tr>
<td>5,000</td>
<td>1,162,625</td>
<td>119</td>
<td>0.453</td>
<td>924,068</td>
<td>99</td>
<td>0.598</td>
<td>5,031,725</td>
<td>136</td>
<td>0.870</td>
</tr>
<tr>
<td>5,500</td>
<td>1,157,738</td>
<td>116</td>
<td>0.450</td>
<td>917,075</td>
<td>95</td>
<td>0.599</td>
<td>5,031,725</td>
<td>136</td>
<td>0.870</td>
</tr>
<tr>
<td>6,000</td>
<td>1,152,653</td>
<td>113</td>
<td>0.448</td>
<td>915,453</td>
<td>94</td>
<td>0.602</td>
<td>5,029,671</td>
<td>135</td>
<td>0.871</td>
</tr>
<tr>
<td>6,500</td>
<td>1,147,368</td>
<td>110</td>
<td>0.455</td>
<td>906,478</td>
<td>90</td>
<td>0.602</td>
<td>5,024,682</td>
<td>133</td>
<td>0.872</td>
</tr>
<tr>
<td>7,000</td>
<td>1,118,958</td>
<td>98</td>
<td>0.449</td>
<td>879,731</td>
<td>80</td>
<td>0.605</td>
<td>5,006,042</td>
<td>126</td>
<td>0.886</td>
</tr>
<tr>
<td>7,500</td>
<td>1,106,156</td>
<td>93</td>
<td>0.449</td>
<td>861,873</td>
<td>75</td>
<td>0.600</td>
<td>4,989,727</td>
<td>122</td>
<td>0.886</td>
</tr>
<tr>
<td>8,000</td>
<td>968,970</td>
<td>64</td>
<td>0.449</td>
<td>737,633</td>
<td>52</td>
<td>0.600</td>
<td>4,843,455</td>
<td>97</td>
<td>0.916</td>
</tr>
</tbody>
</table>
5.10. SYMMETRY ANALYSIS OF TRANSFER JOURNEYS

Figure 5.8 is the representation of the calculated degrees of symmetry (see Table 5.8). The chart shows the degree of symmetry on the vertical axis and the cut-off point on the horizontal axis.

![Degree of Symmetry Chart](image)

Figure 5.8: Degree of Symmetry Chart - Single Journeys

5.10.6 Statistical Summary of Results

The descriptive statistics provide a more detailed analysis of the degree of symmetry of transfer and single journeys (All Day). The lower and upper bound of the 95% confidence interval of transfer journeys is much smaller than that from the single journey analysis. Single journeys are therefore more symmetric than transfer journeys which should be incorporated when applying the degree of symmetry within other transport models.

Figure 5.9 shows the degree of symmetry plotted against the sum of the route interchange. The result of these plots is an emerging pattern outlining that the larger the count of each route combination pair the higher the symmetry seems to be. This is mainly connected to the small sample effect. The higher variability is due to high count of observations.
Single Journeys

The descriptive statistics shown in Table 5.9 provide a more detailed analysis of the degree of symmetry of single journeys (All Day). The range of lower and upper bound of the 95% confidence interval is much smaller than that from the transfer journey analysis. The median of 0.9300 indicates that ignoring outliers of the lower range of the results would lead to a much higher total degree of symmetry. Outliers could be seen as route combinations with a very low degree of symmetry. This low degree of symmetry may be caused by the nature of the route combination. For example, one of the routes could only run in the morning but not in the evening. The skewness of the sample is -1.904 with a standard error of 0.191. This again underlines that outliers in the lower range may bias the degree of symmetry.

<table>
<thead>
<tr>
<th></th>
<th>Single Journeys</th>
<th>Transfer Journey</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>.8382</td>
<td></td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>.9300</td>
<td></td>
</tr>
<tr>
<td><strong>Minimum</strong></td>
<td>.11</td>
<td></td>
</tr>
<tr>
<td><strong>Maximum</strong></td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td>3.927E-02</td>
<td></td>
</tr>
<tr>
<td><strong>Skewness</strong></td>
<td>-1.904</td>
<td></td>
</tr>
<tr>
<td><strong>Std. Deviation</strong></td>
<td>.1982</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.10(a) presents the data including some of the descriptive statistics in form of a histogram. The skewed graph underlines what has been stated in the previous paragraph. The very low degrees of symmetry may be caused by routes that do not have an equilibrium with regard to trip opportunities in one direction and trip opportunities in the opposite direction.
5.11. STATISTICAL MODEL TO ANALYSE SYMMETRY

Figure 5.10: Histogram of Degree of Symmetry of all Individual Routes

Transfer Journeys

The median of 0.8 (see Table 5.10) is smaller than the median of single journeys (0.93) but still indicates that ignoring routes with a low degree of symmetry would lead to a higher overall degree of symmetry. This is also shown in the skewness statistic. The total number of observation N (see Figure 5.10(b)) is much larger for transfer journeys than for single journeys as the focus was on route combinations, which results in smaller numbers of passengers for most observations.

Table 5.10: Descriptive Statistics of the Degrees of Symmetry of Transfer Journeys

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>.5727</td>
<td>.0</td>
<td>1.00</td>
<td>-0.55</td>
</tr>
<tr>
<td>Median</td>
<td>.6792</td>
<td>.0</td>
<td>1.00</td>
<td>-0.55</td>
</tr>
<tr>
<td>Variance</td>
<td>.1345</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>.3668</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.10(b) presents a histogram of all degree of symmetry results. Figure 5.10(a) and Figure 5.10(b) clearly show the differences between single and transfer journeys.

5.11 Statistical Model to Analyse Symmetry

It was already identified that there the volume of passengers travelling in one direction tends to be similar to the volume of passengers travelling the opposite direction. Figure 5.11 shows this in form of a scatterplot which displays the volumes of passengers for each direction. Outliers can generally be explained by routes that only travel in one direction. What is left to show is
whether this tendency to symmetry is statistical significant or not. This will be attempted in the following paragraphs.

An initial attempt tried to fit a loglinear model to the two-dimensional contingency table (as shown by Agresti (2002); Wrigley (1985)). A goodness-to-fit test failed as the data is a very sparse matrix with many null values (over 45%). The following section focuses on an alternative model that investigates symmetry. As already shown, using absolute differences of transfer journey boardings would misinterpret the dataset. It is therefore not suitable to use the difference values directly as the information of how many boardings were recorded would be lost. Therefore the Equation 5.10.3 is used to calculate the weighted differences which were termed 'Degree of Symmetry'. For this test a cut-off point of 500 was chosen to gain statistical power. This simply means that all transfer journey combination pairs with less than 500 boardings (per month) were omitted.

![Figure 5.11: Scatterplot of Route Symmetry](image)

(a) Single Journeys  
(b) Transfer Journeys

Asymmetry of route combination \( i \) and \( j \) is defined as \( a_{ij} = 1 - S_{ij} \).

The Degree of Symmetry \( S \) was calculated using absolute values although it is assumed that the sign of the transfer route combination differences is random. Therefore the sign of each value of \( S \) was changed randomly.

Figure 5.12 shows a histogram of this dataset.

The Jaque-Bera hypothesis test of normal distribution shows that \( H_0 \) as normal distribution can be accepted with \( p = 0.1506 \) (alpha = 0.01). (Note: \( H_0 \) can not be rejected for all datasets for which the cut-off point is below 500.)
The arithmetic mean and the sample standard deviation are 0.0077 and 0.1228 respectively. This shows that the mean value is close to zero which was as expected as the calculated values were randomly distributed around zero.

The following normal distribution assumptions can be applied (Hazewinkel, 2007):

- $-1\cdot\sigma$ ... $1\cdot\sigma$ = 68% of random values will be in the interval
- $-2\cdot\sigma$ ... $2\cdot\sigma$ = 95% of random values will be in the interval
- $-3\cdot\sigma$ ... $3\cdot\sigma$ = 99.7% of random values will be in the interval

Let us assume that symmetry is to be defined as $S \geq 0.9$. Therefore if for 95% of the measurements the asymmetry value is below 20% then symmetry can be assumed.

Now it can be tested whether the data come from a normal distribution with variance 0.01 ($H_0$) against the alternative that the data come from a normal distribution with a different variance ($\alpha = 0.001$). The test indicates that $H_0$ has to be rejected with a p-value of $1.477\text{e-9}$. When the test is re-run with double variance ($\sigma = 0.14$) $H_0$ cannot be rejected (p-value=0.9998).

Figure 5.12: Histogram of asymmetric values of $s$
When accepting $S = 0.86$ as symmetric then it can be said that 68% of the route combination pairs are symmetric. Also, 95% of all values of $S$ are greater than 0.72.

Furthermore the data was analysed using a paired t-test to show whether $H_0$: mean differences = 0. This was done by extracting a route combination for which the daily boardings in one direction were compared with the boarding in the opposite direction. This was done for a period of 4 months. Two route combination pairs were selected with low symmetry and two with high symmetry. The low symmetry route combination pairs had t-values of -11.114 and 10.369 leading to p-values of 0.00 and 0.00 respectively. The observations of the high symmetry route combination pairs resulted in t-values of -0.032 and -0.047 and a p-value of 0.97 and 0.96 respectively. Although this does not provide proof that the method is correct, it shows that there is no evidence that it is incorrect.

5.11.1 Final Comments to the Degree of Symmetry

It is important not to see the degree of symmetry as a number that can be included into other equations or models. It should be treated as an indicator of the degree of symmetry and might be useful as a way of comparing the symmetry of two routes or networks. The research of the degree of symmetry was carried out for this project with regard to the generation of origin destination (OD) pairs. One option when attempting to generate OD pairs from EFC data is to assume symmetry. However some routes do not have symmetry with regard to journey numbers as shown in this section and the assumption of symmetry would therefore bias or even misrepresent the results. The symmetry equations could be applied using different values for the cut-off point. The OD matrices are then based on routes or route combinations that have a degree of symmetry which is higher than the predefined value (e.g. $S > 0.85$) ensuring that the assumption holds.

Therefore, the proposed set of equations provides a measure of the reliability of the assumption of network/route symmetry for transport analysis.

Passengers often have a route choice using substitutional (or common) routes. These route combinations could in theory also be considered as symmetric. However, substitutional routes were not considered for the study of symmetry. It is expected that the correct inclusion of substitutional routes would yield an increase in degree of symmetry. It is believed that the correct inclusion of substitutional routes at this level requires geographic coordinates.
5.11.2 Summary of Network Symmetry

This section analysed whether there is a symmetric behaviour among passenger journeys at a route level. Almost 8 million records were analysed, and with the use of a set of equations, a definition for degree of symmetry was proposed. The following main conclusions were drawn:

- As shown in the section, electronic fare collection data is a suitable data source to analyse historical passenger journeys on route and system level;
- A one sampled t-test on the mean differences of each route combination pair showed that $H_0(\mu_{Diff} = 0)$ has to be rejected on statistical evidence. Initial analysis however indicates that there is a trend to symmetry between two routes. However, perfect symmetry only exists in a few cases;
- A cut-off point $\alpha$ was introduced to determine and reduce the level of bias caused by insignificant route combinations. The value of $\alpha$ has to be chosen carefully and with respect to the period of time that needs to be analysed;
- The set of equations can calculate the degree of symmetry for particular route combinations or for the entire matrix. The degree of symmetry $S$ reaches from 0 to 1 where 0 defines the lowest value of symmetry and 1 defines the highest degree;
- The assumption that there cannot be symmetry throughout a peak time holds for all tests applied to the matrices. Therefore the assumption of symmetry cannot be used for any period less than one day;
- The degree of symmetry function of single journeys is almost constant, showing $S = 0.839$ for $\alpha = 0$ and $S = 0.916$ for $\alpha = 5,000$. As single journeys only consider one route and not a route combination (as used for transfer journeys) the degree of symmetry is more linear and more independent of $\alpha$ as the total number of journeys increases;
- The equations presented in this section can be used as a tool to determine the degree of symmetry $S$ on a route or system level. It may serve as an indicator of the reliability of the assumption that symmetry exists.

5.12 Summary

This section showed that EFC data can serve as a source and foundation for public transport analysis focusing on performance measures, passenger travel patterns and behaviour and system/operator optimisation. The implementation of a specific developed classification algorithm
facilitated the transfer journey analysis. The algorithm classified individual boardings into single journey and transfer journey groups.

The research carried out with regard to the network symmetry identified a gap in the literature. An equation was developed that defined the level of symmetry in a network or network segment. This level of symmetry is more an indicator than an exact percentage. This equation results in a number between 0 and 1 whereby 1 is the highest level of symmetry and 0 is the lowest level of symmetry. The experience gained so far lead to the conclusion that everything above 0.85 can be defined as partial symmetric. Anything below 0.85 is less or not symmetric. This analysis was carried out for transfer journeys where the 0.85 degree of symmetry is reached when the cut-off point is 100 or above. The degree of single journeys reaches from 0.839 (no cut-off point) to 0.196 (cut-off point = 5000). The range of lower and upper bound of the 95% confidence interval for single journeys is much smaller than that from the transfer journey analysis. The median of 0.9300 indicates that ignoring outliers of the lower range of the results would lead to a much higher degree of symmetry;

The following chapter focuses on the core of this thesis which is the identification of passenger destinations at trip level. An algorithm is proposed that estimates destinations of public transport passengers using historical EFC data.
Chapter 6

Design and Development of the Origin/Destination Algorithm

6.1 Introduction

The aim of this chapter is to describe the methodologies and techniques that were applied to extract trip level OD information from an entry-validation EFC system using a novel algorithm; see Hofmann and O'Mahony (2005c,d). The emphasis lies on 'trip-level' OD pairs which means OD pairs that can be directly associated to an individual passenger. Although for this research project most demographic information of passengers is not available, in more recent datasets such information about the actual passenger is often attached to the transactional boarding record, e.g. using smart cards (see Bagchi and White (2005) and Trpanier and Agard (2007b)). The aim of this novel algorithm is therefore not to create an aggregate OD matrix for the network but OD information of individual boardings.

The previous chapters outlined the data structure and the extensions of the database with location parameters (coarse zone and area description, see Section 4.4) and the transfer journey identifier (see Section 4.5). Further, the need and importance of Origin/Destination (OD) information for public transport planning was outlined (see Section 2.9). Throughout the design and development of this algorithm the following parts needed to be addressed:

- Introduce the concept and all associated assumptions that will lead to the actual development of the OD extraction algorithm.
- Develop an algorithm that can determine substitute routes. This is done with a classification algorithm that identifies substitute routes of a network without knowing any details of the network itself;
- Develop the rules that form the rule base using expert domain knowledge,
6. DESIGN AND DEVELOPMENT OF THE ORIGIN/DESTINATION ALGORITHM

- Develop the actual OD algorithm using Rule Based Reasoning in a multi-iterative approach.
- Validate the results using a different dataset and different approaches.
- Identify the limitations and boundaries of the data and the OD algorithm.

Figure 6.1 graphically shows the structure of the entire algorithm whose aim it is to extract and facilitate the analysis of OD pairs. The 'Preparation Stage' and the 'Implementation and Validation Stage' are described in greater detail throughout this chapter. It also mentions briefly the database update stage. The analysis stage is presented in Chapter 7.

6.2 The Concept

OD information is important for various analyses and planning purposes. However, boarding records stored by an entry only validation EFC system do not reveal the location and time of alighting.

**Assumption:** The location of boarding in the evening can be considered the final destination in the morning (Furth, 2000; Richardson, 2003) and therefore facilitates the extraction of a valid OD pair.
This assumption uses the boarding record attributes *Ticket ID* and *Ticket Type ID* which create a unique identifier for each passenger over a certain period of time (depending on the ticket type). This information can be used to explore the path of the passenger and will serve as input data for the OD extraction algorithm. A travel record is defined as the entire set of journeys of one individual passenger throughout the validity period of the ticket. The algorithm has to incorporate patterns and travel records that at first do not result in a valid OD pair due to the complexity of the travel record. After obtaining the estimated OD pairs of the network throughout a period of one month an analysis will follow that incorporates the new findings. The analysis will aim to extract further information with regard to passenger travel behaviour, operational planning, strategic planning and policymaking.

The assumptions used in this thesis are similar to the assumptions used by Barry et al. (2002). They implemented a similar concept on the data collected by the EFC system of the New York Metro (see Section 3.4). There are, however, many structural differences in terms of available data and different levels of complexity between Metro networks and bus networks. Nevertheless, Barry et al. (2002) provided initial evidence that the assumptions are representative and mirror the results obtained by cordon counts. This was described in greater detail in Section 3.4. Validation of the data with regard to this project will be discussed in Section 6.11.

### 6.2.1 The Main Assumption in Greater Detail

The main assumption can be split into two statements:

1. The boarding of journey 1 in one direction can be assumed to be the destination of journey 2 in the opposite direction;
2. The boarding of journey 2 in one direction can be assumed to be the destination of journey 1 in the opposite direction.

If this is true and it is supported by the data recorded in a travel record then one OD pair results from the two individual single journeys. Figure 6.2 shows a diagram that simplifies the assumption for single journeys. $A_O$ is the origin of journey 1, $A_D$ is the destination of journey 1, $B_O$ is the origin of journey 2 and $B_D$ is the destination of journey 2. The assumption states that $A_O = B_D$ and $B_O = A_D$. Since $A_O$ and $B_O$ are the two known locations and $B_O = A_D$ and $A_O = B_D$ all locations ($A_O$, $A_D$, $B_O$ and $B_D$) are now known. Therefore one OD pair can be extracted: $A_O/B_O$ and $B_O/A_O$. The newly identified OD pairs not only consist of location
parameters but also include all other information that was recorded at the time the passenger boarded such as date, time, route, stop, etc.

Transfer journeys on the other hand add to the complexity due to the two legs that were recorded. This means that four boardings were recorded to represent a return transfer journey. Figure 6.3 shows the stages of a one-transfer return journey consisting of 4 individual boardings (two for each journey – decoded as A, B, C and D). \( A_D/B_O \) and \( C_D/D_O \) are the transfer nodes where the passenger changed bus in order to reach a final destination. Although \( A_D \) and \( B_O \) may have slightly different locations they do not influence the OD pair to be extracted. As stated in Section 4.5 it is assumed that the differences in location of \( A_D \) and \( B_O \) are within walking distance (same applies to \( D_O \) and \( C_D \)).

The assumption that the boarding location of the second journey represents the location of alighting of the first journey can be applied to the transfer journey scenario in a similar fashion as it was applied to the single journey scenario. \( A_O, B_O, C_O \) and \( D_O \) are the known locations. The aim is to determine the locations of \( A_D \) and \( D_D \). \( B_D \) and \( C_D \) can also be determined and become important in the analysis stage when transfer nodes, waiting times and in-vehicle times can be calculated (see Section 7.5). The assumption states that \( D_D = A_O, C_D = B_O, B_D = C_O \) and \( A_D = D_O \) which results in the OD pairs \( A_O/D_O, B_O/C_O, C_O/B_O \) and \( D_O/A_O \). Subsequently the OD pairs of interest are \( A_O/C_O \) and \( C_O/A_O \).

Figure 6.2: Visualisation of the Assumption for Single Journeys

![Figure 6.2: Visualisation of the Assumption for Single Journeys](image)

Figure 6.3: Visualisation of the Assumption for Transfer Journeys

![Figure 6.3: Visualisation of the Assumption for Transfer Journeys](image)
6.2. THE CONCEPT

It is important to stress that not all consecutive journeys are return journeys. Naturally not every second journey is a return journey of the first journey. Therefore the need exists to incorporate further decision parameters. Information about direction, date and route of the journey need to be included when deciding whether boarding records actually belong to an OD pair or not. It is further important to differentiate between single and transfer journeys as their data structure is different (2 boardings for single journeys and 4 boardings for transfer journeys).

6.2.2 Trip Chaining

As introduced in Section 3.4 several other efforts sought to extract OD information from EFC datasets. The characteristics of the infrastructure differ greatly between bus and metro networks. The metro network has stations from which a passenger can take several metro lines. Which line is actually boarded cannot be determined with certainty. Bus networks on the other hand are different as the ticket is generally validated on board the bus and the exact route number is stored along with other journey and passenger related information. A further missing data attribute in Metro networks is often the direction identifier. Whereas this is recorded by many bus network EFC systems, the direction cannot always be determined when analysing metro EFC data. Therefore without exit validation on metro networks the data recorded has significantly less semantics than the EFC data from a bus network. One of the main assumptions used by Barry et al. (2002), Zhao (2004), and Trepanier et al. (2007a) assumes that the passenger’s destination of the last trip of the day is the boarding location of the first journey. Trepanier et al. (2007a) incorporates the first journey of the same day and the first journey of the following day to infer the destination of the last journey. This assumption could not be confirmed and proved invalid for the Dublin Bus dataset. The analysis of the next paragraph provides evidence for this. Furthermore, it is noteworthy that this project sees trip chains as linking of activities through the day and not, like indicated in other projects, simply interchange points en route in a ‘one-way’ trip. This is further defined in Bagchi and White (2005).

It could be assumed that at least in most cases if the first journey of the day has the directional attribute of ‘inbound’ then the last journey of the day has to have the directional attribute ‘outbound’ and vice versa (especially in a mostly radial network). Figure 6.4 shows three different scenarios with the directional identifier of each journey. All three scenarios focus on three journeys per day. The scenario (a) and (c) show passenger paths that would be in favour of the assumption. Scenario (b) on the other hand shows a travel path that clearly indicates the problem with the assumption.
Table 6.1 shows a cross-tabulation between the boardings per day per passenger and whether the directional attribute was the same or different for the first and the last journey of the day.

<table>
<thead>
<tr>
<th>Boardings per Day</th>
<th>Direction same</th>
<th>Direction different</th>
<th>Total</th>
<th>True</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N/A</td>
<td>N/A</td>
<td>321,226</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>61,072</td>
<td>275,822</td>
<td>336,894</td>
<td>82%</td>
</tr>
<tr>
<td>3</td>
<td>54,059</td>
<td>74,981</td>
<td>129,040</td>
<td>58%</td>
</tr>
<tr>
<td>4</td>
<td>24,563</td>
<td>74,107</td>
<td>98,670</td>
<td>75%</td>
</tr>
<tr>
<td>5</td>
<td>11,811</td>
<td>19,668</td>
<td>31,479</td>
<td>62%</td>
</tr>
<tr>
<td>6</td>
<td>4,679</td>
<td>9,438</td>
<td>14,117</td>
<td>67%</td>
</tr>
<tr>
<td>7</td>
<td>1,975</td>
<td>3,429</td>
<td>5,404</td>
<td>63%</td>
</tr>
<tr>
<td>8</td>
<td>881</td>
<td>1,619</td>
<td>2,500</td>
<td>65%</td>
</tr>
<tr>
<td>9</td>
<td>370</td>
<td>667</td>
<td>1,037</td>
<td>64%</td>
</tr>
<tr>
<td>10</td>
<td>202</td>
<td>333</td>
<td>535</td>
<td>62%</td>
</tr>
<tr>
<td>&gt;10</td>
<td>210</td>
<td>354</td>
<td>564</td>
<td>63%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>159,822</strong></td>
<td><strong>460,418</strong></td>
<td><strong>941,466</strong></td>
<td></td>
</tr>
</tbody>
</table>

If the directional attribute of the first journey is different to the directional attribute of the last journey then it could be assumed that the assumption of trip chaining might be applicable. If the directional attribute is the same for both journeys the assumption cannot be verified. The percentage in the last column of the table represents the likelihood that the assumption is true (‘direction different’ over ‘Total’). One and two boardings per day are each responsible for one third of all boardings. Two boardings per day show that in 82% of all cases the directional attributes of the first journey is different for the first and last journey of the day. However, when analysing the category of three boardings per day this percentage plummets to 58%. This means that 42% of all passengers that carried out three boardings per day travel in the same direction for the first and last boarding of the day. The data in Table 6.1 show that the assumption is more likely to be correct for an even number of boardings per day. Overall, 25% of all passengers who carry out two or more boardings per day provide evidence against the assumption used.
by Barry et al. (2002), Zhao (2004), Wilson et al. (2005), and Trepanier et al. (2007a). It can therefore be concluded that the trip chaining assumption of the last boarding (see Section 3.4) as it was used for the Chicago, New York city, and Gatineau dataset cannot be used for the EFC data of this project. Initially it was thought that this assumption may only work on metro/heavy rail data. However, a recent paper by Trepanier et al. (2007a) applied this assumption to a bus EFC dataset where it lead to a 20% share of all OD estimations. One explanation could be the type of the network as it serves a population of 240,000 over an area of 598 km$^2$ compared to approximately 115 km$^2$ and 1 million passengers in Dublin.

In conclusion to this section it can be said that the passengers of the Dublin transport network often do not return to their original destination with the last journey recorded. Therefore it was not possible to use the same assumption of trip chaining as was used in other similar projects. It must therefore be concluded that this assumption depends on the structure and characteristics of the public transport network.

### 6.3 The Data Source

The data source used by the algorithm consists of the EFC data which were imported into a database (see Section 4.3). Figure 6.5 shows boarding records of two different passengers. In one case the ticket ID is 3276 and in the other it is 3727.

The following data attributes are included in each file:

- **Boarding ID**
- **Ticket Type ID**
- **Tally ID**
- **Journey Type**
- **Date of Boarding**
- **Time of Boarding**
- **Origin**
- **Destination**
- **Transfer OD Pair**

![Figure 6.5: Data Input Format for the OD Algorithm](image)
Boarding ID is an internal number that serves the unique identification of each of the 46 million boardings stored in the database.

Ticket Type ID is a three digit code representing the type of ticket such as 'Monthly Adult'.

Ticket ID is a unique number that is assigned to each ticket. Figure 6.5 shows boarding records of two different passengers. In one case the Ticket ID is 3276 and in the other it is 3727.

Journey Type represents the identifier whether the boarding was part of a transfer journey or not. '0' represents a transfer journey and '1' represents a single journey. This identifier will be important throughout the OD extraction algorithm.

Date of Boarding specifies the date the boarding took place.

Stage of Boarding is a two-digit number that represents the bus stop on a particular route.

Time of Boarding represents the actual time the passenger boarded the bus.

Route represents the number of the route.

Direction represents the direction of the route. '0' is outbound and '1' is inbound.

Coarse Zone represents aggregated traffic zones. The Greater Dublin Area (GDA) consists of 21 different zones (see Appendix A and Appendix B).

Area Description represents 131 different areas consisting of the names of city centre and urban area names.

Keeping the run-time of the algorithm in mind, the data had to be organised to conform to the requirements of the algorithm. Each iteration of the algorithm is only concerned with one particular passenger. It therefore seemed logical to extract the data by ticket type. The data of each ticket type (56 different ticket types exist) represented one file. Within each file the data were sorted in ascending order by Ticket ID, Date of Boarding and Time of Boarding. Each file therefore consisted of all boardings of a particular ticket type in timely sequential order for each Ticket ID.

Figure 6.5 shows a highlighted single journey records that built a valid OD pair if, and only if routes '83' and '18' are substitutional. The highlighted transfer journey example shows four boarding records which form an OD pair. The passenger travels in the morning from Area 1 to Area 3 transferring in the City Centre South. In this instance the passenger used identical routes and the algorithm does not need to check for substitutional routes. The determination of substitutional routes is therefore important for the overall functionality of the OD extraction algorithm and is explained in Section 6.4.
6.4 Substitutional Routes

6.4.1 Introduction

Each alternative route a passenger can use to get to his/her final destination can be termed as *substitutional route*. It is vital for the successful implementation of the OD extraction algorithm that substitutional routes are incorporated as they are used for classifying journeys into different OD classes which will be introduced in Section 6.5. It is therefore important to understand which routes are substitutional so that the freedom of passengers’ route choice can be incorporated. Figure 6.6 shows a few examples of substitutional routes. In order to reach *Location B* from *Location A* a passenger can use *Route 1* or *Route 3*. Therefore these two routes are substitutional routes. A further characteristic of substitutional routes are that they do not necessarily have to run parallel. In Figure 6.6 *Route 4* does not run parallel along *Route 1*. However due to their common point of intersection at *Location A* and *C* these two routes become also substitutional routes for the purpose of this algorithm as in reality the passengers would have the route choice between the two routes.

![Figure 6.6: Examples of Subroute Scenarios](image)

There are two different ways of generating a list of substitutional routes:

- Obtain the information from the network operator or extract them from route maps;
- Create a heuristic algorithm which estimates substitutional routes based on already recorded EFC data from the network.

The first method seems not only work intensive but also favours human error throughout the definition of these substitutional routes. Many routes may be substitutional for a very short
distance only and therefore could be overlooked. Especially within the city centre area it was identified that each route has multiple other routes that could be taken to reach the passenger’s destination. The second disadvantage is that a manually compiled list is specific to the city for which it was created. To accumulate a complete list for any particular city will therefore take a considerable amount of time, will increase project costs and imply a high rate of error. Further this method consists of the substitutional routes the transport planner believes to be correct but does not necessarily reflect the passenger behaviour.

The second approach on the other hand takes advantage of the passenger boarding data that has already been accumulated. This is based on the premise that substitutional route combinations are used in greater frequency than non-substitutional route combinations. This assumption holds when comparing results from the algorithm with a list of manually compiled substitutional routes. The following sections describe the procedure and validation of dynamically defining substitutional routes using a newly developed identification method.

6.4.2 The Execution

The following steps were carried out to classify whether a route combination is in fact a substitutional route combination or not.

1. The algorithm is based on the premise that substitutional route combinations are used in greater frequency than non-substitutional route combinations. The substitutional route identification algorithm (SRIA) generates a frequency table containing a count of all route combinations that were recorded from single journeys that took place on the same day and in different directions carried out by the same passenger. The sample data shown in Table 6.2 displays only a small subset of the frequencies of route combinations. It shows the route combination consisting of Route 1 (R1) and Route 2 (R2), the number of occurrences and the calculated values for Solution 1 and Solution 2 of the route combinations which will be introduced in point 2 of this section. The data used as learning input consisted of boarding records collected throughout the month of October 1999. For example, the SRIA identified that 776 passengers carried out a potential return journey using the route combination 66 and 66A. The frequency table contains 6,645 different route combination pairs which are used to calculate the sums of all journeys from each route. 353,654 potential passenger route combination pairs were identified;
Table 6.2: Sample Data from the Substitutional Route List

<table>
<thead>
<tr>
<th>Route 1 ( (T_{Ri}) )</th>
<th>Route 2 ( (T_{Rj}) )</th>
<th>Occurrences ( (T_{R_{ij}}) )</th>
<th>Solution 1 ( S_1(R_{ij}) )</th>
<th>Solution 2 ( S_2(R_{ij}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>66 66A 776 5.43 20.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66 66 3182 15.77 50.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66 66B 749 5.28 20.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66 67A 1300 7.84 29.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66 38 17 0.22 0.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66 66A 543 8.74 50.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>etc...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total of 66</td>
<td>-</td>
<td>11,680</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total of 66A</td>
<td>-</td>
<td>3,377</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

2. Two parameters will be the decisive factors in identifying substitutional routes: the number of occurrences of the route combination pair in the frequency table and a calculated weight which is the result of two different approaches. The number of occurrences of a particular route combination pair is not a satisfactory decision factor for defining substitutional routes because it would penalise small and less frequent routes as their total passenger numbers would be smaller. The aim was therefore to include a second decision parameter which contributes to successfully determine substitutional pairs focusing on the reduction of prediction errors. This lead to the development of two different equations which represent a weighted indicator that determines together with the number of occurrences whether the route combination was in fact a substitutional route or not. Two different solutions were developed and tested:

**First Solution**

The aim of this approach is to weight the number of occurrences by the total number of occurrences of each of the two routes. The first set of values is calculated by dividing the total number of occurrences of one particular route combination (e.g. \( R_1 \) and \( R_2 \)) by the total sum of occurrences of the first route and the total sum of occurrences of the second route minus the number of the particular route combination (because it was counted twice). The result is then multiplied by 100 to make it more readable by moving the decimal place to the right (see Equation 6.4.1).

\[
S_1(R_{ij}) = \frac{T_{R_{ij}}}{T_{R_i} + T_{R_j} - T_{R_{ij}}} \times 100 \tag{6.4.1}
\]

where \( S_1(R_{ij}) \) is the calculated weight of Solution 1 of the route combination \( R_{ij} \). \( R_{ij} \) is a route combination with the parameters \( i \) and \( j \) which represent two particular routes. \( T_{R_i} \) and \( T_{R_j} \) are the total number of occurrences of \( R_i \) and \( R_j \) respectively. \( T_{R_{ij}} \) is the
total number of occurrences of the particular route combination pair. In Table 6.2 the Solution 1 weight of the combination $66/66A$ is therefore

$$S_1(R_{66/66A}) = \frac{776}{11,680 + 3,377 - 776} \times 100 = 5.43$$  (6.4.2)

This weighted value is calculated for all identified combination pairs.

The theoretical range of values is $0 < S_1(R_{ij}) \leq 100$. Whereas the lower range of this scale occurs frequently the upper range is a more theoretical value and only occurs when $T_{Rij}$, $T_{Ri}$ and $T_{Rj}$ are of similar value.

Table 6.3 shows the results of a sensitivity analysis of the function when small (1), medium (50) and large (10,000) numbers are used for $T_{Rij}$, $T_{Ri}$ and $T_{Rj}$.

<table>
<thead>
<tr>
<th>$TR_{ij}$</th>
<th>$\Sigma TR_i$</th>
<th>$\Sigma TR_j$</th>
<th>$S_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>10,000</td>
<td>0.009951</td>
</tr>
<tr>
<td>1</td>
<td>10,000</td>
<td>10,000</td>
<td>0.005</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>10,000</td>
<td>0.5</td>
</tr>
<tr>
<td>10,000</td>
<td>10,000</td>
<td>10,000</td>
<td>100</td>
</tr>
</tbody>
</table>

This table shows the behaviour of $S1$ when the values for $TR_{ij}$, $\Sigma TR_i$ and $\Sigma TR_j$ change. Equal value will produce perfect symmetry whereas largely different values will produce a low symmetry value. It is noteworthy that the distribution of the calculated weights is similar to the distribution of the number of occurrences indicating that the sample is representative.

**Second Solution**

The second set of ratios is calculated by dividing the total number of occurrences of one particular route combination (e.g. $R_{66}$ and $R_{66A}$) by the total number of occurrences of $R_{66/66}$ plus the total number of occurrences of $R_{66A/66A}$. $R_{66/66}$ and $R_{66A/66A}$ are both the ideal route combination pairs as the return journey was not taken by a substitutional route.

$$S_2(R_{ij}) = \frac{T_{R_{ij}}}{T_{Ri} + T_{Rj}} \times 100$$  (6.4.3)

where $S_2(R_{ij})$ is the calculated weight of Solution 2 of the route combination $R_{ij}$. $T_{Ri}$ and $T_{Rj}$ are the total number of occurrences of $R_{ri}$ and $R_{jj}$ respectively. $T_{Rij}$ is the total number of occurrences of the particular route combination pair.

Using the data from Table 6.2 the weight of the combination $66/66A$ is calculated as
follows:

\[ S_2(R_{66/66A}) = \frac{776}{3,182 + 543} \times 100 = 20.83 \]  

(6.4.4)

This relative weighted value is calculated for all identified combination pairs. The range of possible values is \( S_2(R_{ij}) > 0 \). A frequency analysis of the calculated weights showed that 99.8% of all calculated values are below or equal to 50. A value of 50 is achieved when the weight is calculated for the same routes (i.e. \( S_2(R_{66A/66A}) \)). A value above 100 is calculated when the number of occurrences of a route combination \( R_{ij} \) is larger than the combined number of occurrences of \( TR_{ii} \) and \( TR_{jj} \). The largest calculated weight for the dataset of this project is 91.89. This was calculated as follows:

\[ S_2(R_{16/16A}) = \frac{2,344}{1,836 + 715} \times 100 = 91.89 \]  

(6.4.5)

In this case \( T_{R16/16A} \) had a total number of occurrences of 2,344. The denominator consisted of \( T_{R16/16} \) and \( T_{R16A/16A} \) which had a total number of occurrences of 1,836 and 715 respectively. This means that the substitutional route combination \( R_{16/16A} \) is more frequently used than the optimal route combinations \( R_{16/16} \) or \( R_{16A/16A} \).

Table 6.4 shows a sensitivity analysis of the function of \( S_2 \) when small (1), medium (50) and large (10,000) numbers are used for \( T_{Rii}, T_{Rij} and T_{Rijj} \).

<table>
<thead>
<tr>
<th>( T_{Rij} )</th>
<th>( T_{Rii} )</th>
<th>( T_{Rijj} )</th>
<th>( S_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>50</td>
<td>1.961</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>10000</td>
<td>0.00995</td>
</tr>
<tr>
<td>1</td>
<td>10,000</td>
<td>10,000</td>
<td>0.005</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>50</td>
<td>98.04</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>10,000</td>
<td>0.498</td>
</tr>
<tr>
<td>50</td>
<td>10,000</td>
<td>10,000</td>
<td>0.25</td>
</tr>
<tr>
<td>10,000</td>
<td>1</td>
<td>1</td>
<td>500,000</td>
</tr>
<tr>
<td>10,000</td>
<td>1</td>
<td>50</td>
<td>19,608</td>
</tr>
<tr>
<td>10,000</td>
<td>50</td>
<td>10,000</td>
<td>99.50</td>
</tr>
<tr>
<td>10,000</td>
<td>10,000</td>
<td>10,000</td>
<td>50</td>
</tr>
</tbody>
</table>

3. After calculating each route combination weight for Solution 1 and Solution 2 it was necessary to determine what values of these parameters classify the route combination pair as substitutional routes. Two cut-off conditions were defined to identify the substitutional routes.

- Minimum number of occurrences. This means that the number of occurrences of a particular route combination has to be greater than the flexible pre-defined value which will be determined in Section 6.4.4;
6. DESIGN AND DEVELOPMENT OF THE ORIGIN/DESTINATION ALGORITHM

- The values of $S_1$ and $S_2$ calculated in step 3 have to be greater than a flexible predefined value which will also be determined in Section 6.4.4.

4. The final list is compiled by ignoring route combinations that are not classified as substitutional routes.

6.4.3 Statistics of Number of Occurrences and Calculated Weights

This section provides details about the descriptive statistics of the 'Number of Occurrences' attribute and of the values calculated with Solution 1 and Solution 2. It is noteworthy that there are no negative values and therefore the values are skewed to the zero mark.

Statistics of the Number of Occurrences Attribute

Table 6.5 shows descriptive statistics of the 'Number of Occurrences' attribute. The distribution is extremely skewed as many route combination pairs do not occur frequently. The algorithm is based on this pattern.

Table 6.5: Descriptive Statistics of the Number of Occurrences Attribute

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>6,645</td>
</tr>
<tr>
<td>Mean</td>
<td>53.22</td>
</tr>
<tr>
<td>Median</td>
<td>4.00</td>
</tr>
<tr>
<td>Mode</td>
<td>1</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>358.47</td>
</tr>
<tr>
<td>Variance</td>
<td>128,498.64</td>
</tr>
<tr>
<td>Skewness</td>
<td>20.072</td>
</tr>
<tr>
<td>Minimum</td>
<td>1</td>
</tr>
<tr>
<td>Maximum</td>
<td>14,185</td>
</tr>
<tr>
<td>Percentiles</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>2.00</td>
</tr>
<tr>
<td>50</td>
<td>4.00</td>
</tr>
<tr>
<td>75</td>
<td>15.00</td>
</tr>
</tbody>
</table>

Figure 6.7 shows the relationship between the number of occurrence attributes and the calculated weights using the equations for solution 1 and solution 2. There is a strong correlation between these two attributes which was expected. The weights only serve as fine tuning of the algorithm so that less frequent routes are not penalised.
6.4. SUBSTITUTIONAL ROUTES

Figure 6.7: Relationships between *Number of Occurrences* and *Weights*

**Statistics of Weights Calculated with Solution 1**

Table 6.6 shows the various descriptive statistics after the weights were calculated for the 6,645 different route combination pairs using *Solution 1*. The range of $S_i(R_{ij})$ is between 0 and 100 with an arithmetic mean of 0.6359. This indicates the skewed distribution of the dataset.

<table>
<thead>
<tr>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Variance</th>
<th>Median</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,645</td>
<td>0</td>
<td>100.00</td>
<td>0.6359</td>
<td>3.342</td>
<td>11.17</td>
<td>0.06</td>
<td>196.935</td>
</tr>
</tbody>
</table>

**Statistics of Weights Calculated with Solution 2**

Table 6.7 shows the various descriptive statistics after the weights were calculated for the 6,645 different route combination pairs using *Solution 2*. The range of $S_2(R_{ij})$ is between 0 and 91.89 with an arithmetic mean of 2.256. This also indicates the skewed distribution of the dataset. The standard deviation (SD) of Solution 2 is 8.591 and therefore larger than the SD of Solution 1 resulting in a greater spread of the calculated weights.

<table>
<thead>
<tr>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Variance</th>
<th>Median</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,645</td>
<td>0.00</td>
<td>91.89</td>
<td>2.2563</td>
<td>8.5909</td>
<td>73.804</td>
<td>0.23</td>
<td>29.140</td>
</tr>
</tbody>
</table>

6.4.4 Finding the Optimal Cut-off Points

This section is concerned with finding the optimal cut-off points. The previous sections calculated the weights which now have to be evaluated. The cut-off points are responsible for
defining whether a route combination pair can be defined as a set of substitutional routes or not. The analysis measured the impact of changing cut-off parameters on the total number of identified substitutional routes, number of errors and impact of errors in percent. An analysis was conducted for both solutions with changing values for the two cut-off points. The aim was to find the cut-off points that would produce the most accurate prediction of substitutional routes. The following values were applied to

- calculated weights: 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0 and 4.5;
- minimum number of occurrences: 10, 20, 30, 40, 50, 60, 100 and 150.

These changing parameters were applied to the dataset. One scenario required both cut-off parameters to be true (AND) whereas the other scenario only required one cut-off parameter to be true (OR) in order to classify the route combination pair as substitutional routes. Four different main scenarios exist:

**Solution 1 - OR:** This includes the calculation of the substitutional route ratios using Equation 6.4.1. Only one of the two cut-off parameters has to be satisfied.

**Solution 1 - AND:** This includes the calculation of the substitutional route ratios using Equation 6.4.1. Both of the two cut-off parameters have to be satisfied.

**Solution 2 - OR:** This includes the calculation of the substitutional route ratios using Equation 6.4.3. Only one of two the cut-off parameters has to be satisfied.

**Solution 2 - AND:** This includes the calculation of the substitutional route ratios using Equation 6.4.3. Both of the two cut-off parameters have to be satisfied.

In order to prove that the prediction of the algorithm was correct or incorrect a simple random sample was extracted from the entire population. The population size of all route combinations was 6,645. The sample size was calculated using the formula introduced by (Dillman, 2000) (also see 4.5.4).

The sample size for a sample with a 95% confidence interval and a sampling error of 5% is therefore

\[
N_s = \frac{(6,645)(0.5)(1 - 0.5)}{(6645 - 1)(0.05/1.96)^2 + (0.5)(1 - 0.5)} = 363 \text{ cases}
\]  

(6.4.6)

A thorough manual comparison with information provided by the urban bus operator showed whether the route combination pair is in fact a substitutional pair or not. These substitutional routes were then compared with the prediction made by the algorithm. The following three parameters were calculated and analysed:

- Number of Subroutes: This shows the number of identified substitutional routes for the entire population;
6.4. SUBSTITUTATIONAL ROUTES

- Number of Errors: This indicates how many errors were made by the prediction algorithm on the sample;
- Relative percentage of Errors: This evaluation parameter shows the impact of the wrongly identified substitutonal routes on the overall prediction procedure. It is based on the sample and calculated using Equation 6.4.7.

The No of Subroutes shows how many substitutational route pairs were identified under changing parameters. The Number of Errors identifies how many errors the algorithm made when predicting substitutational routes. It compared the manual defined result of the 363 instances of the simple random sample with the predicted value of the algorithm. However, a wrong prediction has different impact as a less frequent route combination pair which was falsely defined as substitutational route has less impact than a more frequent route combination pair that was wrongly identified. False prediction can either be false negative or false positive. Failure to define a frequently used route combination as substitutational route would have considerably more impact on the overall OD results than missing a route combination that has only been recorded a few times and vice versa. The analysis therefore calculated the impact of all errors with regard to the number of occurrences of passenger route combinations. This can be defined as

\[ \text{Error}(\nu, \omega) = \frac{\sum \text{Error}(R_{ij})}{T_{\text{Total}}} \times 100 \]  

(6.4.7)

where \(T_{\text{Error}}(R_{ij})\) is the number of occurrences where the route combination pair was either wrongly identified or failed to be identified. \(T_{\text{Total}}\) is the total number of all occurrences of the sample in this case 30,044. \(\omega\) is the chosen cut-off point for the calculated weight and \(\nu\) is the chosen cut-off point for the number of occurrences. Assuming that the chosen cut-off weight was \(\omega = 1.5\), the number of occurrences had to be greater than \(\nu = 50\) and \(\sum \text{Error}(R_{ij}) = 603\) then the following equation can be applied:

\[ \text{Error}(50, 1.5) = \frac{603}{30,044} \times 100 = 2.01\% \]  

(6.4.8)

The result is the relative percentage of identified errors from the sample. This is seen as an indicator of the impact of the wrongly identified route combination pairs. The result obtained in Equation 6.4.8 indicates that 2.01% of all boarding records were wrongly identified.

The results tables of this analysis are shown in Appendix C. Solution 2 with the OR option is the most robust model as changing parameters (cut-off boardings and percentages) do not have a significant impact on the error ratio. In theory the best model with regard to the impact of errors is Solution 2 with the AND option as the calculated error value is only 1.75% for the cut-off points 50 for the number of occurrence and 1 for the calculated weight. This might be the ideal solution for this sample. However, considering the sampling method with a sample
error of 5% and a confidence interval of 95% it was decided to choose a more robust model
where changing parameters within a certain range do not have a large impact on the OD pair
extraction.

From this point forward all examples and demonstrations will use the following scenario:
Solution 2 where either of the two cut-off points has to be true (OR). The cut-off points include
a minimum of 50 passenger boardings and a calculated cut-off weight of each route combination
pair of $S_2(R_{ij}) = 1.5$ or larger (using Equation 6.4.3).

The analysis can not provide an optimal value for both of the cut-off parameters for all
datasets. It solely serves for the purpose of this project. However, the methodology described
above can be applied to identify the cut-off points for other datasets. The correct cut-off points
depend on the data and the transport network. It was therefore necessary to explore some of the
routes manually to define the values of the cut-off parameters. Optimistic definitions of values
may include substitutional routes which are in fact not substitutional but also includes all 'real'
substitutional routes. A pessimistic approach may result in the exclusion of substitutional route
because the cut-off points were too high.

This analysis was also carried out for the months April and September. The list of substi­tutional routes for each of these months was almost identical. The impact error analysis also
proved that different months do not significantly change the list of substitutional routes. How­
ever, the lists of substitutional routes could change more significantly over periods of time.

One option which was not pursued due to the relatively low estimation error would have
been to include information on a passenger level such as route choice patterns of individual
passenger records. More detailed spatial information such as geographic coordinates could also
improve the algorithm. But, considering the low impact error of the algorithm it was decided
not to pursue this problem of substitutional routes any further.

As this method analyses actual passenger behaviour rather than making inferences directly
from network and timetable data, these inferences could include some habitual diversions. This
would explain why less evening return journeys match the optimal criteria due to, e.g. late night
shopping or entertainment. However, the cut-off point should separate the significant from the
insignificant habitual diversions.

To decide in a formal statistical manner whether the difference is significant it is possible to
calculate the probability distribution of the value $S$ given that the routes are not substitutional.
This could be done by using a Monte Carlo simulation of differences under a model that would
need to be defined. The second method would require to analyse the histogram of values of $S$
of a sample of routes that are not substitutional. The cut-off point for determining that routes
are substitutional would then be some high percentile of this histogram such as 95% or 99%tile.
However, considering low impact errors and the relative low overall error rate after comparing the results with the simple random sample this was not further pursued.

6.4.5 Summary

The proposed algorithm to define substitutional routes is an efficient and fast way of creating a list of substitutional routes as it is not network specific. It can therefore be applied to any transport network for which the required data are available. The flexibility in deciding the values of the two cut-off parameters turns the algorithm into a tool that can successfully identify the substitutional routes of a transport mode. Should the data be collected from different modes the algorithm needs to be changed so that one list of substitutional routes is created incorporating all modes. However the algorithm is most suitable for bus networks as these are generally more complex in comparison with light rail or heavy rail networks.

The real advantage of the algorithm is however the identification of substitutional routes based on the behaviour of the network’s passengers. Sometimes the passenger may accept a short walk to access a bus service which also serves his/her final destination. The transport planner may not be aware of these behavioural aspects which could be important to public transport passengers. Failing to identify a substitutional route could lead to unidentified OD pairs which in turn can bias the results obtained by the OD extraction algorithm. It is therefore important that all substitutional routes be identified. Consequently the strongest advantage of using this algorithm is that it incorporates the travel behaviour of passengers which is reflected by the stored boarding records of the EFC system.

6.5 Development of the Rule Base

6.5.1 Introduction

The rule base should consist of all physically and logically possible attribute combinations that could have been recorded for a passenger during the validity period of the ticket. The travel record for each customer reflects all journeys that were carried out by a particular passenger. These records have to be analysed by the OD extraction algorithm which then infers the OD pairs for all logical boarding combinations that occurred.

A technique called Decision Tables is used to identify each of the possible rules. A list of all structural patterns is identified which forms the foundation of the OD extraction algorithm.

6.5.2 Decision Tables

Decision tables are generally considered to be a set of variables or attributes which lead to an action, policy or alternative (Fernandez del Pozo et al., 2003). Decision makers mostly use
decision tables to search for the best recommendation for a certain case or attribute (Fernandez del Pozo et al., 2003; Hewett and Leuchner, 2003). They are further used in knowledge-based decision support systems (DSS) as decision tables have the ability to represent complex logical relationships in an explicable and comprehensible manner (Cragun and Steudel, 1987; Slowinski, 1992; Kohavi and Sommerfield, 1998; Hewett and Leuchner, 2003). Decision tables also facilitate to further process the data or information efficiently due to the fact that they can be used with other representation models such as tabular knowledge bases, some relational databases, rule sets or decision trees (Hewett and Leuchner, 2003). Decision tables are not only helpful when procedural knowledge needs to be represented but also for the implementation of knowledge bases (Santos-Gomez and Darnell, 1992).

For the purpose of this project, decision tables are mainly used as a programming methodology to support the fast execution of the algorithm. They are also used as a technique to ensure completeness, consistency and correctness of all alternatives. Furthermore they eliminate redundancy of alternatives.

The decision tables are therefore useful in validating knowledge (Hewett and Leuchner, 2003). The final decision tables of this project will then be translated into a set of rules which build the knowledge base for the OD extraction process.

### 6.5.3 Determining the Rule Attributes

The use of the recorded attributes introduced in the following paragraphs are fundamental to the successful inference of OD pairs as they form the rules the main algorithm will be based upon. When comparing single journeys, two boarding records are compared with each other (A and B). Transfer journeys consist of four individual boarding records (boarding records A and B for journey 1 and boarding records C and D for journey 2 - the return journey). When comparing transfer journeys the individual boarding records are paired into records A/D and B/C.

The following decision attributes were identified for building the rules and ultimately the rule base. All attributes are justified using existing literature or prior analysis as part of this research:

**Date** \( (d) \) ⇒ The value for this attribute can either be 1 (same day) or 2 (different day). This is similar to Barry et al. (2002) and Zhao (2004) although the cut off point for each day was 4a.m. in their study whereas it is 12p.m. for this study as regular Dublin Bus services stop to operate at 12p.m.. Implementations of this algorithm for other networks need to consider this aspect and make appropriate changes with regard to the definition of a 'day'. At a later stage this attribute will form a new rule (2nd iteration of the OD Algorithm) where it is in fact partially ignored arguing that identified repeated patterns of boardings and alightings can also be used to infer OD pairs (see Section 6.7). Trepanier et al. (2007a) also analysed and compared \( d + 1 \) and \( d - 1 \) with the premise that the passenger’s boarding location in the morning is the location of
alighting of the last journey of the previous day. The model presented as part of this research addresses this possibility during the second iteration of the OD algorithm in a slightly different approach (see Section 6.7). However, it was already shown that not all journeys start in the morning where they have ended in the evening.

\textbf{Route} \( (w_s) \Rightarrow \) Single journeys: Three values can be assigned to this decision attribute: 1 (both boardings took place on the same route), 2 (both boardings took place on a substitutional route) and 3 (the boardings took place on two different non-substitutional routes). For the projects described in Barry et al. (2002) and Zhao (2004) the route was not recorded by the EFC system. It was the algorithm that inferred the route based on the trip chain data and the station the boarding record was created. The algorithm was therefore not able to clearly identify the route when substitutional routes for the route segment existed. Trepanier et al. (2007a) uses the route to infer the destination of a journey. However, trip chaining is used which means that substitutional routes do not need to be considered. The algorithm proposed in this research can on the other hand correctly identify each route as this is part of the passenger boarding record. This is of particular importance when for example passenger load levels need to be extracted and analysed.

\textbf{Route A/D or Route B/C} \( (w_{T1}, w_{T2}) \Rightarrow \) Transfer journeys: Three values can be assigned to this decision attribute: 1 (both boardings took place on the same route), 2 (both boardings took place on a substitutional route) and 3 (the boardings took place on two different non-substitutional routes). As all similar projects used the assumption of trip chaining, transfer routes did not need to be identified separately. It was shown that trip chaining can not be applied to Dublin’s EFC data which forced the separate consideration of transfer journeys. Trepanier et al. (2007a) use similar assumptions as are used in the transfer journey section of this thesis. These assumptions include that the alighting stop from the first route and the boarding stop of the second route are in close vicinity. A tolerance value of 2km was set by Trepanier et al. (2007a) to determine transfer journeys. Transfer journeys were therefore identified as part of the OD inference process and not separately as it was done for this project.

\textbf{Direction} \( (c_s) \Rightarrow \) Single journeys: Two values can be assigned to this decision attribute: 1 (the two boarding records have different directions) and 2 (both boardings have the same direction). Barry et al. (2002) and Zhao (2004) do not have a recorded value for the directional attribute. This again can only be inferred when analysing the trip chain. The Dublin Bus EFC system recorded a directional identifier for each passenger boarding record. This is one of the most important attributes as the proposed model does not use the trip chaining approach. OD information can only be determined when a corresponding 'return journey' has been found. This 'return journey' has to travel into the opposite direction. Trepanier et al. (2007a) use the direction attribute to calculate the possible remaining alighting stops but do not check it against the
direction of the following boarding.

**Direction A/D or Direction B/C** \( (c_{T1}, c_{T2}) \Rightarrow \) Transfer journeys: Two values can be assigned to this decision attribute: 1 (the two boarding records have different directions) and 2 (both boardings have the same direction). The direction attribute is checked for all four boarding records to see whether the passenger actually travelled into the opposite direction during the return journey.

**Journey** \( (v) \Rightarrow \) This attribute focuses on the type of journey. Two values can be assigned to this decision attribute: 1 (single journey) and 2 (transfer journey with one transfer). Barry et al. (2002), Zhao (2004), and Trepanier et al. (2007a) do not have to differentiate between single and transfer journeys (at least not directly). Due to the main trip chaining assumption there is no need to identify the journey type before commencing with the analysis. This attribute is of importance as two different procedures exist: one for single journeys and one for transfer journeys. However, the iterative classification algorithm (see Section 4.5.2) set this attribute previously.

**Repeated** \( (/) \Rightarrow \) This attribute focuses on whether the two journeys were repeated with identical parameters such as route (or substitutional routes), stage and direction which is important with regard to the likelihood that the extracted destination is in fact correct. Three values can be assigned to this attribute: 1 (the OD pair was repeated on the same day), 2 (the OD pair was repeated within the validity period of the ticket) and 3 (the OD pair was not repeated). Barry et al. (2002), Trepanier et al. (2007a) and Zhao (2004) do not analyse this. The reasoning behind this attribute is that inferred OD pairs that occurred frequently throughout the validity period of the passenger’s ticket are more likely to be correctly inferred as non-reoccurring OD pairs. This is a derivable attribute which is generated after the potential OD pair has been identified.

**Order** \( (o) \Rightarrow \) This attribute focuses on the order of the boarding records as they are compared. 1 (the boarding records are in consecutive order which means that they were recorded immediately after each other) and 2 (the boarding records were not in consecutive order). Due to the trip chaining assumption Barry et al. (2002), Trepanier et al. (2007a) and Zhao (2004) do not need to include this parameter into their algorithm. This attribute shows if passengers travelled between their 'main' return journeys (e.g. during lunch).

After defining all available data attributes it is possible to compile the rule base which takes place initially by using decision tables. For the purpose of the presentation of decision tables and due to the structural difference between single (2 boardings) and transfer journeys (4 boardings) it was decided to split the decision table into two separate tables. One for single journeys and one for transfer journeys.

**Single Journeys**
Table 6.8 summarizes a list of all possible conditions and their values. Journey type $v$ has to be single.

Table 6.8: Conditions and Possible Values for Single Journeys

<table>
<thead>
<tr>
<th>Condition</th>
<th>Possible values of the conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day $d$</td>
<td>(Same Day = 1, Following Day = 2)</td>
</tr>
<tr>
<td>Route $w_s$</td>
<td>(Same Route = 1, Substitute Route = 2, Different Route = 3)</td>
</tr>
<tr>
<td>Direction $c_s$</td>
<td>(Different direction = 1, Same Direction = 2)</td>
</tr>
<tr>
<td>Repeated $f$</td>
<td>(Same Day = 1, Period of Ticket Duration = 2, Not Repeated = 3)</td>
</tr>
<tr>
<td>Order $o$</td>
<td>(Consecutive = 1, Non-consecutive = 2)</td>
</tr>
</tbody>
</table>

The number of theoretical possible classes is the product of all alternatives of each conditional attribute (i.e. Day, Route, Direction, Repeated and Order). There are $2 \times 3 \times 2 \times 3 \times 2 = 72$ possible combinations. All illogical combinations which do not lead to an OD pair have to be omitted as the only interest is to obtain valid OD pairs.

The following elimination rules apply for single journeys leaving a total of 12 remaining combinations:

1. IF Direction = \{Same Direction\} THEN Journey $\neq$ \{Return Journey\} $\Rightarrow$ 36 eliminations;
2. IF Day = \{Following Day\} THEN Journey $\neq$ \{Return Journey\} $\Rightarrow$ 18 eliminations;
3. IF Route = \{Different Route\} THEN Journey $\neq$ \{Return Journey\} $\Rightarrow$ 6 eliminations.

To (1): If the two boarding records have the same value for the direction attribute it cannot be a return journey and therefore no valid OD pair can be obtained.

To (2): If the two boarding records took place on a different day then it cannot result in an OD pair.

To (3): If the two boarding records have a different value of the route attribute and were not identified as substitutional routes then no OD pair can be obtained.

Table 6.9 shows the 12 remaining combinations after eliminating the mutually exclusive and illogical combinations, the elimination of redundant conditions and the definition of the different classes for each combination.
Table 6.9: Final Rule Base for Single Journeys

<table>
<thead>
<tr>
<th>Rule</th>
<th>Day $d_S$</th>
<th>Route $w_S$</th>
<th>Direction $c_S$</th>
<th>Repeated $f_S$</th>
<th>Order $o_S$</th>
<th>Class $y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_{S1}$</td>
<td>Same Day</td>
<td>Same route</td>
<td>Different</td>
<td>Same day</td>
<td>Consecutive</td>
<td>$S1$</td>
</tr>
<tr>
<td>$r_{S2}$</td>
<td>Same Day</td>
<td>Same route</td>
<td>Different</td>
<td>Same day</td>
<td>Non-consecutive</td>
<td>$S2$</td>
</tr>
<tr>
<td>$r_{S3}$</td>
<td>Same Day</td>
<td>Same route</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Consecutive</td>
<td>$S3$</td>
</tr>
<tr>
<td>$r_{S4}$</td>
<td>Same Day</td>
<td>Same route</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Non-consecutive</td>
<td>$S4$</td>
</tr>
<tr>
<td>$r_{S5}$</td>
<td>Same Day</td>
<td>Same route</td>
<td>Different</td>
<td>Not repeated</td>
<td>Consecutive</td>
<td>$S5$</td>
</tr>
<tr>
<td>$r_{S6}$</td>
<td>Same Day</td>
<td>Sub route</td>
<td>Different</td>
<td>Same day</td>
<td>Non-consecutive</td>
<td>$S6$</td>
</tr>
<tr>
<td>$r_{S7}$</td>
<td>Same Day</td>
<td>Sub route</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Consecutive</td>
<td>$S7$</td>
</tr>
<tr>
<td>$r_{S8}$</td>
<td>Same Day</td>
<td>Sub route</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Non-consecutive</td>
<td>$S8$</td>
</tr>
<tr>
<td>$r_{S9}$</td>
<td>Same Day</td>
<td>Sub route</td>
<td>Different</td>
<td>Not repeated</td>
<td>Consecutive</td>
<td>$S9$</td>
</tr>
<tr>
<td>$r_{S10}$</td>
<td>Same Day</td>
<td>Sub route</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Non-consecutive</td>
<td>$S10$</td>
</tr>
<tr>
<td>$r_{S11}$</td>
<td>Same Day</td>
<td>Sub route</td>
<td>Different</td>
<td>Not repeated</td>
<td>Consecutive</td>
<td>$S11$</td>
</tr>
<tr>
<td>$r_{S12}$</td>
<td>Same Day</td>
<td>Sub route</td>
<td>Different</td>
<td>Not repeated</td>
<td>Non-consecutive</td>
<td>$S12$</td>
</tr>
</tbody>
</table>

Transfer Journeys

Table 6.10 shows a list of all conditions and their values. The number of theoretical possible classes is the product of all alternatives of each conditional attribute. There are $2 \times 3 \times 3 \times 2 \times 3 \times 2 \times 3 \times 2 = 432$ possible combinations. All illogical combinations which do not lead to an OD pair can be omitted.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Possible values of the conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day - $d$</td>
<td>(Same Day = 1, Following Day = 2)</td>
</tr>
<tr>
<td>Route of A/D - $w_{T1}$</td>
<td>(Same Route = 1, Substitute Route = 2, Different Route = 3)</td>
</tr>
<tr>
<td>Route of B/C - $w_{T2}$</td>
<td>(Same Route = 1, Substitute Route = 2, Different Route = 3)</td>
</tr>
<tr>
<td>Direction of A/D - $c_{T1}$</td>
<td>(Different Direction = 1, Same Direction = 2)</td>
</tr>
<tr>
<td>Direction of B/C - $c_{T2}$</td>
<td>(Different Direction = 1, Same Direction = 2)</td>
</tr>
<tr>
<td>Repeated - $f$</td>
<td>(Same Day = 1, Period of Ticket Duration = 2, Not Repeated = 3)</td>
</tr>
<tr>
<td>Order - $o$</td>
<td>(Consecutive = 1, Non-consecutive = 2)</td>
</tr>
</tbody>
</table>

The following elimination rules apply for transfer journeys leaving a total of 24 possible remaining combinations:

1. IF Dir (A/D) = \{Same Dir.\} THEN Journey $\neq$ \{Return Journey\} $\Rightarrow$ 216 eliminations;
2. IF Dir (B/C) = \{Same Dir.\} THEN Journey $\neq$ \{Return Journey\} $\Rightarrow$ 108 eliminations;
3. IF Day = \{Different Day\} THEN Journey $\neq$ \{Return Journey\} $\Rightarrow$ 54 eliminations;
4. IF Route (A/D) = \{Diff. Route\} THEN Journey $\neq$ \{Return Journey\} $\Rightarrow$ 18 eliminations;
5. IF Route (B/C) = \{Diff. Route\} THEN Journey $\neq$ \{Return Journey\} $\Rightarrow$ 12 eliminations.

To (1): If the values of the direction attribute of transfer boarding records A and D are the same no OD pair can be obtained for this transfer journey.
To (2): If the values of the direction attribute of transfer boarding records B and C are the same no OD pair can be obtained for this transfer journey.

To (3): If the two boarding records took place on a different day then no OD pair can be obtained for this transfer journey.

To (4): If the values of the route attribute of journey records A and D are different then no OD pair can be obtained.

To (5): If the values of the route attribute of journey records B and C are different then no OD pair can be obtained.

Table 6.11 summarizes the 24 remaining combinations for transfer journeys after eliminating the mutual exclusive and illogical combinations, the elimination of redundant conditions and the definition of the different classes for each combination.

### 6.5.4 Identified Rules

After eliminating all redundancies, contradictions and mutual exclusive alternatives one class was assigned to each of the attribute combinations. The type of class depends on the composition of the values of each attribute. Each class leads to an action and consequently defines the strength of the OD pair extracted by the OD extraction algorithm. Single journey classes consist of classes $y_{S1}$ to $y_{S12}$ and transfer journey classes consist of classes $y_{T1}$ to $y_{T24}$.

The rule base $R$ consists of rules $r_i$. Each classification rule can be expressed as

$$r_i : (\text{Condition}_i) \longrightarrow y_i.$$  

The left hand side of the rule is known as precondition or rule antecedent which contains a conjunction of attribute tests such as

$$\text{Condition}_i = ((A_1 \text{ op } v_1) \land (A_2 \text{ op } v_2) \land ... \land (A_k \text{ op } v_k)),$$

where $(A_j \text{ op } v_j)$ is an attribute-value pair and $\text{op}$ defines the logical operator. The rule consequent represents the right hand side of the rule which contains the inferred class $y_i$. For example,

$$r_{S1}: (d_s = \text{'same'} \land w_s = \text{'same'} \land c_s = \text{'diff.'} \land f_s = \text{'same'} \land o_s = \text{'consecutive'}) \longrightarrow y_{S1}.$$

Tables 6.9 and 6.11 show all rules $r$ and their corresponding classes $y$. Throughout the OD algorithm all 32 identified rules are checked for each potential OD pair records.
The information to which class a passenger boarding pair belongs will be carried forward and also added to the existing database. It will be valuable during the analysis of the results. The classes introduced above only focus on an individual OD pair. However, it would also be useful to analyse all OD pairs per individual passenger. This second stage of the OD extraction algorithm will be demonstrated in Section 6.7.

### 6.6 The Initial OD Algorithm

This section outlines how the algorithm functions. Appendix C shows all procedural diagrams of the algorithm in greater detail.

We define

\[ B = \{b^j\}, \]  

(6.6.1)

to be the set of all boarding records with \( b^j \) defining boarding records for passenger \( j \). We further define \( b^j \) to be:

\[ b^j = \{b^j_{dwc} \}, \]  

(6.6.2)

where \( b^j_{dwc} \) is the \( i \)th boarding record with day \( d \), route \( w \), direction \( c \) and journey type \( v \).
All boarding records \( B \) are ordered in ascending order of passenger \( j \), date \( d \) and time \( t \). There is no relationship between individual passengers when inferring OD pairs and therefore the algorithm only considers the boarding records of each individual passenger \( b^j \). \( S \) and \( T \) stand for single journey and transfer journey respectively. The variables \( f \), \( o \), \( r \) and \( y \) are defined in Table 6.11. The following pseudo code of the algorithm outlines its structure.

For each file \( b \) in \( B \)
For each passenger \( j \) in file \( b \)
  For each boarding record \( i \) in \( b^j \)
    IF \( b^{ji}(v) = \) Single Journey Then
      \( k = i + 1 \)
      For each boarding \( b^{jk}_S \) where \( k > i \) and \( b^{jk}(v) = \) Single Journey
        Compare Date \( b^{ji}_S(d) \) and \( b^{jk}_S(d) \)
        Compare Routes \( b^{ji}(w) \) and \( b^{jk}(w) \)
        Compare Direction \( b^{ji}(c) \) and \( b^{jk}(c) \)
        Compare Repeat \( b^{ji}(f) \) and \( b^{jk}(f) \) in \( b^j \)
        Compare Order \( b^{ji}(o) \) and \( b^{jk}(o) \)
        Check rules \( r \) in Rule Base \( R_S \)
        IF rule found THEN
          Output OD pair and class \( y_S \)
          break
        END
      END
    Elseif \( b^{ji}(v) = \) Transfer Journey Then
      \( k = i + 1 \)
      IF \( b^{jk}(v) = \) Transfer Journey Then
        \( l = k + 1 \)
        For each boarding \( b^{jl} \)
          IF \( b^{jm}(v) = \) Transfer Journey Then
            \( m = l + 1 \)
            Compare Date \( b^{ji}_T(d) \) and \( b^{jm}_T(d) \)
            Compare Routes \( b^{ji}_T(w) \) and \( b^{jm}_T(w) \)
            Compare Direction \( b^{ji}_T(c) \) and \( b^{jm}_T(c) \)
            Compare Date \( b^{ji}_T(d) \) and \( b^{jm}_T(d) \)
            Compare Routes \( b^{ji}_T(w) \) and \( b^{jm}_T(w) \)
            Compare Direction \( b^{ji}_T(c) \) and \( b^{jm}_T(c) \)
            Compare Repeat \( b^{ji}_T(f) \), \( b^{jm}_T(f) \), \( b^{jk}_T(f) \) and \( b^{jl}_T(f) \) in \( b^j \)
            Compare Order \( b^{ji}_T(o) \), \( b^{jm}_T(o) \), \( b^{jk}_T(o) \) and \( b^{jl}_T(o) \)
          END
      END
    END
6. DESIGN AND DEVELOPMENT OF THE ORIGIN/DESTINATION ALGORITHM

Check rules $r$ in Rule Base $R_T$
IF rule found THEN
Output OD pair and class $y_r$
break
END
END
END
END
END
END

Of particular importance is the fact that only one rule at most can match the result of the comparison as the rules are mutually exclusive. Therefore only one rule can be triggered to identify and classify the OD pair. It is therefore not a fuzzy algorithm.

The initial algorithm does not incorporate already identified OD pairs and therefore focuses on two (single journeys) or four (transfer journeys) passenger boarding records only. This is a limitation in itself as surely the overall recorded passenger behaviour over the validity period of the ticket contains further information that could lead to a higher and more accurate identification of OD pairs. A second stage algorithm was therefore developed to take advantage of recurring patterns of passengers’ boarding records. This addition to the first iteration of the OD extraction algorithm is explained in the following section.

6.7 Extended OD Extraction Algorithm

6.7.1 Introduction

Knowing the information detailed in the previous section and having the boarding patterns of all passengers facilitates an analysis with regard to the performance of the OD extraction algorithm. As the aim was to extract the initial OD set with the least amount of assumptions the algorithm may extract more OD pairs when introducing new assumptions. This section will analyse entire individual passengers’ boardings and OD record data (from the first iteration) in order to determine areas of improvements. It could be argued that there is travel information and knowledge in the entire boarding records of each passenger which will be further explored in this section.
This extension of the first iteration of the OD extraction algorithm uses a merged dataset consisting of boarding records that led to OD pairs and boarding records that did not lead to OD pairs. The aim is to determine why the algorithm could not infer more OD pairs and ultimately what assumptions could be employed to achieve a greater OD extraction ratio.

Throughout October 1999 323,652 different passenger tickets were identified. These tickets recorded almost 2.3 million unlinked trips. The journey pattern of 109,846 tickets identified an odd number of boarding records. Considering that the algorithm requires at least two boarding records (even number) to form an OD pair this immediately results in 109,846 trips that cannot be assigned a final destination unless new rules can be formed. The following sections discuss potential rules to improve the OD extraction ratio.

### 6.7.2 One Boarding Record per Ticket

The boarding records showed that 58,572 individual tickets recorded only one journey which makes it impossible to assign a final destination to the record. 21,175 of these tickets are 1-Journey tickets. No reasonable assumption would allow one to infer the passenger’s destination at trip level.

There is anecdotal evidence in London that working OAPs who got a travel pass, which is only valid after the morning peak, use a single ticket to get to work during the morning peak period. For the return journey the travel pass is used.

Therefore it is only possible to improve the algorithm for tickets that recorded more than one boarding record throughout the validity period of the ticket. Networks that allow the re-charging of cards (such as smart cards) will often eliminate tickets with such a low utilisation rate and therefore improve the potential OD extraction ratio. Rechargeable smartcards such as the PAYG Oyster cards have the effect that these once off boardings on a ticket are greatly reduced.

### 6.7.3 Two Boarding Records per Ticket

162,087 individual tickets recorded only two unlinked trips which makes it difficult to assign a single OD pair as there is no re-occurring pattern. 57,081 OD pairs were recorded on tickets that consisted of two unlinked trips. Transfer journey (linked trip) records could not be used to define an OD pair as 4 boardings are necessary. The question arises as to which further assumption would improve the OD extraction ratio. When analysing the two boarding records detailed in Table 6.12 and Table 6.13 it becomes apparent that the algorithm would have assigned an OD
pair if, and only if, both boardings took place on the same day. The two single journeys were carried out on the same route and in different directions. The boarding location is also different. Therefore if the date had been the same the algorithm would have inferred an OD pair. A possible assumption could therefore be that in the case of two identical passenger boardings the date can be ignored for the second iteration of the OD extraction algorithm.

<table>
<thead>
<tr>
<th>ID</th>
<th>Ticket Type</th>
<th>Ticket ID</th>
<th>Journey Type</th>
<th>Date</th>
<th>Stage</th>
<th>Time</th>
<th>Route</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>5255468</td>
<td>310</td>
<td>84</td>
<td>1</td>
<td>13/10/99</td>
<td>75</td>
<td>9:24</td>
<td>90</td>
<td>1</td>
</tr>
<tr>
<td>5729410</td>
<td>310</td>
<td>84</td>
<td>1</td>
<td>18/10/99</td>
<td>25</td>
<td>18:55</td>
<td>90</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.12: Ticket with only Two Boarding Records - Example 1

<table>
<thead>
<tr>
<th>ID</th>
<th>Ticket Type</th>
<th>Ticket ID</th>
<th>Journey Type</th>
<th>Date</th>
<th>Stage</th>
<th>Time</th>
<th>Route</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>4342062</td>
<td>310</td>
<td>8093</td>
<td>1</td>
<td>01/10/99</td>
<td>92</td>
<td>19:09</td>
<td>102</td>
<td>0</td>
</tr>
<tr>
<td>4454787</td>
<td>310</td>
<td>8093</td>
<td>1</td>
<td>04/10/99</td>
<td>00</td>
<td>9:00</td>
<td>102</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.13: Ticket with only Two Boarding Records - Example 2

When implementing this algorithm two more classes were added to the single categories. Class $y_{S13}$ represents all cases where the new assumption was implemented and the route ID was the same. Class $y_{S14}$ represents all cases where the route number was not the same for the two boardings but the routes were substitutional.

This second iteration of the main algorithm with the new assumption resulted in the following aggregated numbers of OD pairs:

- Class $y_{S13}$ resulted in an additional 4,551 cases. This is equivalent to approximately 0.61% of all single journeys that were assigned a destination following the above stated assumption;
- Class $y_{S14}$ resulted in an additional 4,150 cases. This is equivalent to approximately 0.57% of all single journeys that were assigned a destination.

The new assumption could successfully be applied for 8,701 of these tickets. Furthermore, 95% of identified OD pairs (classes $y_{S13}$ and $y_{S14}$) were journeys recorded with a two journey ticket. The new assumption states that the date of the boardings can be neglected when only two boardings were recorded on any particular ticket. This could theoretically be extended to tickets with more than two boardings. It is however believed that there is not enough evidence when analysing various passenger travel records to enforce this assumption. It will therefore not be applied to tickets that recorded more than two boardings throughout their life span.
6.7. EXTENDED OD EXTRACTION ALGORITHM

Table 6.14: Weekly Ticket with several Boarding Records

<table>
<thead>
<tr>
<th>No</th>
<th>Case</th>
<th>ID</th>
<th>Ticket Type</th>
<th>Ticket ID</th>
<th>Journey Type</th>
<th>Date</th>
<th>Time</th>
<th>Bus Stage</th>
<th>Route ID</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4473196</td>
<td>410</td>
<td>172937</td>
<td>1</td>
<td>04/10/99</td>
<td>63</td>
<td>6:53</td>
<td>15B</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4463509</td>
<td>410</td>
<td>172937</td>
<td>1</td>
<td>04/10/99</td>
<td>12</td>
<td>19:59</td>
<td>29A</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4555819</td>
<td>410</td>
<td>172937</td>
<td>1</td>
<td>05/10/99</td>
<td>63</td>
<td>12:22</td>
<td>15B</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4603552</td>
<td>410</td>
<td>172937</td>
<td>1</td>
<td>05/10/99</td>
<td>25</td>
<td>22:55</td>
<td>15B</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4684476</td>
<td>410</td>
<td>172937</td>
<td>1</td>
<td>06/10/99</td>
<td>26</td>
<td>17:21</td>
<td>15</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4830990</td>
<td>410</td>
<td>172937</td>
<td>1</td>
<td>07/10/99</td>
<td>64</td>
<td>12:51</td>
<td>15</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4805009</td>
<td>410</td>
<td>172937</td>
<td>1</td>
<td>07/10/99</td>
<td>25</td>
<td>23:22</td>
<td>15</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>4913715</td>
<td>410</td>
<td>172937</td>
<td>1</td>
<td>08/10/99</td>
<td>63</td>
<td>16:58</td>
<td>15B</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4979149</td>
<td>410</td>
<td>172937</td>
<td>1</td>
<td>09/10/99</td>
<td>63</td>
<td>8:02</td>
<td>15B</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>4974921</td>
<td>410</td>
<td>172937</td>
<td>1</td>
<td>09/10/99</td>
<td>25</td>
<td>20:44</td>
<td>15B</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

6.7.4 Three or more Boarding Records per Ticket

It becomes more interesting when considering tickets that consist of more than 2 validations throughout their life span. An emerging pattern can often be identified for those tickets that have several identified OD pairs. The following discussion will outline what further assumptions could improve the OD extraction algorithm.

Analysing the records of a weekly adult ticket in Table 6.14 highlights the problems that exist when not using further assumptions. In this case the destination could only be identified for the record pairs 3 and 4, 6 and 7, and 9 and 10. A destination for each of the remaining records was not possible to determine as no boarding record of the return journey exists.

Record 1 of Table 6.14 for example is exactly the same as record 3 apart from the date the boarding was recorded. The same applies to records 5/7 and 8/9. It could be assumed that in these cases the passenger carried out the exact same journey without, for whatever reason, boarding a bus for the return journey. Following this assumption it would be possible to infer more passenger destinations. It is however unclear how to proceed when the destination of these similar boarding records were inferred differently. In the case of the passenger data of Table 6.14 it is not a problem as the records 3 and 9 were inferred with the same destination. Additionally, records 4 and 10 have identical boarding locations. If, however, the boarding locations of the second boarding of an OD pair vary then the assumption weakens. The algorithm may allow a certain amount of leeway with regard to the boarding location. For example it could be argued that differences such as two to three bus stages in either direction are acceptable. In summary the following assumption may improve the OD pair extraction ratio: All records that were not identified as part of an OD pair and therefore have no inferred destination and which
are identical (apart from the date of recording) with boarding records of which the location of alighting is known may be assigned with the same destination as of the already inferred OD pair. When multiple possibilities with varying destinations exist the boarding time may be considered to decide the most similar pair.

Applying this assumption to the records presented in Table 6.14 would have the following impact:

- Boarding record 1 is identical to records 3 and 9. Both of these records have the same location of alighting which is bus stage 25 on route 15B. This is therefore also the inferred location of record 1;
- Boarding record 2 is not identical to any of the remaining records and can therefore not be assigned a destination;
- Boarding record 5 is identical to record 7. The original OD extraction algorithm assigned the location of alighting as bus stage 64 on route 15. This is therefore also the inferred location of record 5;
- Boarding record 8 is identical to records 3 and 9. Both of these records have the same location of alighting which is bus stage 25 on route 15B. This is therefore also the inferred location of record 8.

Using the new assumption it would be possible to assign three new locations of alightings to the presented records 1, 5 and 8. (see Table 6.14). When implementing this new assumption two more classes were added to the single categories. Class $y_{S15}$ represents all cases where the new assumption was implemented and the route ID was the same. Class $y_{S16}$ represents all cases where the route number was not the same for the two boardings but the routes were substitutional.

The extended algorithm assigned a destination to 138,519 records. It is important to note that in these two cases the destination was inferred without having recorded a return journey and therefore it cannot be considered as an OD pair but simply as a record of which the destination is known. This new assumption inferred the destination of 9.2% of all single journeys. 51,219 single journey destinations were assigned to class $y_{S15}$ where the two routes had to be the same. The remainder of 87,300 were assigned to class $y_{S16}$ where the route ID’s of the single journey records were substitutional.

An analysis with the aim to identify a pattern when exactly passengers do not travel back to their origin was carried out on two class groups (classes 1-14 and 15-16). The analysis showed that mostly the days in mid week (Tuesday, Wednesday and Thursday) were the days when passengers tended to make more return journeys which was as expected. The number of return journeys was 5 times lower on Saturday and Sunday than it was during days midweek.
6.8. ALGORITHM INNOVATIONS

The second analysis based on the results obtained by implementing the new assumption focused on the actual time when passengers boarded the bus. This analysis compared the boarding time of the cases 15-16 with the boarding time of all single journeys. The aim was to identify a change in passenger behaviour. However, no emerging patterns could be identified.

6.7.5 Summary of Algorithm Improvements

Two new assumptions have been added to improve the algorithm. The first assumption states that when a ticket only consists of two boarding records recorded then the date can be ignored for the second iteration of the OD extraction algorithm. The second assumption states that all records that were not identified as part of an OD pair and therefore have no inferred destination and which are identical (apart from the date of recording) with boarding records of which the location of alighting is known may be assigned with the same destination as of the already inferred OD pair. The two new assumptions have contributed considerably to the overall ratio of identified OD pairs. Although the two new additions are not as strong as the main assumption evidence suggests that they are correct. Unfortunately it is not possible to verify them in their entirety.

This extended algorithm works in particular well for weekly and monthly tickets where an improvement of OD estimation of up to 37% could be achieved. There is considerable less improvement for limited journey tickets.

In summary it can be concluded that the two new assumptions improved the main algorithm and also showed that assumptions do not necessarily have to be based on the main dataset but can be based on result sets too.

6.8 Algorithm Innovations

This section focuses on the differences of the proposed algorithm in comparison to existing algorithms.

One of the main innovations is that the main assumption is not based entirely on trip chaining but on the actual entire boarding record of the return journey. Trip chaining would result in many errors as shown in Section 6.2.2 for Dublin’s bus network. None of the existing algorithm actually did any analysis on the trip chaining assumption directly. It becomes clear when comparing existing literature that there are fundamental differences in metro and bus networks when inferring OD pairs at trip level. The complexity of bus networks (more stops/stages) and the offered number routes mainly contribute to this difference. Furthermore the proposed algorithm directly incorporates route and direction attribute values which is only done to some extent by Trepanier et al. (2007a). All but the proposed algorithm infer the location entirely based the
following journey disregarding any details of that journey other than the date. This is done by assuming that (1) public transport passengers do not use private transportation (car, motorbike, bicycle, etc.) between trip segments, (2) no passenger walks a long distance in order to board at a different station to the one he/she alighted and (3) the last trip of the day ends where the first trip of the day was recorded (Wilson et al., 2005). The Dublin Bus EFC data indicate and provide evidence that these assumptions cannot be applied in their entirety to all boarding records. More reasoning needs to be applied to make each decision of the OD pair based on the attribute values of the journeys belonging to this OD pair. This is done by identifying and implementing the rule base.

Routes, direction and journey type (single or transfer) do not come into consideration for all metro based approaches. This is a clear advantage of bus EFC systems as these attributes are mostly recorded and can therefore be incorporated in the method of inferring destinations at trip level. As mentioned throughout this chapter all rules of the first iteration of the OD extraction algorithm incorporate route, direction and journey type. Therefore the more attributes are integrated when inferring destinations the more informed can the decision of destination can be made.

The proposed algorithm uses a second iteration to further infer OD pairs by analysing the inferred OD pairs (first iteration) with non-inferred boarding records. This is done by comparing the attribute values of non identified OD pairs with similar attribute combinations of inferred OD pairs. Only Trepanier et al. (2007a) does this to some extent.

All OD pairs are attributed to one of 36 classes (first iteration) and an additional 4 classes (second iteration) which all in turn can be used to analyse the strength of the inferred OD pair. Furthermore, the approach of rule based reasoning is advantageous when new rules need to be implemented as these can easily be added to the rule base without, in general, changing the inference engine.

Substitutional routes were solely considered by the proposed algorithm and were automatically identified and also used when inferring OD pairs. This is not applicable to existing other OD extraction algorithms as the trip chaining assumption is applied. The substitutional identification algorithm is novel and unique in its reasoning.

The proposed algorithm is the only one that makes a difference between single and transfer journeys. Transfer journeys have different characteristics and therefore need to be treated separately. This is done by determining different rules as outlined throughout this chapter.

The algorithm assigns a different class to inferred OD pairs when they occur repeatedly. This added information attribute contributes to the representativeness of OD pairs as OD pairs that occurred more frequently throughout the validity period of the analysed ticket are clearly
stronger. This can further be used when analysing the validity of the OD extraction algorithm.

Essentially when not using the trip chaining assumption the necessary algorithm to infer OD pairs becomes more complicated as simply not every second journey’s boarding location was the location of alighting of the previous journey. One of the biggest problems is to actually validate the inferred results of the algorithm. Barry et al. (2002) validated the results using aggregated cordon count data which makes it difficult to exactly verify their proposed algorithm. Trepanier et al. (2007a) do not mention any validation method for their results. The proposed algorithm was validated using several different approaches which all can be found in Section 6.11. This includes the first validation of OD estimation using historical EFC data on trip level. This was achieved by collecting data specifically for the validation of this algorithm.

6.9 Algorithm Output

Depending on the type of analysis different formats of output files can be chosen. However, the most versatile option is to store the OD pairs in the database. All records that are stored in the database can be outputted in various formats according to the requirements of the analysis. The following information attributes are stored:

6.9.1 Single Journeys

- Unique OD record ID;
- Class number associated with the OD pair;
- Ticket type;
- Ticket ID (Identifies individual passenger);
- Journey type (Single or Transfer);
- Date of the two boardings;
- Boarding record ID of first boarding;
- Bus stop ID of first boarding;
- Time of boarding of first boarding;
- Route ID of first boarding;
- Direction of first boarding;
- Coarse zone of first boarding;
- Area description of first boarding;
- Boarding record ID of second boarding;
- Bus stop ID of second boarding;
- Time of boarding of second boarding;
- Route ID of second boarding;
- Direction of second boarding;
- Coarse zone of second boarding;
- Area description of second boarding;
- Number of repeated similar OD records found.
6.9.2 Transfer Journeys

The 4 boarding records needed to form a return transfer journey are encoded with A & B for the transfer in one direction and C & D for the transfer journey in the opposite direction.

- Unique OD record ID;
- Class number associated with the OD pair;
- Ticket type;
- Ticket ID (Identifies individual passenger);
- Journey type (Single of Transfer);
- Date of the boarding records;
- Boarding record ID of boarding A, B, C, and D;
- Bus stop ID of boarding A, B, C, and D;
- Time of boarding of boarding A, B, C, and D;
- Route ID of boarding A, B, C, and D;
- Direction of boarding A, B, C, and D;
- Coarse zone of boarding A, B, C, and D;
- Area description of boarding A, B, C, and D;
- Number of repeated similar OD records found.

Having these data elements available it is now possible to commence further analysis as each record defines an OD pair. The analysis of the generated OD pairs will be presented in Chapter 7.

6.10 Testing of the Code and Performance of the Algorithm

This section briefly describes the testing methodology that was applied to test whether the code is working as it was intended. This simply included a number of tests that determine whether all OD pairs are identified using the code of the OD extraction algorithm.

Initially a synthetic dataset was created of which the OD pairs were known in advance simply by manually applying the assumption that the destination of the return journey was the destination of the first journey. The synthetic dataset consisted of 500 boarding records that represented all possible cases (possible and impossible). The following scenarios were tested:

- At least one of each of the 12 + 4 single journey classes;
- At least one of each of the 24 transfer journey classes;
- Several single journeys that consisted of the same return journey so the ‘Repeat’ attribute could be tested;
- Several transfer journeys that consisted of the same return journey so the ‘Repeat’ attribute could be tested;
- Consecutive and non-consecutive boarding records that led to an OD pair;
- Boarding records that used substitutional routes to build an OD pair;
• Boarding records that were not part of an OD pair;
• At least one of all non-valid class types that were initially eliminated.

The results obtained of the algorithm were then manually compared with the previously known outcome. It was possible to conclude that the algorithm is working in accordance with the implemented rules and the main assumptions.

The processing time of the current algorithm mainly dependents on the total number of passenger records and the total boardings recorded of each passenger ticket. The more boardings belong to a ticket the longer the algorithm takes to process the data. The process speed can be increased by using multiple PCs, each using a different data set (Single program multiple data paradigm).

The above statements are based for execution of the method on the same PC.

6.11 Validation of the Results

The dataset has no exit validation information of any of the records and it was therefore difficult to validate the OD pairs at trip level in order to know whether they in fact occurred or not. This section reports on the attempts to validate the algorithm using different reasoning and data sources. This includes the employment survey the DTO carried out in 2001, followed by a simple random sample. Then the analysis of repeated OD pairs will be further investigated to see whether any patterns emerge that could be used to strengthen the validity of the OD results. Finally, a small travel diary study was implemented to collect some data which were then used to validate the OD extraction algorithm.

6.11.1 Survey Results

The DTO survey recorded a total of 64,456 valid responses of which 10,990 were public transport users who provided details about their travel pattern. A total of 83% made a return journey also using a bus as their public transport mode. The remainder use cars, walk or use other forms of transport. A large amount of passengers (96%) of these 83% made no detour throughout their morning journey. This slightly decreases for the evening journey where 82% made no detour. Furthermore, 91.2% of bus passengers travelled with Dublin Bus.

The first validation using the survey data focuses on the time difference of the two journeys (morning and evening) under the premise that these should be similar on an aggregate level. Figure 6.8 shows two histograms representing the results obtained from both analyses. Considering that the DTO survey was aimed at employees only the peak section from 400 to 900 minutes time difference should be analysed in greater detail. This would represent part-time and
full-time employee journeys. Within this interval the extracted journey times and their relative frequencies are very similar.

![Figure 6.8: Time Difference of Journeys](image)

The mean and standard deviation resulting of the OD algorithm sample are 435 and 220 respectively (N=537,705). The survey data resulted in a mean of 534 with a standard deviation of 145 (N=10,990). Figure 6.8a clearly shows the much higher frequencies of journey lengths in the lower region. A followup analysis ignored journey lengths below 400 and above 900 for both samples (N=288,081 for the OD algorithm and N=9,759 for the DTO survey sample). This lead to a mean of 585 with a standard deviation of 102 for the OD algorithm sample and a mean of 573 with a standard deviation of 73 for the survey sample. Although a weak indication but it could be argued that the similar means of both samples provide evidence that the two samples originate from a similar distribution.

A stronger statistical analysis followed using OD matrices from both data sources. This second validation method compares relative frequencies of OD pairs obtained from the algorithm and the survey database. If the relative frequencies of OD pairs from both datasets are similar then it could be argued that the OD algorithm works correctly with regard to obtaining the correct OD locations.

Two OD matrices were compiled; one from each dataset. The sample sizes of OD algorithm and survey data are different which lead to the use of relative frequencies instead of number of journeys by OD pair. Figure 6.9 shows two 3-dimensional bar charts displaying the relative
OD frequencies of the OD algorithm and survey data. The most noticeable difference are the journeys originating from the city centre (Zone 1). The increased number of such journeys in the data set coming from the OD algorithm could be explained by the different focus the data were collected for. The survey clearly targeted employees who travel to and from work whereas the OD algorithm dataset also incorporates social journeys as well as non-work journeys.

![3-D Bar Charts of OD Pairs](image)

Figure 6.9: 3-D Bar Charts of OD Pairs

Initial statistical test showed a correlation of OD pairs of the two datasets when comparing the relative frequencies. This can be seen when plotting the relative frequencies onto a scatter graph with a linear line fitted (slope = 0.937)(see Figure 6.10a). Furthermore a 95% confidence interval was fitted onto the graph. The linear correlation of the data has an adjusted $R^2$ of 0.756 with a standard error of 0.005. Figure 6.10a clearly shows a pattern of increasing variances at the lower region of the graph. A weighted least squares analysis where the weights used were the inverse variances did not remove this pattern. When removing all less significant OD pair results by applying a cut-off value of 0.002 to all relative frequencies the pattern could be removed which is shown in Figure 6.10b. This scatter plot also shows the 95% confidence interval as well as a linear fitted line with slope = 0.933. The adjusted $R^2 = 0.703$ with a standard error of 0.006. Figure 6.10c shows the scatter graph where $\log(\text{rel.freq.}) - \log(1 - \text{rel.freq.})$ was applied to each relative frequency.
The model and the data show that the survey data has a more complete OD matrix than the OD algorithm. The survey data OD matrix has only 80 OD combinations where no passenger travelled to and from compared to 153 in the OD algorithm matrix (out of a total of 324 combinations). This may indicate that the OD algorithm does not identify more complex journeys where passengers have to board more than two buses to get to their final destination. It could also indicate small errors with regard to the manual assignment of stage locations. With geographic coordinates this would be eliminated entirely.

In summary it could be concluded that there is visual and statistical evidence that the two OD matrices are similar.

### 6.11.2 Validation Using a Simple Random Sample

A weaker form of validation is a simple random sample of passengers to analyse their recorded travel data. The aim of this study was to analyse randomly selected travel records of a number
of passengers to see whether the OD pairs could be correct and whether any other reasoning can be applied to provide evidence that the OD algorithm extracted the correct OD information or, on the other hand, why it did not identify the OD pair.

This validation method analysed 883 passenger records belonging to 100 randomly selected tickets. This random subset was then checked for false positives and negatives. A total of 190 OD pairs could be extracted of which 28 were transfer journeys. The algorithm failed to assign a destination to the remaining 435 boarding records. The following could be extracted:

- 4 single OD pairs may have been wrongly identified due to wrong substitutional route assignment;
- 2 transfer OD pairs were wrongly assigned a destination due to lack of substitutional route knowledge;
- 3 single OD pairs were missed to be identified due to wrong assignment of a transfer journey;
- 2 transfer pairs were missed to be identified due to wrong assignment of a transfer journey;
- 85 boarding records were the only journey carried out on the day and were not similar to any of the identified OD pairs;
- 41 boarding records were the only record assigned to the ticket;
- 124 boarding records could not be assigned a destination as the directional attribute indicated that the only two journeys made that day were taken into the same direction.
- 6 OD pairs could have been assigned if combinations of single and transfer journeys were considered by the algorithm;
- 124 transfer journey records did not have a matching return transfer journey;
- It was not possible to classify the remaining unassigned 43 records.

The following performance measures can be calculated:

\[
\text{recall} = \frac{\text{number of OD identified}}{\text{number of true OD pairs in dataset}} \quad (6.11.1)
\]

and

\[
\text{precision} = \frac{\text{number of true OD identified}}{\text{number of OD pairs identified}} \quad (6.11.2)
\]

The higher the values for recall and precision the better the performance of the algorithm. Considering a split between transfer and single journey then approximately 357 OD pairs are contained in the sample. Therefore \( \text{recall} = \frac{184}{357} = 0.52 \) and \( \text{precision} = \frac{184}{190} = 0.968 \) As the aim of the algorithm was to achieve a high precision rate by not introducing too many assumptions this result is very good. Furthermore, the algorithm successfully extracts simple trip chains
but has problems with complex trip chains where the passenger carries out a multi-purpose trip chain.

Although this method was a somewhat weaker form of validation it still highlights the weaknesses and strengths of the algorithm. In particular, the algorithm fails when passengers made a single journey to reach their destination in one direction and then replace this single journey by a transfer journey to make the return to their origin.

6.11.3 Repeated OD Pairs
Repeated OD pairs indicate a strong representativeness as the return journey is obviously part of a passenger's routine. It could be argued that the more often a journey is repeated the higher the probability is that the inferred OD destination is in fact correct. The algorithm counted how often each OD pair was repeated and included the variable as part of the output information. Table 6.15 shows the cross-tabulation of the various cases and the number of recorded repeats for single journeys. The values of Table 6.15 are represented in aggregated format (in increments of 5).

It could be argued that more repeatedly identified OD pairs show that the OD extraction algorithm functions well. For example was it possible to determine, when analysing weekly adult and student tickets, that the algorithm identified 75% of OD pairs as repeated OD pairs.

<table>
<thead>
<tr>
<th>OD Cases</th>
<th>Repeated OD Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-1</td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>16,997</td>
</tr>
<tr>
<td>2</td>
<td>1,144</td>
</tr>
<tr>
<td>3</td>
<td>67,226</td>
</tr>
<tr>
<td>4</td>
<td>5,988</td>
</tr>
<tr>
<td>5</td>
<td>64,118</td>
</tr>
<tr>
<td>6</td>
<td>3,015</td>
</tr>
<tr>
<td>7</td>
<td>2,401</td>
</tr>
<tr>
<td>8</td>
<td>202</td>
</tr>
<tr>
<td>9</td>
<td>52,395</td>
</tr>
<tr>
<td>10</td>
<td>6,537</td>
</tr>
<tr>
<td>11</td>
<td>66,739</td>
</tr>
<tr>
<td>12</td>
<td>6,779</td>
</tr>
</tbody>
</table>

Furthermore there is a pattern with regard to the number of repeats per week day. It could be argued that passenger's carry out repeated journeys during the week but break out of this routine at weekends. A total of almost 50% of journey OD pairs were not repeated at Saturday and Sunday compared to 36% on Tuesday, Wednesday, and Thursday.
6.11.4 Travel Diary Study

The lack of a results set to the main recorded EFC dataset is certainly one of the main prohibitors with regard to proving the validity of the algorithm and its inferred results. It was therefore decided to collect some public transport data in a small scale travel diary study where 40 people (4 adults and 36 students) agreed to participate. The aim was to get approximately 384 records as this is the required sample size for a study with a 5% sampling error and a 95% confidence interval.

The importance of this study was communicated and daily reminder emails were sent to all participants. Furthermore a brief outline how to record the data and to what detail was provided. The participants recorded their travel data daily via a purposely developed web site where each of the participants had a unique user account. The web site was designed with the minimisation of entry errors in mind. This included a check that the date entered was within the 7 day test period but not in the future, the route was an existing route and the time was in a valid format. Additionally they were given a log book to record each journey detail as it occurred. After logging on to the web site account the user could enter date, time, route, origin, and destination. The system also recorded the entry date and the order of entry automatically without the knowledge of the participants. The direction as well as boarding stage was added manually (using a map) to each of the recorded journeys after the survey finished. Already entered journeys were visible to the participant. A pilot study was used to test the web site.

The aim of this study was to obtain a dataset which is as realistic as possible and of which the destination of each record is known. It was then possible to run the algorithm on the small set of recorded data and in turn produce OD pairs which could then be compared to the destination provided by the participants.

A total of 327 records were created by 27 participants. After checking the data records individually and adding bus stage encoding the location of the origin and destination in terms of area description and coarse zone was determined. Following this data preparation stage the transfer journey algorithm was applied followed by the OD extraction algorithm. The following results were obtained:

The algorithm inferred a destination to 172 records with a total of 12 transfer journeys and 62 single journey OD pairs. The wrong stage was inferred to 8 single journey OD pairs. The error was small and only inferred 1-3 stages (100-400 meters) away from the actual destination. This was mainly due to walking parts of the distance either because no service was available at the time or due to a different initial purpose such as a short visit at a convenience store. This did not occur to any of the transfer OD pairs. Furthermore, a wrong destination was assigned to 4 single journeys. Once the assumption that leads to class $y_{16}$ caused this inference error and three times a wrong substitutional route combination lead to the misinterpretation. One of the
transfer OD pairs was incorrectly inferred due to an incorrect substitutional route. This could have been avoided using geographic coordinates while inferring substitutional routes. None of the single OD pairs were entirely wrong identified. Five participants only recorded one journey over the entire test period and no further estimation of the destination was possible. A total of 58 journeys were made in one direction only. Out of these 58 journeys, 48 were made in the evening going to a suburban centre or the city centre. As no return journey was recorded it was not possible to find a simple trip chain to infer the destination. A total of 75 records could not be inferred with a destination as no return journey matched with the same or substitutional route. The remaining 17 boarding records seemed to have been part of a complex trip chain where 3 boarding records were carried out. As earlier identified, these complex trip chains are impossible to be identified using the proposed algorithm. The algorithm by, for example, Barry et al. (2002) or Zhao et al. (2007) would infer these much better. One conclusion, which will be followed up later on, could well be that a combination of the proposed algorithm with one of the earlier introduced methods may extract a better OD estimation result.

Equations 6.11.1 and 6.11.2 are used to calculate the performance measures recall and precision. The measure for \( \text{recall} = \frac{172}{327} = 0.53 \) with a \( \text{precision} = \frac{160}{172} = 0.93 \). This precision value is correct when small error such as wrong identification by a couple of stages is ignored. Otherwise \( \text{precision} = \frac{144}{172} = 0.84 \). These results are similar to the simple random sample validation above. The precision value is high which was as expected due to the initial direction to concentrate on the identification of simple trip chains. It would be interesting to see how the algorithm performs with less dated EFC data. White (2008) states that 80% of trips consist of simple trip chains. It is therefore expected that when the utilisation of magnetic/smart cards is higher that the recall value is also higher.

### 6.11.5 Validation Based on the OD Results

It could be argued that the results themselves could prove the validity of the algorithm by analysing some performance measures or key aspects. The following are two such analyses where once the time difference in hours between the first journey and the second (return) journey is calculated. This is followed by a brief section that explores the distribution of boarding times of first and second journeys.

Figure 6.11 also shows the distribution of time differences between two boardings. The results are broken down by the categories child, student and adult. The vertical axis shows the percentage of each category. The most obvious finding from this analysis is that the time difference between both boarding records carried out by children is between 7 and 9 hours the highest. Students have an almost even distribution between 3 and 10 hours whereas adults have a small peak at hour 3 and hour 10.
Figure 6.11: Time Difference between OD Pair Boarding Records by User Category

Figure 6.12: Boarding Time Distribution of First and Second Boarding Time Records

Figure 6.12 shows the distribution of boarding times for both first and second boarding time of each OD pair. Therefore the total frequency of each line is the same. There is a steep continuous increase of boardings between 6.30 a.m. and 8.10 a.m.. The distribution of the second boarding times is not as defined and stretches from 2.30 p.m. to 6.30 p.m..

Section 6.12 focuses on the limitations and boundaries of the algorithm and its assumptions in greater detail. Furthermore, it outlines several scenarios where the algorithm would fail to function correctly.
6.12 Limitations and Boundaries of Data and the Algorithm

Analysis that is based on assumptions always has limitations and boundaries. This section deals with identifying these limitations and boundaries. This is mainly done focusing on the data that the algorithm is based on and then the algorithm itself. It is commonly known that the results can only be as good as the data. Therefore even a perfect algorithm using no assumptions cannot provide true results if the data used by the algorithm are not correct. In addition it is important that the necessary data attributes are available.

6.12.1 Main Assumption

The main assumption of the algorithm was thoroughly described and analysed throughout this chapter. As mentioned before the aim of this project was to create OD pairs that did in fact occur. Trip chaining was therefore not applied although it would have provided a much higher OD identification rate. But, the results set would not be as representative (see Section 6.2). Each added assumption will take away from this representativeness of the OD information. Even with the main assumption used in this project there are several scenarios when the algorithm actually produces the wrong result. For example, it cannot be identified whether a passenger walked half way towards his final destination before boarding a bus and then boarded a bus to his/her final destination in the evening. In this case his/her final destination would be inferred as the boarding location of the morning. Naturally this scenario could also entail a short walk in the evening before boarding the bus which in turn would bias the algorithm’s estimation of the destination of the journey in the morning. It therefore has to be acknowledged that the prediction of the destination and therefore all following derived attributes such as in-vehicle time, waiting time or arrival time may not be entirely correct. Another example where the algorithm would fail to infer the OD pair is when the passenger made a transfer journey to get to his/her final destination but, for whatever reasons, made the return journey without transferring. Therefore any combination of transfer and single journeys will not lead to an OD pair extraction.

6.12.2 Assumption – Subroutes

The next limitation is the assumption implied with regard to the substitutional route identification. There is a small error when classifying alternative routes as mentioned in Section 6.5. This small error may be carried forward into the OD analysis. It would be easy to create a scenario that disproves the assumption especially as individual passenger information is not incorporated into each OD estimation using geographic coordinates.

This assumption may have a negative impact on defining the passenger’s destination when the two routes were in fact not substitutional or when the algorithm failed to define two routes as substitutional.
6.12.3 Transfer Journeys

The need to know which boardings were part of a transfer journey was outlined throughout the chapter. As the procedure for identifying OD pairs is different for single and transfer journeys the correct classification of the two journey types has a direct impact on the OD inference. Therefore, wrongly identified (either false positive or false negative) passenger boarding records will fail to produce a correct result with regard to the OD inference.

6.12.4 Data Source – Lack of Data

The data provided from Dublin Bus is a comprehensive dataset with a minimum of recording errors caused by the reading device. However some missing data blocks were identified throughout the development of the OD estimation algorithm and the analysis stage. The first problem with the data is that geographic coordinates are not recorded. This means that the time and location of boarding depends on whether the bus driver keyed in the correct bus stage or not (see Section 4.6.1). This is not as much of a problem for the actual estimation of the OD information but more for the following analysis stage. There is no need for a bus driver to update his/her location on the system if no passenger boards the vehicle. However, this lack of information can cause a problem throughout the calculation of derived attributes such as in-vehicle time or passenger arrival time. Geographic coordinates that are stored along with time stamps could solve this problem without any interaction of the bus driver. AVL in combination with Real Time Passenger Information (RTPI) is implemented on many networks. However, the coordinates of the vehicle are generally only stored temporarily. It is therefore advised to store records when a bus passes a bus stop. A continuous recording of such information could further be used to exactly present the flow and path of the bus including dwell time at the bus stop, speed and passenger miles for a network, route or route segment.

The second missing dataset was the representation of the entire network and its routes and bus stops with geographic coordinates. This limited the algorithm to aggregated spatial identifiers. Derivable data attributes such as distance of bus stops when transferring onto another bus service were not possible to calculate due to the lack of geographic coordinates.

6.12.5 Small Sample

Only 14% of Dublin Bus’ boarding records were created using a magnetic strip card. The remainder were cash paying customers. This may cause a sampling problem when aggregated performance measures need to be produced. Furthermore it could lead to problems when findings from magnetic strip card data are considered for implementation on a network wide level. It is not analysed and with the given dataset it cannot be analysed whether findings from the magnetic strip card data is representative for all passengers. Surveys for example indicate that
cash paying passengers make less transfer journeys than magnetic strip card holders. Furthermore, it is believed that the general travel behaviour of magnetic strip card holders is different to cash paying customers.

However, integrated ticketing mainly using smart cards has progressed significantly over the last couple of years and in some bus networks (e.g. London) cash payments are responsible for less than 5% of all transactions (TFL, 2007). Rahbee (2003) states that 91% of all rail trips are non-cash transactions.

### 6.12.6 Unknown Routes and Ticket Types

The database contains routes and ticket types which are not defined. This leads to problems in the analysis stage when results need to be interpreted correctly which becomes more difficult when part of the information is not known. However this is solely a problem that occurred because Dublin Bus did not actively participate in this project.

### 6.13 Summary

This chapter outlined the methodology that was used to extract OD pairs from EFC datasets. Initially the concept and the assumptions which the OD extraction algorithm is based upon were stated. The main assumption is that the boarding location for a journey in one direction is the location of alighting in the opposite direction of the journey. A number of attributes such as date, route (sub route) and direction are used to infer the OD information.

Substitutional routes were defined and their importance with regard to the OD extraction algorithm was outlined (see Section 6.5). An algorithm for dynamically identifying substitutional routes was developed. Two solutions were proposed and tested with the results focusing on the 'error of impact'. A sample was used to test the two approaches. Two cut-off parameters were introduced; 'number of occurrences' and 'calculated weight'. A detailed study introduced a method to calibrate these cut-off points in such a manner that the errors of impact are minimised. The cut-off points for this project were > 50 passenger combination pairs for each route combination pair and a calculated weight of 2.0 or larger. Solution 2 (OR) was identified as the most suited algorithm to calculate the weights.

Travel classes which were introduced in Section 6.5 are all logically possible attribute combinations. Decision Tables were used to identify the set of combinations that later built the rule base. A different class type was assigned to each travel scenario which will later identify the type of OD pair. 12 different combinations were identified for single journeys and 24 different types were identified for transfer journeys (see Table 6.9 and Table 6.11).
The algorithm structure is outlined in Sections 6.6 and 6.7. This part of the chapter uses the previously defined techniques and algorithm together to build the OD extraction algorithm. The individual processes were described in detail.

This was followed by the validation of the algorithm using various methods all producing promising results. The limitations and boundaries of all assumptions were explored to see what impact they may have on the project results.

In order to further improve the OD ratio the following information may be necessary:

- Geographic coordinates would certainly contribute to the application of new assumptions. It is therefore important to incorporate such information for further stages of the development;
- If smart cards are employed, data attributes such as the passenger’s address and other demographic information could be used to improve or add new assumptions that will lead to the improvement of the algorithm. It could be imagined that assumptions also change depending on ticket type and age group of the passengers. The assumption which was introduced in Section 6.7.3 focused on a special scenario where only 2 boardings were carried out per ticket. This could be extended to base further assumptions on the actual patterns of the customers. For example, students have different travel patterns than school children and working adults. More information on inferring passenger behaviour using historical smartcard data was shown by Morency et al. (2006) and Trepanier et al. (2007a) and Trpanier and Agard (2007b).

The following chapter focuses on the analysis of the results as well as the estimation of various performance measures.
Chapter 7

OD Results and Their Application to Determine In-Vehicle Time

7.1 Introduction

The previous chapters described the methodology to extend the database with information that was previously unknown. This included spatial information (i.e. location of each bus stop), a transfer journey identifier and destination information of over 1 million individual passenger journeys.

This chapter focuses on these results which are further detailed and explored. In particular the newly created OD information will be used to demonstrate possible usage to support transport planners throughout the decision making process. This will be done by proposing a method to determine in-vehicle time and its variability using the newly inferred destination attribute as well as an analysis on total travel time and waiting time of transfer journeys.

As mentioned in previous sections the EFC data of Dublin Bus have a rather low content of magnetic card users and therefore some of the applications of the data may not be as representative as those of transport networks with a higher percentage of magnetic card utilisation. Therefore, the proposed method to extract in-vehicle time serves to demonstrate an example of the capabilities of the dataset. The actual calculated performance measures can not be applied to the Dublin Bus network due to the reason mentioned above and due to the dated dataset.

7.2 Summarised Results of the OD Extraction Algorithm

This section will provide general figures about the outcome of the OD extraction algorithm mainly with regard of its performance. There were 2,274,096 magnetic card boardings recorded in October 1999, of which 1,498,685 were single journeys and 775,411 were transfer journeys.
7.2. SUMMARISED RESULTS OF THE OD EXTRACTION ALGORITHM

While there are numerous different ticket types available not all of them favour the OD extraction. For example, tickets which are only valid for one journey cannot be used for OD extraction as data of a return journey are missing. These tickets were removed for the purpose of the following analyses.

7.2.1 Results by Ticket Type

Table 7.1 shows a summary of the results received from the OD extraction algorithm grouped by Ticket Type ID, which together with the Ticket Name identify the analysed ticket. Total Journeys, Single Journeys and Transfer Journeys show the total number of boardings that were recorded in the month of October 1999 grouped by each ticket type. The columns Total Inferred Destinations, Inferred Single Destinations and Inferred Transfer Destinations show the total numbers of destinations that were identified by the algorithm. Percentage Single shows the ratio of identified single destinations and total number of single journeys. Inferred Single Destinations was derived by multiplying the identified OD pairs from class 1 to 14 by two and adding the number of identified destinations of classes 15 and 16. Percentage Transfer shows the ratio of identified Inferred Transfer Destinations and total number of transfer journey records. Inferred Transfer Destinations was derived by multiplying all identified transfer journey OD pairs by four as it represents a return journey which consists of four boarding records. The column percentage total indicates the total inferred destinations over the total possible destinations.

The table is in descending order of Percentage Total omitting any tickets that have less than 1,100 records or are single use tickets.

As mentioned above, some ticket types suit the OD extraction more than others. More passenger boardings per ticket generally result in a better result with regard to the estimation of OD information. The nature of the algorithm is that the more routine travel patterns a passenger has the more likely a destination can be applied to the boarding record. The following conclusions can be drawn from the summarised result shown in Table 7.1:

- Foreign student tickets show a high total percentage of OD estimation. The highest percentage with 78% means that out of 10 journeys 7.8 could be inferred with a destination. The algorithm is programmed in such a way that the greater the routine behaviour of a passenger the better the achieved OD results will be. It could therefore be assumed that foreign students have on average a higher routine behaviour than the remaining passengers. This could be caused by their smaller social network. On average, 1, 2, 3, and 4
Foreign Student tickets achieve an OD estimation rate of 71% based on over 70,000 individual boarding records;

Table 7.1: Summary Figures of OD Extraction Results

<table>
<thead>
<tr>
<th>Ticket Name</th>
<th>Total Journeys</th>
<th>Single Journeys</th>
<th>Transfer Journeys</th>
<th>Total Inferred Destinations</th>
<th>Inferred Single Destinations</th>
<th>Inferred Transfer Destinations</th>
<th>Percentage Single</th>
<th>Percentage Transfer</th>
<th>Percentage Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign Student-3-week</td>
<td>736.6</td>
<td>680.6</td>
<td>83.0</td>
<td>5979</td>
<td>5891</td>
<td>88</td>
<td>87%</td>
<td>11%</td>
<td>78%</td>
</tr>
<tr>
<td>Foreign Student-4-week</td>
<td>5250.5</td>
<td>4280.3</td>
<td>970.2</td>
<td>37119</td>
<td>35387</td>
<td>1732</td>
<td>83%</td>
<td>18%</td>
<td>71%</td>
</tr>
<tr>
<td>Foreign Student-2-week</td>
<td>9992</td>
<td>8116</td>
<td>1876</td>
<td>6692</td>
<td>6548</td>
<td>144</td>
<td>81%</td>
<td>8%</td>
<td>67%</td>
</tr>
<tr>
<td>Foreign Student-1-week</td>
<td>2001</td>
<td>1682</td>
<td>319</td>
<td>1317</td>
<td>1249</td>
<td>68</td>
<td>74%</td>
<td>21%</td>
<td>66%</td>
</tr>
<tr>
<td>10-JourneyFeeder-adult</td>
<td>5176</td>
<td>5140</td>
<td>36</td>
<td>3102</td>
<td>3102</td>
<td>0</td>
<td>60%</td>
<td>0%</td>
<td>60%</td>
</tr>
<tr>
<td>Unknown</td>
<td>2869</td>
<td>1960</td>
<td>909</td>
<td>1565</td>
<td>1365</td>
<td>200</td>
<td>70%</td>
<td>22%</td>
<td>55%</td>
</tr>
<tr>
<td>Weekly Student Travelwide</td>
<td>55503</td>
<td>35316</td>
<td>20187</td>
<td>30033</td>
<td>24101</td>
<td>5932</td>
<td>68%</td>
<td>29%</td>
<td>54%</td>
</tr>
<tr>
<td>Monthly Adult Travelwide</td>
<td>8747</td>
<td>5266</td>
<td>3481</td>
<td>4695</td>
<td>3975</td>
<td>720</td>
<td>75%</td>
<td>21%</td>
<td>54%</td>
</tr>
<tr>
<td>Monthly Adult SH* Bus/Rail</td>
<td>2847</td>
<td>23264</td>
<td>9583</td>
<td>16606</td>
<td>14890</td>
<td>1716</td>
<td>64%</td>
<td>18%</td>
<td>51%</td>
</tr>
<tr>
<td>Monthly Student SH* Bus/Rail</td>
<td>218262</td>
<td>153266</td>
<td>64996</td>
<td>111969</td>
<td>10377</td>
<td>11592</td>
<td>65%</td>
<td>18%</td>
<td>51%</td>
</tr>
<tr>
<td>Unknown</td>
<td>54495</td>
<td>36531</td>
<td>17964</td>
<td>27830</td>
<td>23722</td>
<td>4108</td>
<td>65%</td>
<td>23%</td>
<td>51%</td>
</tr>
<tr>
<td>Weekly Student Cityzone</td>
<td>484123</td>
<td>302990</td>
<td>181533</td>
<td>247331</td>
<td>190923</td>
<td>48308</td>
<td>66%</td>
<td>27%</td>
<td>51%</td>
</tr>
<tr>
<td>Annual Adult Travelwide</td>
<td>10159</td>
<td>6028</td>
<td>4131</td>
<td>5174</td>
<td>4430</td>
<td>744</td>
<td>73%</td>
<td>18%</td>
<td>51%</td>
</tr>
<tr>
<td>Weekly Adult Bus</td>
<td>18676</td>
<td>10990</td>
<td>7686</td>
<td>9257</td>
<td>7285</td>
<td>1972</td>
<td>66%</td>
<td>26%</td>
<td>50%</td>
</tr>
<tr>
<td>Monthly Adult (Aer Lingus)</td>
<td>45790</td>
<td>25946</td>
<td>19844</td>
<td>22918</td>
<td>18434</td>
<td>4484</td>
<td>71%</td>
<td>23%</td>
<td>50%</td>
</tr>
<tr>
<td>Annual StaffBus</td>
<td>12875</td>
<td>9685</td>
<td>3190</td>
<td>6340</td>
<td>5856</td>
<td>484</td>
<td>60%</td>
<td>15%</td>
<td>49%</td>
</tr>
<tr>
<td>Annual CIE OAP* Bus/Rail</td>
<td>3568</td>
<td>2614</td>
<td>954</td>
<td>1676</td>
<td>1520</td>
<td>144</td>
<td>58%</td>
<td>12%</td>
<td>47%</td>
</tr>
<tr>
<td>Annual Bus/Rail</td>
<td>3824</td>
<td>2757</td>
<td>1067</td>
<td>1765</td>
<td>1633</td>
<td>132</td>
<td>59%</td>
<td>12%</td>
<td>46%</td>
</tr>
<tr>
<td>Weekly Adult SH* Bus/Rail</td>
<td>43203</td>
<td>31621</td>
<td>11582</td>
<td>18895</td>
<td>17623</td>
<td>1272</td>
<td>56%</td>
<td>11%</td>
<td>44%</td>
</tr>
<tr>
<td>Weekly Adult Cityzone</td>
<td>471577</td>
<td>240519</td>
<td>231058</td>
<td>202337</td>
<td>142865</td>
<td>59472</td>
<td>59%</td>
<td>26%</td>
<td>43%</td>
</tr>
<tr>
<td>Weekly Adult MH Bus/Rail</td>
<td>2216</td>
<td>1361</td>
<td>855</td>
<td>892</td>
<td>700</td>
<td>192</td>
<td>51%</td>
<td>22%</td>
<td>40%</td>
</tr>
<tr>
<td>Schoolchild 2-Journey</td>
<td>113440</td>
<td>108586</td>
<td>4854</td>
<td>45738</td>
<td>45694</td>
<td>44</td>
<td>42%</td>
<td>1%</td>
<td>41%</td>
</tr>
<tr>
<td>Rambler (3 Day Bus only)</td>
<td>272158</td>
<td>140815</td>
<td>131343</td>
<td>109877</td>
<td>76201</td>
<td>33676</td>
<td>54%</td>
<td>26%</td>
<td>40%</td>
</tr>
<tr>
<td>Weekly Adult GH* Bus/Rail</td>
<td>2605</td>
<td>1669</td>
<td>936</td>
<td>970</td>
<td>862</td>
<td>108</td>
<td>52%</td>
<td>12%</td>
<td>37%</td>
</tr>
<tr>
<td>Adult 2-Journey(23 Stages)</td>
<td>29093</td>
<td>28611</td>
<td>482</td>
<td>10728</td>
<td>10724</td>
<td>4</td>
<td>37%</td>
<td>1%</td>
<td>37%</td>
</tr>
<tr>
<td>Scholar 2-Journey</td>
<td>19762</td>
<td>18993</td>
<td>769</td>
<td>7405</td>
<td>7405</td>
<td>0</td>
<td>39%</td>
<td>0%</td>
<td>37%</td>
</tr>
<tr>
<td>Adult One Day Travelwide</td>
<td>32602</td>
<td>14720</td>
<td>17882</td>
<td>12097</td>
<td>6837</td>
<td>5260</td>
<td>46%</td>
<td>29%</td>
<td>37%</td>
</tr>
<tr>
<td>Family One Day Travelwide</td>
<td>8096</td>
<td>4017</td>
<td>4079</td>
<td>3030</td>
<td>1898</td>
<td>1132</td>
<td>47%</td>
<td>28%</td>
<td>36%</td>
</tr>
<tr>
<td>Adult Bus/Rail SH* -Day</td>
<td>3076</td>
<td>2338</td>
<td>738</td>
<td>1103</td>
<td>1011</td>
<td>92</td>
<td>43%</td>
<td>12%</td>
<td>36%</td>
</tr>
<tr>
<td>Family Bus/Rail SH* -Day</td>
<td>9444</td>
<td>1329</td>
<td>615</td>
<td>707</td>
<td>607</td>
<td>100</td>
<td>46%</td>
<td>16%</td>
<td>36%</td>
</tr>
<tr>
<td>Weekly Adult LH* Bus/Rail</td>
<td>1300</td>
<td>692</td>
<td>608</td>
<td>467</td>
<td>343</td>
<td>124</td>
<td>50%</td>
<td>20%</td>
<td>36%</td>
</tr>
<tr>
<td>Adult 2-Journey(12 Stages)</td>
<td>34719</td>
<td>34078</td>
<td>641</td>
<td>12570</td>
<td>12566</td>
<td>4</td>
<td>37%</td>
<td>1%</td>
<td>36%</td>
</tr>
<tr>
<td>Adult 2-Journey(7 Stages)</td>
<td>98210</td>
<td>89840</td>
<td>3370</td>
<td>32525</td>
<td>32525</td>
<td>0</td>
<td>36%</td>
<td>0%</td>
<td>35%</td>
</tr>
<tr>
<td>Adult 2-Journey(3 Stages)</td>
<td>78710</td>
<td>71361</td>
<td>7349</td>
<td>24569</td>
<td>24565</td>
<td>4</td>
<td>34%</td>
<td>0%</td>
<td>31%</td>
</tr>
<tr>
<td>Adult 2-Journey(23+ Stages)</td>
<td>5121</td>
<td>5053</td>
<td>68</td>
<td>1602</td>
<td>1602</td>
<td>0</td>
<td>32%</td>
<td>0%</td>
<td>31%</td>
</tr>
</tbody>
</table>

* OAP - Old Age Pensioner; SH - Short Hop; MH - Medium Hop; LH - Long Hop; GH - Giant Hop

- Generally, monthly tickets result in a higher total percentage of OD estimation than weekly tickets. Monthly tickets often recorded more passenger boardings than weekly tickets and therefore provide the algorithm more data to extract routine behaviour;
- 2-Journey tickets have a relative low percentage of transfer journeys. The average total percentage of OD estimation is approximately 36% which is considerably lower than the total average of 46%. Rechargeable smart cards are nowadays very common and might prevent the frequent occurrence of only two journeys on an individual ticket. 2-Journey
tickets were used to make over 16% of the network's magnetic card boardings;

- Due to small errors in the EFC data, unique IDs were duplicated and therefore result in some OD pairs. Although this is an error it does not cause any problems with regard to the OD estimation. The algorithm proved to be working as in this case only 18 errors occurred which is less than 0.00005% and therefore neglectable;

- The OD extraction percentage is higher for single journeys than it is for transfer journeys. This is due to the increased complexity of transfer journeys as four boarding records have to be compared and matched. Including all ticket types the average OD extraction percentage of single journeys is 57% compared to 24% of transfer journeys. This considerable difference could also be caused by scenarios where the passenger makes a transfer journey to the desired location and then decides to return via a through bus peak-time only service. The algorithm currently does not facilitate the OD extraction of such scenarios;

- Some ticket types are unknown. It could be assumed that if the number of transfer journeys of the unknown tickets is very low then they are more than likely 2-Journey tickets. Tickets with higher transfer journeys are likely to be weekly or monthly tickets;

- It is noteworthy to add that these figures are average figures of several thousand different passengers. Many passenger tickets have an OD extraction percentage of over 95% and in some cases even all journeys of a passenger were assigned a destination. Furthermore not all recorded boardings have a return journey due to car pooling or other modes of transport that offer an alternative to the bus services. For these journeys it is not possible to assign a destination without violating existing or adding more assumptions.

A total of 46% of all records were assigned with a destination. There are several reasons for this not being higher. The following is a list of these:

- The main concern lies with the dated data set. Since 1999 the usage of magnetic strip cards has increased considerably in most networks. Therefore, the project data set used for this study contains an unusually high percentage of transfer journeys (over 34%) which in turn are much more difficult to assign a destination as single journeys. Furthermore, in 1999 the pricing structure of weekly and monthly tickets of magnetic strip cards was such that it only favoured passengers who made more than 10 weekly single journeys. Again, this caused the number of passengers that carried out single routine journeys to and from work to pay cash. Introductions such as the pay as you go (PAYG) Oyster card in London could reduce this effect and therefore allow for an improved OD
7. OD RESULTS AND THEIR APPLICATION TO DETERMINE IN-VEHICLE TIME

estimation ratio. Most frequently used weekly and monthly tickets achieve a higher identification of destinations than other tickets. This is reflected in the summary table above where it can be seen that the main weekly and monthly tickets achieved an estimation rate of around 65% for single journeys and over 50% for all journeys.

- The number of tickets was greatly reduced in recent years which will further contribute to a higher destination estimation ratio.

- This study only focused on data of one month. However, there are tickets that span over two months which results in some records of passengers that were recorded in either September or November being disregarded. This mainly matters for the classes 13, 14, 15 and 16 as these analyse past boarding behaviour which is not available in its entirety due to the neglect of the neighbouring months.

This is also addressed in Section 7.3 when the results are compared to other algorithms.

7.2.2 Results by OD Classes

The algorithm was designed in such a way that each OD pair was assigned to an OD class. There were 16 classes for single journey OD pairs and 24 classes for transfer OD pairs. This section shows how the inferred OD pairs were spread over the various classes and what this means with regard to their representativeness. Summary figures will also be included. The various classes were introduced and discussed in Section 6.5.

Single Journeys

All identified OD pairs were assigned to the 16 classes (S1-S16) which is shown in Table 7.2. The last two columns indicate how many OD pairs were assigned to each class and the percentage of this figure with respect to the total number of inferred single journey OD pairs. 92% of all single journey OD pairs classes were recorded in consecutive timely sequential order. 71% of all identified single journey OD pairs were repeated either on the same day or throughout the period of ticket validity.

Classes where the return journey records were in consecutive order and were also repeated are the best indicators for a strong OD pair. It simply means that the passenger undertook the same return journey frequently and did not travel anywhere else in between the two OD boardings. Table 7.3 shows the strong and weak OD classes. The table provides a summary of identified ODs of which some were categorised into four different groups. The first two rows show the strongest two OD categories where the passenger used the same or a substitutional
route, made the journey repeatedly and did not travel between the two OD boardings. The records that could be assigned to this category were responsible for 54% of all identified OD pairs ignoring classes 13-16 as the repeated and/or order attribute are not checked for these classes). The bottom two rows of Table 7.3 show the two weakest categories. In these cases the passenger’s journeys were not repeated and were not in consecutive order which indicates that another recording took place between the two OD boardings leaving room for the algorithm to ‘misinterpret’ the records.

Table 7.2: Single Journey OD Results Grouped by Classes

<table>
<thead>
<tr>
<th>Class</th>
<th>Day</th>
<th>Route</th>
<th>Direction</th>
<th>Repeated</th>
<th>Order</th>
<th>OD Pairs</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Same</td>
<td>Same</td>
<td>Different</td>
<td>Same</td>
<td>Consecutive</td>
<td>28,180</td>
<td>5.7%</td>
</tr>
<tr>
<td>S2</td>
<td>Same</td>
<td>Same</td>
<td>Different</td>
<td>Same</td>
<td>Non-consecutive</td>
<td>1,910</td>
<td>0.4%</td>
</tr>
<tr>
<td>S3</td>
<td>Same</td>
<td>Same</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Consecutive</td>
<td>91,449</td>
<td>18.6%</td>
</tr>
<tr>
<td>S4</td>
<td>Same</td>
<td>Same</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Non-consecutive</td>
<td>8,448</td>
<td>1.7%</td>
</tr>
<tr>
<td>S5</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Not repeated</td>
<td>Consecutive</td>
<td>64,118</td>
<td>13.0%</td>
</tr>
<tr>
<td>S6</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Not repeated</td>
<td>Non-consecutive</td>
<td>3,015</td>
<td>0.6%</td>
</tr>
<tr>
<td>S7</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Same</td>
<td>Consecutive</td>
<td>3,388</td>
<td>0.7%</td>
</tr>
<tr>
<td>S8</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Same</td>
<td>Non-consecutive</td>
<td>286</td>
<td>0.1%</td>
</tr>
<tr>
<td>S9</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Consecutive</td>
<td>62,332</td>
<td>12.7%</td>
</tr>
<tr>
<td>S10</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Non-consecutive</td>
<td>7,824</td>
<td>1.6%</td>
</tr>
<tr>
<td>S11</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Not repeated</td>
<td>Consecutive</td>
<td>66,739</td>
<td>13.6%</td>
</tr>
<tr>
<td>S12</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Not repeated</td>
<td>Non-consecutive</td>
<td>6,779</td>
<td>2.0%</td>
</tr>
<tr>
<td>S13</td>
<td>Diff</td>
<td>Same</td>
<td>Different</td>
<td>NA</td>
<td>NA</td>
<td>4,551</td>
<td>1.4%</td>
</tr>
<tr>
<td>S14</td>
<td>Diff</td>
<td>Sub</td>
<td>Different</td>
<td>NA</td>
<td>NA</td>
<td>4,150</td>
<td>0.9%</td>
</tr>
<tr>
<td>S15</td>
<td>Diff</td>
<td>Same</td>
<td>Different</td>
<td>Repeated</td>
<td>NA</td>
<td>51,219</td>
<td>10.4%</td>
</tr>
<tr>
<td>S16</td>
<td>Diff</td>
<td>Sub</td>
<td>Different</td>
<td>Repeated</td>
<td>NA</td>
<td>87,300</td>
<td>17.8%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>491,688</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

The algorithm therefore works very well with the identification of OD pairs consisting of single journeys. Over 54% of all inferred OD pairs are considered to be representative due to their consecutive order and their repetitive nature. It could also be argued that classes 15 and 16 are strong as the OD information that falls into these classes was only possible because very similar OD pairs were already identified. However, this was not quantitatively included in this analysis.

Table 7.3: Strong and Weak OD Classes for Single Journeys

<table>
<thead>
<tr>
<th>Route</th>
<th>Repeated</th>
<th>Order</th>
<th>Identified ODs</th>
<th>Percentage of all ODs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same</td>
<td>Yes</td>
<td>Consecutive</td>
<td>119,629</td>
<td>35%</td>
</tr>
<tr>
<td>Sub</td>
<td>Yes</td>
<td>Consecutive</td>
<td>65,720</td>
<td>19%</td>
</tr>
<tr>
<td>Same</td>
<td>No</td>
<td>Non-consecutive</td>
<td>3,015</td>
<td>1%</td>
</tr>
<tr>
<td>Sub</td>
<td>No</td>
<td>Non-consecutive</td>
<td>6,779</td>
<td>2%</td>
</tr>
</tbody>
</table>
Table 7.4 shows the results of inferred transfer OD pairs grouped by all 24 classes. The last two columns show how many OD pairs were assigned to each class and the percentage of this figure with respect to the total number of inferred transfer journey OD pairs. A total of 86% of all transfer journey OD pairs were recorded in consecutive timely sequential order and 47% were repeated either on the same day or throughout the ticket validity period. The considerably lower percentage of the transfer journey OD pairs which were repeated is probably caused by the added complexity of transfer journey records. Furthermore it could be argued that many transfer journeys are not part of passengers’ daily routine and are ‘once-off’ journeys that were simply not repeated. This could include journeys such as non-work trips, for example, outpatient hospital visits or visiting friends or family.

### Table 7.4: Transfer Journey OD Results Grouped by Classes

<table>
<thead>
<tr>
<th>Case</th>
<th>Day</th>
<th>Route A/D</th>
<th>Route B/C</th>
<th>Direction A/D</th>
<th>Direction B/C</th>
<th>Repeated Order</th>
<th>OD Pairs</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>Same</td>
<td>Same</td>
<td>Same</td>
<td>Same</td>
<td>Same</td>
<td>Consecutive</td>
<td>Same Day</td>
<td>813</td>
</tr>
<tr>
<td>T2</td>
<td>Same</td>
<td>Same</td>
<td>Same</td>
<td>Different</td>
<td>Different</td>
<td>Non-consecutive</td>
<td>Same Day</td>
<td>65</td>
</tr>
<tr>
<td>T3</td>
<td>Same</td>
<td>Same</td>
<td>Different</td>
<td>Different</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Consecutive</td>
<td>5,024</td>
</tr>
<tr>
<td>T4</td>
<td>Same</td>
<td>Same</td>
<td>Different</td>
<td>Different</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Non-consecutive</td>
<td>343</td>
</tr>
<tr>
<td>T5</td>
<td>Same</td>
<td>Same</td>
<td>Different</td>
<td>Different</td>
<td>Different</td>
<td>Not repeated</td>
<td>Consecutive</td>
<td>4,680</td>
</tr>
<tr>
<td>T6</td>
<td>Same</td>
<td>Same</td>
<td>Different</td>
<td>Different</td>
<td>Different</td>
<td>Not repeated</td>
<td>Non-consecutive</td>
<td>308</td>
</tr>
<tr>
<td>T7</td>
<td>Same</td>
<td>Sub</td>
<td>Same</td>
<td>Different</td>
<td>Different</td>
<td>Same Day</td>
<td>Consecutive</td>
<td>558</td>
</tr>
<tr>
<td>T8</td>
<td>Same</td>
<td>Sub</td>
<td>Same</td>
<td>Different</td>
<td>Different</td>
<td>Same Day</td>
<td>Non-consecutive</td>
<td>71</td>
</tr>
<tr>
<td>T9</td>
<td>Same</td>
<td>Sub</td>
<td>Same</td>
<td>Same</td>
<td>Same</td>
<td>Difference</td>
<td>Period of ticket</td>
<td>2,864</td>
</tr>
<tr>
<td>T10</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Same</td>
<td>Same</td>
<td>Difference</td>
<td>Period of ticket</td>
<td>269</td>
</tr>
<tr>
<td>T11</td>
<td>Same</td>
<td>Sub</td>
<td>Same</td>
<td>Different</td>
<td>Different</td>
<td>Not repeated</td>
<td>Consecutive</td>
<td>5,355</td>
</tr>
<tr>
<td>T12</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Same</td>
<td>Same</td>
<td>Not repeated</td>
<td>Non-consecutive</td>
<td>657</td>
</tr>
<tr>
<td>T13</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Different</td>
<td>Same</td>
<td>Consecutive</td>
<td>Same Day</td>
<td>664</td>
</tr>
<tr>
<td>T14</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Different</td>
<td>Same</td>
<td>Non-consecutive</td>
<td>Same Day</td>
<td>156</td>
</tr>
<tr>
<td>T15</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Different</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Consecutive</td>
<td>4,845</td>
</tr>
<tr>
<td>T16</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Different</td>
<td>Non-consecutive</td>
<td>783</td>
<td>1.7%</td>
</tr>
<tr>
<td>T17</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Not repeated</td>
<td>Consecutive</td>
<td>4,379</td>
<td>9.5%</td>
<td></td>
</tr>
<tr>
<td>T18</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Not repeated</td>
<td>Non-consecutive</td>
<td>782</td>
<td>1.7%</td>
<td></td>
</tr>
<tr>
<td>T19</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Same</td>
<td>Different</td>
<td>Consecutive</td>
<td>Same Day</td>
<td>626</td>
</tr>
<tr>
<td>T20</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Different</td>
<td>Same</td>
<td>Not repeated</td>
<td>Non-consecutive</td>
<td>259</td>
</tr>
<tr>
<td>T21</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Different</td>
<td>Consecutive</td>
<td>3,145</td>
<td>6.8%</td>
</tr>
<tr>
<td>T22</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Period of ticket</td>
<td>Not repeated</td>
<td>Consecutive</td>
<td>920</td>
<td>2.0%</td>
</tr>
<tr>
<td>T23</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Not repeated</td>
<td>Consecutive</td>
<td>6,416</td>
<td>13.9%</td>
<td></td>
</tr>
<tr>
<td>T24</td>
<td>Same</td>
<td>Sub</td>
<td>Different</td>
<td>Not repeated</td>
<td>Consecutive</td>
<td>2,035</td>
<td>4.4%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>46,017</td>
</tr>
</tbody>
</table>

As demonstrated above, the various OD classes represent different levels of representativeness of each OD pair. Table 7.5 shows the strong and weak OD classes for transfer journeys. The table provides a summary of identified ODs which were categorised in six different groups. The first three rows of Table 7.5 represent the category of strong OD pairs whereas the remaining three rows correspond to categories of weaker OD classes. The route column identifies...
whether the identified OD pair consisted of boarding records with the same route, 1 substitu­tional route or entirely of substitutional routes. The strongest three categories represent 40% of all identified OD pairs. The weakest categories are with 8% considerably lower. However, within the transfer journey OD pairs only 47% were repeated. This shows that a large propor­tion of inferred transfer journey OD pairs belong to the classes that are neither strong nor weak but are more of average representativeness.

<table>
<thead>
<tr>
<th>Route</th>
<th>Repeated</th>
<th>Order</th>
<th>Identified ODs</th>
<th>Percentage of all ODs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same</td>
<td>Yes</td>
<td>Consecutive</td>
<td>5,837</td>
<td>13%</td>
</tr>
<tr>
<td>1 Sub</td>
<td>Yes</td>
<td>Consecutive</td>
<td>8,931</td>
<td>19%</td>
</tr>
<tr>
<td>2 Sub</td>
<td>Yes</td>
<td>Consecutive</td>
<td>3,771</td>
<td>8%</td>
</tr>
<tr>
<td>Same</td>
<td>No</td>
<td>Non-consecutive</td>
<td>308</td>
<td>1%</td>
</tr>
<tr>
<td>1 Sub</td>
<td>No</td>
<td>Non-consecutive</td>
<td>1,449</td>
<td>3%</td>
</tr>
<tr>
<td>2 Sub</td>
<td>No</td>
<td>Non-consecutive</td>
<td>2,035</td>
<td>4%</td>
</tr>
</tbody>
</table>

7.3. COMPARISON OF THE RESULTS TO OTHER ALGORITHMS

This section will compare the results of the proposed algorithm to other similar approaches. Three other algorithms will be compared: Barry et al. (2002), Zhao et al. (2007), and Trepanier et al. (2007a). The first two focus on the estimation of destinations for passengers from a Metro network. The last approach focuses on a bus network.

Barry et al. (2002) achieved a total of 83% which is not further broken down. However, the method of chaining all records together has been applied which was proven throughout this thesis not be possible for the Dublin Bus data. This is only validated at aggregate level when OD results are compared to total cordon counts. Nevertheless, this is the highest OD estimation ratio of all projects. Zhao et al. (2007) identified a total of 50.01% without the assumption that the last journey in the evening has the destination of the first journey in the morning. Including this assumption adds another 21.1% to the result. However, the authors also link all rail trips as trip chains which are then used to determine the destination. Strong assumptions are used to justify this method.
The only other method, other than the proposed algorithm, that is based on bus EFC data has been introduced by Trepanier et al. (2007a). Although the characteristics of the network are considerably different to Dublin Bus' network it offers the best comparison. The proposed method achieved a 46% identification method which is the same as the proposed algorithm estimated. The study is based on smart card data and therefore does not need to consider ticket type when interpreting results. Trepanier et al. (2007a) also applied the assumption that the destination of the last journey of the day is the boarding location of the first journey of the day. This resulted to an additional 20% which is similar to the result obtained by Zhao et al. (2007) (21.1%).

One reoccurring question arises: Is the assumption of the last journey - first journey acceptable. Barry et al. (2002) states that this assumption is valid for 90% of all subway users. This study showed in Section 6.2.2 that the assumption can not be applied in a universal manner. There would be the possibility to make further assumptions when geographic coordinates are included in the dataset.

One clear advantage of the proposed method is the selection of assumptions. A destination is only assigned when a return journey is found. Other approaches assigned a destination based on the following journey regardless of its parameters. The proposed method is therefore very good when simple trip chains are being recorded.

In summary it can be said that the proposed algorithm estimated a similar amount of destinations as the method developed by Trepanier et al. (2007a). There are still two key questions that could not be answered. The first one is how and when it is justifiable to apply the last journey assumption. The second question is with regard to the uncontrolled usage of trip chains to infer destinations. This certainly seems to work for rail networks but as stated above, causes large errors for urban bus networks. One further step to extent this research area would be to apply a combination of algorithms. This could be done by applying other algorithms after the proposed algorithm identified all possible destinations with the aim to address the boarding records that do not have an inferred destination. Lack of necessary data (such as geographic coordinates) and information about algorithms did not allow for this option to be implemented.

7.4 Implementation of the OD Algorithm

For this algorithm to be implemented the following steps need to be addressed:

- EFC data need to be in a relational database having a similar structure as outlined in Chapter 4;
- Transfer journeys need to be identified;
Substitutional routes need to be identified (either manual or using an algorithm such as outlined in Section 6.4);

- The rule base needs to be checked;
- Both iterations of the OD algorithm need to be executed;
- The results should be checked using some form of simple random sample.

Any regular PC with appropriate storage space and memory can process 1 million records in approximately 6 hours. Faster computers or additional nodes can increase the process speed.

7.5 Waiting Time and In-Vehicle Time Analysis

7.5.1 Introduction

Knowing waiting times and total travel times of passengers are important information parameters for effective transport planning. Many current research efforts rely on such information which is often estimated based on survey results. The goal is to minimise the in-vehicle and waiting time of passengers on a network wide level. According to Zhou and Zhong (2005) there are four alternative-specific elements that influence the inner-city route-choice of public transport passengers: travel cost, frequency, in-vehicle time and out-of-vehicle time which is in this case the waiting time. The in-vehicle time is the time the passenger is travelling on the public transport vehicle whereas the out-of-vehicle time is the time the passenger has to wait to board a public transport vehicle. Nielsen (2000) also argues that waiting times are an important parameter for public transport assignment with regard to some passengers that attempt to minimise the number of transfers whereas others want to minimise travel time. One further problem is that waiting times have often different distributions for different sub-modes or network segments. Generally in-vehicle times and waiting times have to be estimated on a network wide level. In practice in-vehicle times and waiting times depend on the network segment, bus route, or sometimes even the route segment. In addition both of these parameters change throughout the different time periods of a day such as morning peak, evening peak and off-peak.

Having information of the destination at trip level as outlined in previous chapters facilitates now to obtain some of the above mentioned parameters (for each passenger) on a network-wide level. For example, the in-vehicle time can be obtained by subtracting the arrival time of the OD pair from the actual boarding time. A similar concept can be used to extract the in-vehicle times and waiting times of transfer journeys. However, the data that are available for this project only allow to infer the waiting time of a passenger with regard to transfer journeys (out-of-vehicle time between the two individual boardings). The following section will expand on in-vehicle and waiting time extraction.
7.5.2 Extending the Database with Arrival Times of Passenger

The OD extraction algorithm provided various information attributes of the two and four boarding records of single and transfer journeys respectively (see Section 6.9). Knowing the inferred destination of a passenger also allows to calculate the time of alighting of each passenger. Figure 7.1 shows an example of two bus journeys in direction 0 and 1 on route x. The numbers along the arrows are the unique bus stop IDs. It further shows two passenger journeys; Journey 1 and Journey 2. The known information is the boarding location and time of journey 1 at bus stop 2. Further are known the boarding location and time of journey 2 at bus stop 24. The assumption that the boarding of the second journey is the location of alighting of the first journey infers that the passenger alighted at bus stop 7 of journey 1 and bus stop 29 of journey 2. This has already been presented in Chapter 6 and the database stores this new data attribute inferred by the OD extraction algorithm. As no geographic coordinates are known the database used the opposite bus stop. Having the information parameter of the location of alighting and also the unique number of the bus journey it is then possible to extract the exact time when the bus stopped at bus stop 7 and 29 respectively. This is mainly possible because the EFC system records a time based on stages meaning the boarding time of a passenger is attached to the stage record that also stores the stage number. Knowing the time when the bus was at a particular stage facilitates the extraction of the arrival time of each record for which a destination was inferred.

This procedure extended the database with arrival times of each journey where it was possible to extract the opposite stage. As mentioned in Section 5.4 some stages of various routes are unknown and it is therefore impossible to extract the opposite stage and consequently the journey length cannot be calculated. The second problem was that, due to the lack of geographic coordinates, the opposite bus stop was only possible to extract accurately when journey 1 and journey 2 were carried out using the same route. This problem is eliminated when geographic coordinates are available and bound to the dataset as it will then be possible to extract opposite

Figure 7.1: Extracting Arrival Times
bus stops for the entire network regardless which route combination the passenger has chosen.

In summary, the following data attributes have been calculated and added to the existing database: Arrival time of Journey 1; Arrival time of Journey 2; Arrival time of Journey 3 (transfer journeys only); Arrival time of Journey 4 (transfer journeys only).

7.5.3 In-Vehicle Time for Single Journeys on a Network Level
As mentioned above the in-vehicle time is an important parameter for various analyses and is also part of performance indicators according to the Transit Capacity and Quality of Service Manual (TCQSM) (TCQSM, 1999). It can be calculated on a route segment, route and system scope and is of interest to the public, decision-makers, public transport managers and transport planners (TCQSM, 1999). The TCQSM refers to the travel time as a measure that is generally an average duration of a passenger journey from origin to destination. It can be used as a performance indicator using the average travel time or its standard deviation (variability). Reducing the standard deviation may result in an increase in service satisfaction from a passenger point of view but not necessarily results in a reduction of the travel time itself.

In-vehicle times for single journeys can simply be derived from subtracting the original boarding time from the inferred arrival time. Therefore

$$IVT_{S(OD_x)} = AT_{(OD_x)} - BT_{(OD_x)}$$  \hspace{1cm} (7.5.1)

where $IVT_{S(OD_x)}$ is the in-vehicle time of a single journey $S$ derived from a particular OD pair $x$, $BT_{(OD_x)}$ and $AT_{(OD_x)}$ are the passenger boarding/alighting times at the origin and destination respectively. For calculating the average the following formula applies:

$$AIVT_{S(OD_x)} = \frac{\sum(AT_{(OD_x)} - BT_{(OD_x)})}{n} for \ all \ OD_x$$  \hspace{1cm} (7.5.2)

where $AIVT_{S(OD_x)}$ is the average in-vehicle time of all OD$_x$ that are in the route segment, route or network depending on the relevant scope.

Figure 7.2 shows the in-vehicle time of all passengers on a network level in the form of a histogram. The mean of the in-vehicle time is 29.5 minutes with a standard error of the mean of 0.034, which is, due to the large sample size, very low by default. The standard deviation is 17.75 minutes which in this case outlines the variability of in-vehicle time. As all routes are incorporated this measure becomes meaningless as some routes are longer than others. The standard deviation becomes more significant when same type of journeys will be analysed as it can then be used as the measure of variability of in-vehicle times. The histogram consists of 272,383 single passenger in-vehicle times. As mentioned previously, it was not possible to calculate the arrival time of all OD pairs as sometimes the opposite bus stop could
not be identified. The 25%, 50% and 75% percentiles of all journeys are 17, 26 and 39 minutes respectively. This means that only 25% of all journeys take between 39 and 120 minutes. A similar skewed pattern of trip length distribution could be found in London (TFL, 2000). As we are measuring times rather than lengths, travel times at peak times will be at a lower speed and hence a greater in-vehicle time can be expected.

![Figure 7.2: Histogram of In-vehicle Time of Single Journeys on a Network Level](image)

On such an aggregate level this performance measure could be used to compare average in-vehicle time over a certain period of time such as October 2004 with October 2005. It would also be useful to compare the performance indicator after significant changes were implemented throughout the network (e.g. Quality Bus Corridors) to evaluate the impact of the adjustments on average in-vehicle time.

Figure 7.3(a) shows the distribution of boarding times of all single journey boarding times of which the in-vehicle time can be extracted whereas Figure 7.3(b) shows the distribution of all recorded single journeys from October 1999. It can be observed that both graphs are almost identical hence proving that the characteristics of the identified OD pairs with regard to boarding times correspond with the overall characteristics of the entire dataset. There were 1,498,685 single journeys recorded in October. The only slight differences are the slopes during the evening peak times which are faintly steeper in Figure 7.3(b). The coarseness of the graphs differs due to the separate levels of sample sizes.
6.4.3 Boarding Times with Known In-vehicle Time

Figure 7.3: Boarding Time Distributions

Figure 7.4 shows the histograms of in-vehicle times for a certain period of the day. These include Morning peak time (7:30 – 8:30), Evening peak time (16:30 – 17:30), Morning off-peak time (10:00 – 11:00), and Afternoon off-peak time (14:00 – 15:00). Each histogram has an accompanying table containing basic statistics such as the number of instances, mean, standard deviation and percentiles (25, 50 and 75).

The morning peak and evening period are almost identical apart from the slightly higher mean value for the evening peak period and a higher number of instances in the morning period. It could be argued that the lower number of occurrences in the evening peak is due to the wider peaking phase of evening journeys as shown in Figure 7.3. The differences in the mean and standard deviation are negligible. The arithmetic mean is for both periods about 34 minutes. 25% of all journeys do not take longer than 20 minutes and 50% of all single journeys take 31 minutes. However 25% of all passenger journeys have an in-vehicle time of over 45 minutes. The morning off-peak period has the lowest mean of in-vehicle time with a value just below 24 minutes per single journey. 25% of all journeys take 14 minutes or less and only 25% take longer than 32 minutes. It also has the lowest number of instances. The afternoon off-peak period which is from 14:00 to 15:00 also shows a considerably lower number than the analysis of peak periods. The mean is 26.5 minutes and the percentiles are 15, 24 and 35 minutes for the 25, 50 and 75 percentile respectively. This means that 25% of all journeys take less than 15 minutes and 25% take longer than 35 minutes.
7. OD RESULTS AND THEIR APPLICATION TO DETERMINE IN-VEHICLE TIME

7.5.4 In-vehicle Time Variability on a Network Level

The in-vehicle time variability is an important measure when assessing customer satisfaction. In this example it states the positive difference in minutes between in-vehicle times of the single boardings of a return journey. It therefore compares the in-vehicle time a passenger experienced in one direction with the in-vehicle time of the journey in the opposite direction. Only those journeys were considered that had a return journey on the same day. Figure 7.5 shows the in-vehicle time variability of single journeys in the form of a histogram. An analysis similar to this can only be done when times of both passenger journeys are known which is difficult to achieve using surveys. The OD estimation algorithm therefore produces a new data set with new possible analyses.

Basic statistics are displayed beside the graph. The OD dataset consisted of 53,751 records for which both in-vehicle times were known. The mean of the positive in-vehicle differences is roughly 9 minutes with a standard deviation of 10.29. The 25th, 50th and 75th quartiles are
7.5. WAITING TIME AND IN-VEHICLE TIME ANALYSIS

2, 6 and 12 minutes respectively. This means that 25% of all passengers’ journeys had an in-vehicle time variability of two minutes or less. 50% had a variability of 6 minutes. 25% of all passenger journeys had an in-vehicle variability of 12 minutes or more. The higher values of the chart could be errors in the dataset as the in-vehicle times were calculated without knowing the exact geographic coordinates. All networks where the geographic coordinates of each bus stop are known will eliminate this error. However, the error even without geographic coordinates seems to be negligible.

7.5.5 Travel Time for Transfer Journeys

Before analysing the in-vehicle time of transfer journeys the peak and off-peak times have to be identified. The line graph displayed in Figure 7.6 shows the boarding times of passengers at the journey’s origin. Comparing this graph to Figure 7.3(b) it can be seen that the peak periods are not exactly the same. It could be assumed that transfer journeys take longer than single journeys and therefore transfer passengers have to commence their journey at an earlier time than single journey passengers. The second argument is that because of the extended journey length the passengers choose a bus service at an earlier time to escape the general peak time. Therefore the peak times for this analysis are chosen as follows:

- Morning peak time (7.00 – 8:00) - Figure 7.7a;
- Evening peak time (16:30 – 17:30) - Figure 7.7b;
- Morning off-peak time (10:00 – 11:00) - Figure 7.7c;
- Afternoon off-peak time (14:00 – 15:00) - Figure 7.7d.
As expected, the travel time of transfer journeys is considerably longer than for single journeys. The travel time that is calculated for this analysis consists of the sum of the length of journey 1, waiting time at transfer node and length of journey 2. This performance measure can only be calculated for OD pairs of which both arrival times were known.

Figure 7.7a shows a histogram of the total journey length on a network wide level during the morning peak period. All required arrival times were available for 7,656 transfer journeys. The arithmetic mean was just above 63 minutes with a standard deviation of 22.4 minutes. Again, because the analysis focuses on the entire network the standard deviation cannot be used as performance indicator. The analysis further shows that 25% of all transfer journeys take 47 minutes or less and 25% take 79 minutes or more. The median lies at 62 minutes. The main part of 50% of all transfer journeys take between 47 and 79 minutes.

Figure 7.7b shows a histogram of the total journey length on a network wide level for the evening peak period. All required arrival times were available for 4,398 transfer journeys. This number is lower than the equivalent of the morning peak because the evening peak period is longer and more spread out. The arithmetic mean is 68 minutes and therefore approximately 5 minutes longer than during the morning peak hour. The standard deviation is also slightly higher with a value of 23.6 minutes. The analysis further showed that 25% of all transfer journeys take 51 minutes or less and 25% take 85 minutes or more. The median lies at 67 minutes which again is 5 minutes more than during the morning peak period. The main part of 50% of all transfer journeys take between 51 and 85 minutes.

Figure 7.7c shows a histogram of the total journey length on a network wide level during the morning off-peak period. All required arrival times were available for 1,744 transfer journeys. The arithmetic mean is almost 59 minutes. The standard deviation is 22.7 minutes. The analysis
further showed that 25% of all transfer journeys take 43 minutes or less and 25% take 74 minutes or more. The main part of 50% of all transfer journeys take between 43 and 74 minutes.

![Histogram of Transfer Journey Lengths](image)

Figure 7.7: Journey Length of Transfer Journeys

Figure 7.7d shows a histogram of the total journey length on a network wide level during the afternoon off-peak period. All required arrival times were available for 2,324 transfer journeys. The arithmetic mean is approximately 65.5 minutes and therefore considerably higher than the calculated mean during the morning off-peak period. The analysis further showed that 25% of all transfer journeys take 48 minutes or less and 25% take 83 minutes or more. The main part of 50% of all transfer journeys take between 48 and 83 minutes. One main reason for the length of journeys during the evening off-peak period is the extended waiting time. Services do not run as frequent which increases the waiting time and therefore also the total journey time.

Waiting time is a further important performance measure and is in practice relatively difficult to determine (at least on a large scale). Surveys are often biased due to misinterpretation of the actual waiting time by the participants. Knowing the arrival times of all transfer journey
boardings it is possible to calculate the waiting time as well as its variability. We define

\[ WT_{T(OD_x)} = BT_{T(OD_{x3})} - AT_{T(OD_{x2})} \]  

(7.5.3)

where \( WT_{T(OD_x)} \) is the waiting time of a particular OD pair \( x \) of transfer journey \( T \). \( BT \) and \( AT \) are the boarding and alighting times respectively where \( x3 \) is the third boarding record and \( x2 \) is the second boarding record of the four record transfer OD pair.

Figure 7.8 shows two histograms that visualise these performance measures on a network level. The waiting time has a mean of almost 17 minutes based on a sample of 30,000 transfer journey boardings. The waiting time variability was only calculated for those transfer journeys of which both waiting times (morning and evening) were known. The difference of these was termed waiting time variability. The required information was available for 6,300 transfer journey pairs. The mean of this variability is almost 19 minutes with a standard deviation of almost 20 minutes. These graphs highlight some margin of error as a waiting time by our predefined constraint cannot be longer than 90 minutes. This error is caused by either wrongly assigned destinations or by incorrect time stamps of the EFC data.

![Figure 7.8: Transfer Journeys Waiting Time and its Variability](image)

This analysis was carried out on a network wide level. Depending on the purpose of a travel time analysis it might be necessary to focus on route level. In that case the performance measure of the travel time variability (standard deviation of travel time) can also be used to analyse changes on a specific route over time. As the EFC data area available on a continuous basis it is possible to calculate the above mentioned performance measure more frequently allowing the operator to make appropriate changes to optimise its services. Furthermore it would be possible for regulatory bodies to monitor the performance of operators.
7.6 Summary

This chapter focused on the results of the algorithm with regard to the estimation ratio and the OD pair assignment to each class. This was followed by a demonstration as to how in-vehicle time of single journeys, variability of in-vehicle time of single journeys, total transfer journey travel time and transfer journey waiting time can be calculated by extracting the estimated arrival times of passenger journeys. This approach showed how decision support for transport planners and decision makers can be improved using the proposed method. The following main conclusions can be drawn:

- The OD results were grouped by ticket type and then analysed as to how many destinations could be inferred by the algorithm. An average of 46% of all boardings (57% single journeys and 24% transfer journeys) were assigned a destination. Some ticket types however reached a OD inference ration of as high as 78% (Foreign student tickets). Rechargeable smart cards will reduce the number of tickets that consist of small boarding numbers. For some tickets the algorithm assigned all boarding records a destination;

- Out of all single journey OD pairs 92% were recorded in consecutive timely sequential order. Furthermore, 59% of all identified single journey OD pairs were repeated either on the same day or throughout the period of ticket validity. Out of all transfer journey OD pairs 86% were recorded in consecutive timely sequential order. Furthermore 47% of all transfer journey OD pairs were repeated either on the same day or throughout the period of ticket;

- The various classes of OD pairs was analysed with regard to strong and weak OD pairs. Single journey OD pairs were classed as 54% strong and 3% weak. For transfer journeys 40% of identified OD pairs were assigned to strong classes whereas 8% were assigned to weak classes;

- The results were compared to other algorithms. It seems that rail/metro network EFC data suit the trip level OD extraction more than for bus EFC data. Equivalent results were obtained between the proposed method and the method by Trepanier et al. (2007a). The universal validation of individual assumptions is still lacking.

- Knowing the destination of passengers it is further possible to extract the estimated time of alighting. This further facilitates an analysis with regard to in-vehicle time, in-vehicle time variability total travel time, waiting time at transfer nodes and waiting time variability. The analysis showed that the network wide average in-vehicle time is almost 30 minutes ($n = 272,383$). The average in-vehicle times of single journeys in the morning peak period and evening peak period were 33.9 and 34.4 minutes respectively. The in-vehicle time during the off peak periods were with 23.9 minutes per journey considerably shorter (morning off-peak). The variability of in-vehicle time when comparing the two records
of each OD pair was 9.1 minutes. The in-vehicle time analysis of transfer journeys had different results. The mean of total journey time (both legs and waiting time at transfer node) was 63 and 68 minutes for morning peak and evening peak periods respectively. A similar difference was found when analysing the morning and afternoon off-peak periods.

The average waiting time at transfer nodes is 17 minutes with an average variability of almost 20 minutes;

The following chapter will summarise the conclusions and main findings of this thesis. It will also outline some future research suggestions.
Chapter 8

Discussion & Conclusions

This chapter summarises the research that was carried out throughout this project, discusses the main results and addresses the objectives that were stated in the introduction. The main findings and their implications are then discussed. Various suggestions for future research precede the final remarks of this thesis.

In general, this thesis proposed a new method to estimate destinations of public transport passengers using historical EFC data of an urban bus operator. Existing literature and methods for rail networks identified the need for methods to be able to infer destinations of bus passengers. The proposed algorithm offers a solution to infer the alightings of passengers based on a set of rules. These rules were then applied to the EFC data set with the focus to determine the stage of alighting. A series of steps needed to be developed prior the execution of the main OD algorithm. These included the correct data migration of the semi-structured, semi-encrypted EFC data, the identification of single and transfer journeys, the identification of substitutional routes and the definition of the rule base which was then used to infer the destinations. It was attempted to include as many methods to validate these methods as possible. A travel diary that recorded information that was similar to the recordings of the EFC system was carried out. The investigation in terms of correctness of the OD extraction algorithm was promising. The study showed a high precision rate with a moderate value for the recall which is similar to the actual results. The applications of the results to the public transport sector are multiple. For demonstration purposes it was shown that knowing the destinations it was further possible to calculate the time of alighting of each passenger journey. This was then used to calculate performance measures such as in-vehicle time, in-vehicle time variability, total transfer journey time, waiting time and waiting time variability.

The following section will focus on the main contributions of this thesis.
8.1 Summary of Research and Main Contributions

The main aim of this thesis was to develop a method that estimates passenger destinations at trip level using historical EFC data that were recorded on an entry only validation urban bus network. The following section will summarise the steps that were developed to achieve this aim and will further outline the specific contributions to the body of knowledge.

8.1.1 Data Migration Framework

A total of 45 million boarding records were available for the project which were migrated into a relational database using a newly developed 4-Phase data migration framework. Such a framework is important to ensure that experience is documented and projects are reliable and can be replicated. It should assist researchers and practitioners during the import process of the data by providing a project management aid. The domain specific model is, although being presented with Oracle tool sets, actually non-proprietary. A more detailed description of this aspect of the research project can be found in Hofmann et al. (2003).

8.1.2 Transfer Journey Classification

A newly developed algorithm classifies journeys from historical EFC data into two different categories: linked trips and unlinked trips. This was achieved by applying a set of assumptions to the boarding records. The algorithm was tested over a four month period. A new assumption was added to differentiate the model from previous research such as Bagchi and White (2003) and Okamura et al. (2003). This assumption stated that a transfer journey classification can only be made if the second boarding is being made on a different route. The argument behind this is that if the same route is taken twice then it was probably a return journey and not necessarily a transfer journey. This is of particular importance when the cut-off point is larger meaning that passengers potentially could make a journey without a single main purpose. This additional assumption improved the transfer journey classification algorithm considerably by increasing its precision value. A simple random sample and a survey was used to show the correctness of this algorithm. Furthermore, a Monte Carlo simulation assessed the robustness of the algorithm when random errors were introduced. The simulation showed that a random added error mostly has linear impact on the error of classifying transfer journeys. Similar project that aimed to determine passenger destinations based on historical EFC data mostly neglected the area of transfer journeys. Partially because the method of chaining most journeys together made the knowledge of linked and unlinked trips unimportant. However, we believe that this is not entirely true. Transfer journeys should be treated differently regardless which approach is used simply due
to their difference of characteristics compared to single journeys. It is therefore important for this and other studies to be able to identify all true transfer journeys that were made on the network. The proposed algorithm has a small and acceptable error rate which demonstrates its clear ability to classify linked and unlinked trips.

8.1.3 Analysis of Data Quality with Regard to Bus Driver Behaviour

Many EFC systems still rely on the bus driver to select the correct bus stage which is then part of the stored boarding record. Little research has been found with regard to the correctness and impact of this issue. Three different methods were used to show that (1) the bus driver generally recorded the correct stage and (2) the quality of the data set is good. The first approach analysed the frequency of recorded bus stops. The second approach analysed boarding distributions on the premise that more people board at busy bus stops. The last approach focused on the arrival time of the vehicle at the bus stop on the premise that the arrival times on bus stops should not be the same for a number of bus stops.

8.1.4 Transfer Journey Analysis

A comprehensive analysis mainly with regard to travel behaviour of transfer passengers was carried out. This contributed considerably to the general understanding of the characteristics of transfer journeys in the Dublin Bus transportation network. Varying travel times, ticket utilisation, and ticket category analysis were investigated.

8.1.5 Network Symmetry

A new measure to calculate route/network symmetry was proposed. The measure of 'Degree of Symmetry' could potentially be used to justify the use of symmetry assumptions. The method was applied to single and transfer journeys with the aim to test the network for the degree of symmetry. The model can be used to compare routes or networks with regard to degree of symmetry. Potentially this could also be used to estimate OD matrices. However, this would only be possible at an aggregate level and was therefore not further pursued as part of this thesis.

8.1.6 Identification of Substitutional Routes

A novel method to identify substitutional routes (common routes) which are alternative routes a passenger can use to get to his/her final destination was proposed. An innovative algorithm was developed that works on the premise that passengers use the combination of substitutional routes more frequently than non-substitutional routes. The results of the algorithm were tested using
a simple random sample and a manual identification of substitutional routes. This method of identifying alternative routes was chosen because it clearly represents what passengers consider to be substitutional routes as the results are based on actual boarding records. Simple random samples validated the method and an acceptable error of 2% was detected.

8.1.7 OD Extraction of Journey Pairs

Most of the above stages were necessary to facilitate the design and development of the OD extraction algorithm that estimates the destination of urban public transport passengers. The main assumption stated that, in simplified terms, the location of the boarding in the evening can be considered the final destination in the morning and vice versa. This statement is further enhanced by considering a multiple array of attributes when inferring a passenger destination at trip level.

The main algorithm used a rule based approach called Rule Based Reasoning (RBR) that uses a reasoning process to connect data to conclusions. It is the formal implementation of the thinking process when the aim is to extract patterns similar to the ones that are emerging from passenger boarding records. Production rules or simply rules are the most common method to represent knowledge. RBR allows domain knowledge to be transferred into rules which are then applied to the dataset. These rules were identified using decision tables that consisted of all theoretical combinations of the values of the recorded attributes. Illogical combinations were then eliminated. The final set consisted of 12 rules for single journeys and 24 rules for transfer journeys. Section 6.8 lists the detailed innovations of this model. The proposed method follows a pessimistic approach. This means that the approach aimed to get a high precision value which lead to a lower total destination identification value (recall).

8.1.8 OD Extraction of Entire Passenger Boarding Records

As the first iteration only analysed boarding pairs a second iteration of the algorithm was proposed. Throughout this second iteration the algorithm and its rule base was extended to facilitate the analysis of all recorded boardings of any given passenger in a parallel manner with the aim to extract common travel behaviour that was stored over the entire ticket validity period. This method further improved the extraction ratio of passenger destinations by adding four more rules. This novel approach means that the more boarding records are recorded by a passenger the higher is the identification rate of destinations. This is of particular importance as many networks now operate with re-chargeable cards that assign boarding records to a passenger for
many months or even years. However, in practice full records are often only stored for several weeks. This may change over time as storage cost of data constantly decreases and applications of EFC data increase.

8.1.9 Validation of the OD Results
The dataset has no exit validation information for any of the records and it was therefore difficult to validate the estimated OD extraction results. The validation was therefore split into several different approaches each using a slightly different view point. The main validation method consisted of data obtained from a small travel diary study that was solely implemented to collect data for this validation process. The second method was to validate the results by comparing various measures and matrices in aggregate form to the results that were based on a large DTO employment survey. Two approaches analysed the actual results to see whether the aggregate statistics are in line with common sense and domain expertise. All validation methods indicated that the results are acceptable and that the error rates are low. Furthermore, one section focused on the limitations and boundaries of both algorithm and data.

8.1.10 Calculation of In-vehicle Time, Total Transfer Journey Time and Waiting Time at Transfer Nodes
One of many applications that result from trip level OD data is the calculation of performance measures at different levels (e.g. route, route segment or network level). This thesis showed this by using the inferred destinations to further extract the arrival time of passengers. Knowing times of boarding and alighting at trip level then facilitates the calculation of in-vehicle time, in-vehicle time variability, total transfer journey time and waiting time at transfer nodes. The outcome of this novel approach can be used as performance or monitoring measures by urban bus operators or regulators. It further shows the potential possibilities and its applications of the extended data base.

8.1.11 Summary
In conclusion to this section it can be stated that the set aims and objectives were mainly achieved. Although the ratio of inferred destinations is low, the precision and accuracy values are high which was one of the main goals. Furthermore, the results are similar to the most recent study by Trepanier et al. (2007a) which was also the first publication that presented a method to infer destinations of bus passengers at trip level based on historical EFC data. However, we believe that Trepanier et al. (2007a) used a more optimistic approach when inferring
destinations. Due to the nature of their algorithm transfer journeys and substitutional routes were not considered.

The validation of the results were promising in particular the travel diary study which was used to simulated real EFC data and tested the OD extraction algorithm. Interesting should be the implementation of the algorithm on newer data as the use of magnetic strip cards increased considerably since 1999 and should therefore also provide a better result.

8.2 Can EFC Data Improve Information Need of Operators?

Section 2.9 reported about a study that focused on the information needs of PTEs, county councils and bus operators (Bagchi, 2003a). Nine statements were presented and the participants had to rank each of them. After having extended and analysed the EFC data set, the following paragraphs will focus on each of the statements in order to evaluate whether EFC data could be used to provide the various entities with the required information.

**Statement 1:** *Having a continuous record of journeys undertaken by individuals using our service and being able to link the records to those individuals.* This is possible using EFC data within reason. In the case where a fare card medium is assigned to a name it is possible to follow the travel path of the passenger. This is increasingly the case for smart cards or annual tickets. When using EFC data generated by magnetic strip cards then the length of continuous records of individual passengers depends on the validity period of the ticket in use.

**Statement 2:** *Identifying more of our customers by name and address not just the travel card holders.* This is not possible as passengers that do not hold a travel card are not traceable within the system.

**Statement 3:** *Recording the origin and destination of all journeys (boarding and alighting points to/from vehicle) undertaken on our buses.* As this thesis and other projects showed, it is possible to extract OD pairs of journeys. The inferred OD pairs provide a dataset that could be used to determine many performance measures or trends as this thesis has shown by introducing several examples.

**Statement 4:** *Having the same journey information available for a larger and more representative proportion of our customers than is possible to obtain through existing surveys.* Again, this is possible. What has not yet been analysed is the difference between a sample consisting of magnetic strip card boardings and the population of all public transport users of a specific network.
Statement 5: *Having a continuous record of journeys undertaken by identified customers using our services so that we can see how people vary their travel behaviour day to day, week to week, etc.* This could be realised by providing a certain amount of passengers with a travel card, linking each unique ticket number to the personal details of a passenger. The data recorded of the participating customers could then be extracted and analysed.

Statement 6: *Being able to identify the pattern of interchange with other buses as part of a customer’s journey from one activity to another (e.g. home to work).* This is only possible when period or multi-journey tickets are used. Chapter 5 focused on the analysis of route interchange. This is probably one of the most advanced analysis areas within this thesis. However, the activity can only be estimated on the basis of the recorded boarding records such as location and destination as well as time and date of the boardings.

Statement 7: *Monitor and measure demand at different time periods and at different places (e.g. interchanges).* This is possible without any great difficulty as long as the data is stored in an accessible format such as a database. Chapter 5 and 7 discussed this matter in some detail.

Statement 8: *Measure more accurately than presently possible how many new customers have started using our services over a period of time.* This is not possible in its entirety. In the case where smart cards are used, new subscriptions could be analysed. However, new customers that pay by cash can not directly be identified.

Statement 9: *Measure more accurately than presently possible how many customers have stopped using our services over a period of time.* This is similar to statement 8. Smart cards that are not used anymore could be analysed. However, this would not result in a definite number of passengers that stopped using the service.

In summary it could be concluded that EFC data can provide information that is explicitly required by County Councils, PTEs and bus operators. In particular, after being able to estimate destinations for a considerable portion of the boarding records.

### 8.3 Main Findings and Future Implications

This section focuses on the main findings of the research presented in this thesis. More detailed findings can also be found at the end of each chapter.

The transfer journey algorithm identified approximately 36% of all unlinked boarding records as linked trips. This additional data attribute is an essential part of information for various operational and passenger behaviour analyses. It was further identified that the 90 minute cut-off
point was the correct choice although this could be reduced to 75 minutes resulting in the loss of only 10% of originally declared transfer journeys.

The data quality section of the paper was considered important because the quality of the results generated can only be as good as the quality of the data themselves. Although the analysis concluded that the recorded data is accurate (within a margin of error) it was also discovered that the dataset is not as complete as it could have been. This is mainly with regard to stages where no passengers boarded which occurs particularly towards the final stages of each route. In summary, although there is frequent human interaction, the system where the bus driver has to indicate the location of the stop does work and the recorded data seem to be accurate. With the growing integration of more complex technology, most systems will not have to face this problem for much longer.

The lack of geographic coordinates caused limitations in many aspects throughout the research and development of the proposed methods. Having locations of each bus stage/stop would therefore certainly be of advantage in particular with regard to the validation of OD pairs, substitutional routes and transfer journeys.

The transfer journey analysis produced summary figures that clearly helped to more fully understand passenger’s transfer behaviour. In particular it highlighted the differences between single journeys and transfer journeys in terms of boarding time, ticket utilisation and usage. The analysis, for example, showed that almost 10% of transfer passengers transfer within 10 minutes after boarding the first bus and 50% of all transfer customers transfer within 34 minutes.

Most transfers originate from the city centre coarse zone 1 (20.80%). Over 60% transfer in the city centre. Smaller clusters have also been discovered which may indicate the need for orbital routes to reduce or eliminate the need of transfer journeys.

The research carried out with regard to the network symmetry identified a gap in the literature. An equation was developed that defined the level of symmetry in a network or network segment.

The ticket type analysis lead to the result that over 53% of all transfer journeys are validated using a Weekly Adult City zone or Weekly Student City zone ticket. The same tickets are only responsible for 32% of single journey validation. The third most used ticket for transfer journeys was the 3 Day - Adult ticket which was responsible for almost 17% of transfer journey validations compared to almost 9% for single journey validations. The four most common tickets to validate transfer journeys were responsible for 76% of all validations.
The time analysis where transfer journeys were compared with single journey boarding times shows that the morning peak of transfer journeys is clearly from 7.00 to 8.00 whereas the single journey morning peak is from 8.00 to 9.00. The evening peak of transfer journeys is from 16.00 to 17.00 while the peak time of single journeys lies between 17.00 and 18.00. This has been established through the initial peak time analysis as well. Overall the run of the two curves is very similar.

The following main conclusions can be drawn with regard to the improvement of the algorithm and the results themselves:

Foreign student tickets show a very high total percentage of OD estimation. The highest percentage with 78% means that out of 10 journeys 7.8 could be inferred with a destination. The algorithm is designed in such a way that the greater the routine behaviour of a passenger the better will be the OD results achieved. It could therefore be assumed that foreign students have on average a higher routine behaviour than the remaining passengers. This could be caused by their smaller social network. On average almost 7.1 out of 10 journeys can be inferred using all foreign student tickets

2-Journey tickets have a relatively low percentage of transfer journeys. The average total percentage of OD estimation is 36% which is considerably lower than the total average of 46%.

The tickets with a percentage of OD estimation of less than 1% are tickets that in theory should not result in any OD pairs. However due to small errors in the EFC data, unique IDs were duplicated and therefore result in some OD pairs. Although this is an error, it does not cause any problems with regard to the OD estimation. The algorithm proved to be working as in this case only 18 errors occurred which is less than 0.00005% and therefore acceptable.

The OD extraction percentage is higher for single journeys than it is for transfer journeys. This is due to the increased complexity of transfer journeys as four boarding records have to be compared and matched. Including all ticket types the average OD extraction percentage of single journeys is 57% compared to 24% of transfer journeys.

It is noteworthy to add that these figures are average figures of several thousands of different passengers. Many passenger tickets have an OD extraction percentage of over 95%. Furthermore not all recorded boardings have a return journey due to car pooling or other modes of transport that offer an alternative to the bus services. For these journeys it is not possible to assign a destination without violating existing or adding more assumptions.

Out of all single journey OD pairs 92% were recorded in consecutive timely sequential
order. Furthermore, 59% of all identified single journey OD pairs were repeated either on the same day or throughout the period of ticket validity. Over 54% of all inferred single journey OD pairs are considered to be representative due to their consecutive order and their repetitive nature.

Out of all transfer journey OD pairs, 86% were recorded in consecutive timely sequential order. Furthermore 47% of all transfer journey OD pairs were repeated either on the same day or throughout the period of ticket.

The strongest three categories represent 40% of all identified OD pairs. The weakest categories are considerably lower with 8%. However, within the transfer journey OD pairs only 47% were repeated. This shows that a large proportion of inferred transfer journey OD pairs belong to the cases that are neither strong nor weak but are more of average representativeness.

Two new assumptions have contributed considerably to the overall ratio of identified OD pairs and increased the total amount of passenger journeys that could be assigned a destination from 873,004 to 1,026,880. This is equivalent to an identification rate of all boardings of almost 46%. Considering only single journeys then the average identification reaches 57%. In general, the longer the ticket validity period is the higher the percentage of improvement. This is mainly caused by the new assumption introduced in Section 6.7.4.

As expected, the main destination is the city centre. As most of the inferred destinations were assigned to single journeys, this makes sense considering the radial nature of the urban bus network. The other pattern that occurs is the increased frequency of journeys where origin and destination were within the same zone. Furthermore two clusters were identified. One spans coordinates of the zones 11 to 14 (origin) and 5 to 7 (destination). The second cluster includes the zones 5 to 7 (origin) and 11 to 14 (destination). The zones of these clusters include busy suburban centres that offer work, shopping and leisure activities and are therefore frequently visited by Dublin Bus passengers.

Knowing the destination of passengers it is further possible to extract the estimated time of alighting. This further facilitates an analysis with regard to in-vehicle time, total travel time, waiting time at transfer nodes and in-vehicle time variability. The analysis showed that the network wide average in-vehicle time is almost 30 minutes \((n = 272,383)\). The average in-vehicle times of single journeys in the morning peak period and evening peak period were 33.9 and 34.4 minutes respectively. The percentiles were approximately the same. The in-vehicle time during the off peak periods were with 23.9 minutes per journey considerably shorter (morning
off-peak). The variability of in-vehicle time when comparing the two records of each OD pair was 9.1 minutes. The in-vehicle time analysis of transfer journeys had different results. The mean of total journey time (both legs and waiting time at transfer node) was 63 and 68 minutes for morning peak and evening peak periods respectively. A similar difference was found when analysing the morning and afternoon off-peak periods.

8.4 Suggestions for Further Research

The main research effort described throughout this thesis only addressed the OD estimation for a single operator based on single mode data. Most cities have multi-modal and multi-operator networks. Furthermore not all required data attributes were available. Consequently the data needs for advanced analyses need to be addressed in much greater detail. This section elaborates briefly on these important issues.

8.4.1 What Data are Really Needed?

As already mentioned throughout this thesis: the analysis results can only be as good as the data that are used to obtain them. Furthermore data attributes have to be available in order to incorporate them directly or indirectly into any analysis. For example, this project could not avail of exact spatial parameters such as geographic coordinates of all bus stages. If EFC data are used in commercial projects to improve operational or strategic planning and decision making a more detailed segregation of areas of the public transport network will be critical.

Furthermore this project’s data source only consisted of data records that were recorded with magnetic strip cards. The data recorded by cash paying public transport passengers were not available. An interesting question that may be addressed in future research projects is how much the cash records differ from magnetic strip card (or smart card) recordings. If the difference is negligible then all results obtained from the results based on magnetic strip card data can be applied to the entire population. Efforts such as the introduction of off-bus ticketing to cover less frequent users (e.g. Oyster PAYG) might result in such data becoming more representative of the bus user market as a whole.

Another issue that needs to be addressed in relation to the data needs are standards that have to be implemented by all public transport operators regardless of transport mode of the entire network. If the aim is to develop more complex algorithms that incorporate several operators and modes then the data attributes have to be standardised so that they can be merged into one big database. This is especially necessary in the following cases:
Other stakeholders such as public transport regulators or departments of transport show interest in analysing the data; 
One operator is in charge of more than one mode.

It is important to realise that future automatic data collection systems which support public transport operations can be designed so that required data are available. For example, AVL systems can be incorporated into EFC systems and therefore generate exact spatial and temporal data of the journey of buses.

8.4.2 Development of a Multi-modal/Multi-operator Algorithm

The OD estimation algorithm which was proposed in this thesis was only based on data from one operator and one mode (urban bus operator). Although Dublin Bus was more or less the main urban bus operator for the GDA and the bus transport mode is the main public transport mode this does not apply to many cities. Multi-modal and multi-operator environments are the norm in most large cities that have an innovative public transport network. The project which was researched by MIT in Boston using Chicago data used urban rail data and urban bus data to estimate OD pairs on a passenger level. However, the proportion of the urban bus data was solely used to infer the previous heavy rail station where the passenger alighted. In reality public transport passengers that live in a city with a multi-modal network will often use two or more modes to travel between destinations. It is therefore important that OD estimation of public transport passengers has to involve all major modes of transport. The following issues may cause problems which have to be solved before OD pairs from a multi-modal and multi-operator network can be estimated:

- Data of the various operators and modes need to be compatible. This means that the format of the recorded data needs to be similar so that they can be imported into a common database which then serves as the common source of data for future analyses;
- The values of the data attributes used by the various automatic data collection systems need to be standardised for all modes and operators. For example it is necessary that bus stops have a unique ID which needs to be used by all operators that serve the stop. Also directional identifiers (inbound-outbound) need to be standardised so that smooth data import with no errors is possible;
- A problem that is of a less technical nature is data sensitivity from an operator point of view. There have to be incentives to individual operators as well as the entire public
transport operators of a network. Public transport regulators are an upcoming trend to ensure that all operators and all modes work together in an efficient and effective manner. A survey of operators’ opinions and expectations might be a solution to form a better picture.

8.5 Final Remarks

There is no doubt that the public transport sector will follow the trend that many other industry and service sectors have successfully pursued for several years. Data recording systems such as EFC systems will in future be designed with the data needs in mind and not solely to serve the public transport operations directly. In an age where public transport in urban areas is one of the corner stones of and contributors to socio-economic developments, the potential for improvements cannot be disregarded. Large data sources certainly have the ability to improve public transport networks by facilitating the transport planner with more information about passenger behaviour and travel patterns as well as operational performance of the public transport fleet.


M. Hofmann and M. O'Mahony. Transfer Journey Identification and Analyses from Electronic Fare Collection Data. IEEE Intelligent Transportation Systems, Vienna, Austria, 2005b.


Appendices
## Appendix A

### Coarse Zone Description

<table>
<thead>
<tr>
<th>Coarse Zone</th>
<th>Coarse Zone Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>City Centre (within Canal Ring)</td>
</tr>
<tr>
<td>2</td>
<td>Dublin Port Area</td>
</tr>
<tr>
<td>3</td>
<td>North East City (Clontarf, Raheny, Ayrfield)</td>
</tr>
<tr>
<td>4</td>
<td>North West City (Cabra, Finglas, Ballymun)</td>
</tr>
<tr>
<td>5</td>
<td>South East City (Rathmines)</td>
</tr>
<tr>
<td>6</td>
<td>South West City (Kilmartin, Walkingtown, Kimmage)</td>
</tr>
<tr>
<td>7</td>
<td>Fingal West (Blanchardstown / Castleknock)</td>
</tr>
<tr>
<td>8</td>
<td>Fingal East (Portmarnock, Malahide, Donabate, Swords, Airport)</td>
</tr>
<tr>
<td>9</td>
<td>Fingal North West (Naul, Ballybohill, Oldtown)</td>
</tr>
<tr>
<td>10</td>
<td>Fingal North East (Rush, Lusk, Skerries, Balbriggan)</td>
</tr>
<tr>
<td>11</td>
<td>South Dublin - Lucan, Clondalkin</td>
</tr>
<tr>
<td>12</td>
<td>South Dublin - Tallaght</td>
</tr>
<tr>
<td>13</td>
<td>South Dublin - Saggart, Rathcoole, Bohernabrena</td>
</tr>
<tr>
<td>14</td>
<td>Dun Laogharie / Rathdown - North</td>
</tr>
<tr>
<td>15</td>
<td>Dun Laoghaire / Rathdown - South</td>
</tr>
<tr>
<td>16</td>
<td>Meath</td>
</tr>
<tr>
<td>17</td>
<td>Kildare</td>
</tr>
<tr>
<td>18</td>
<td>West Wicklow</td>
</tr>
<tr>
<td>19</td>
<td>East Wicklow</td>
</tr>
<tr>
<td>20</td>
<td>Louth</td>
</tr>
<tr>
<td>21</td>
<td>Externals</td>
</tr>
</tbody>
</table>
Appendix B
Coarse Zone Map
Appendix C

Substitutional Routes

The following tables Table 2 to Table 5 show the results of the four different analysis scenarios. It shows the cut-off points of *Number of Boardings* on the main horizontal header and the cut-off values of the calculated weights (using Equations 6.4.1 and 6.4.3) on the vertical header. Within each of these cut-off values there are three further categories:

- No of Subroutes;
- Number of Errors;
- Relative percentage of Errors.

Figure 1, Figure 2, Figure 3 and Figure 4 show the cut-off boardings on the x-axis and the impact on wrongly identified boardings due to errors in estimating substitutional routes on the y-axis (in %).

Table 2 shows the various values for Solution 1 using the OR option. The highlighted cells indicate the cut-off point combinations where the percentages of relative errors were below 2.5%. This is also shown when graphically displaying the data (see Figure 1). This model is robust where the *Number of Occurrences* is between 50 and 60 and calculated weights lie between 1 and 4.5.

Table 3 shows the various values for Solution 1 using the AND option. The highlighted cell indicates the cut-off point combination where the percentage of relative errors was below 2.5%. Only one cut-off combination resulted in a low impact error. This is also shown when graphically displaying the data (see Figure 2). Figure 2 shows a high variation between the different parameters and it is therefore impossible to identify a logical pattern. This model is not robust and does not result in low impact errors.

Table 4 shows the various values for Solution 2 using the OR option. The highlighted cells indicate the cut-off point combinations where the percentages of relative errors were below 2.5%. This is also shown when graphically displaying the data (see Figure 3). This model is robust where the *Number of Occurrences* is between 50 and 100 and calculated weights lie between 1 and 4.5.

Table 5 shows the various values for Solution 2 using the AND option. The highlighted cells indicate the cut-off point combinations where the percentages of relative errors were below 2.5%. This is also shown when graphically displaying the data (see Figure 4). Figure 4 shows a similar pattern as Figure 2. Although the 0.5 to 2.5 cut-off points are relatively similar the
Table 2: Solution 1 - OR

<table>
<thead>
<tr>
<th>Cut-off points for 'Number of occurrence'</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>100</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>2,162</td>
<td>1,467</td>
<td>1,193</td>
<td>1,079</td>
<td>1,015</td>
<td>980</td>
<td>941</td>
<td>931</td>
</tr>
<tr>
<td>Number of Subroutes</td>
<td>75</td>
<td>41</td>
<td>30</td>
<td>26</td>
<td>23</td>
<td>23</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>6.15</td>
<td>4.67</td>
<td>3.72</td>
<td>3.26</td>
<td>2.98</td>
<td>2.51</td>
<td>3.22</td>
<td>3.22</td>
</tr>
<tr>
<td>1</td>
<td>2,139</td>
<td>1,391</td>
<td>1,076</td>
<td>901</td>
<td>785</td>
<td>720</td>
<td>620</td>
<td>589</td>
</tr>
<tr>
<td>Number of Subroutes</td>
<td>75</td>
<td>43</td>
<td>30</td>
<td>25</td>
<td>19</td>
<td>20</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.92</td>
<td>4.38</td>
<td>3.34</td>
<td>2.76</td>
<td><strong>1.9</strong></td>
<td><strong>2.09</strong></td>
<td>3.06</td>
<td>3.91</td>
</tr>
<tr>
<td>1.5</td>
<td>2,137</td>
<td>1,381</td>
<td>1,049</td>
<td>865</td>
<td>743</td>
<td>676</td>
<td>529</td>
<td>473</td>
</tr>
<tr>
<td>Number of Subroutes</td>
<td>74</td>
<td>42</td>
<td>29</td>
<td>24</td>
<td>19</td>
<td>20</td>
<td>26</td>
<td>28</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.89</td>
<td>4.5</td>
<td>3.32</td>
<td>2.74</td>
<td><strong>2.03</strong></td>
<td><strong>2.22</strong></td>
<td>3.73</td>
<td>4.55</td>
</tr>
<tr>
<td>2</td>
<td>2,135</td>
<td>1,375</td>
<td>1,039</td>
<td>852</td>
<td>725</td>
<td>657</td>
<td>491</td>
<td>412</td>
</tr>
<tr>
<td>Number of Subroutes</td>
<td>74</td>
<td>42</td>
<td>29</td>
<td>24</td>
<td>19</td>
<td>20</td>
<td>26</td>
<td>29</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.89</td>
<td>4.35</td>
<td>3.32</td>
<td>2.74</td>
<td><strong>2.03</strong></td>
<td><strong>2.22</strong></td>
<td>3.73</td>
<td>4.89</td>
</tr>
<tr>
<td>2.5</td>
<td>2,134</td>
<td>1,374</td>
<td>1,035</td>
<td>845</td>
<td>717</td>
<td>648</td>
<td>473</td>
<td>378</td>
</tr>
<tr>
<td>Number of Subroutes</td>
<td>74</td>
<td>42</td>
<td>29</td>
<td>24</td>
<td>19</td>
<td>20</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.89</td>
<td>4.35</td>
<td>3.32</td>
<td>2.74</td>
<td><strong>2.03</strong></td>
<td><strong>2.22</strong></td>
<td>3.73</td>
<td>5.77</td>
</tr>
<tr>
<td>3.0</td>
<td>2,130</td>
<td>1,369</td>
<td>1,030</td>
<td>839</td>
<td>711</td>
<td>637</td>
<td>454</td>
<td>355</td>
</tr>
<tr>
<td>Number of Subroutes</td>
<td>73</td>
<td>41</td>
<td>28</td>
<td>23</td>
<td>18</td>
<td>20</td>
<td>27</td>
<td>32</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.89</td>
<td>4.34</td>
<td>3.32</td>
<td>2.73</td>
<td><strong>2.03</strong></td>
<td><strong>2.41</strong></td>
<td>4.15</td>
<td>6.20</td>
</tr>
<tr>
<td>3.5</td>
<td>2,130</td>
<td>1,369</td>
<td>1,030</td>
<td>839</td>
<td>711</td>
<td>637</td>
<td>451</td>
<td>351</td>
</tr>
<tr>
<td>Number of Subroutes</td>
<td>72</td>
<td>40</td>
<td>27</td>
<td>23</td>
<td>18</td>
<td>20</td>
<td>28</td>
<td>33</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.89</td>
<td>4.34</td>
<td>3.32</td>
<td>2.73</td>
<td><strong>2.03</strong></td>
<td><strong>2.41</strong></td>
<td>4.41</td>
<td>6.46</td>
</tr>
<tr>
<td>4.0</td>
<td>2,130</td>
<td>1,369</td>
<td>1,030</td>
<td>838</td>
<td>710</td>
<td>635</td>
<td>448</td>
<td>347</td>
</tr>
<tr>
<td>Number of Subroutes</td>
<td>72</td>
<td>40</td>
<td>27</td>
<td>23</td>
<td>18</td>
<td>20</td>
<td>28</td>
<td>33</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.89</td>
<td>4.34</td>
<td>3.32</td>
<td>2.73</td>
<td><strong>2.03</strong></td>
<td><strong>2.41</strong></td>
<td>4.41</td>
<td>6.46</td>
</tr>
</tbody>
</table>

higher cut-off points result in a high error percentage. This model is not robust for any cut-off point combinations.
**Figure 1**: Solution 1 - OR - Impact Errors of False Predictions

**Table 3**: Solution 1 - AND

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>No of Subroutes</th>
<th>No of Errors</th>
<th>Relative Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>23</td>
<td>2.96</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>24</td>
<td>3.88</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>24</td>
<td>5.05</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>28</td>
<td>5.39</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>557</td>
<td>2.27</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>557</td>
<td>4.04</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>437</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>437</td>
<td>5.39</td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>28</td>
<td>5.05</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>28</td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>28</td>
<td>8.08</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>28</td>
<td>11.89</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>297</td>
<td>5.05</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>297</td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>297</td>
<td>8.08</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>297</td>
<td>11.89</td>
</tr>
<tr>
<td>2.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>269</td>
<td>5.05</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>269</td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>269</td>
<td>8.08</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>269</td>
<td>11.89</td>
</tr>
<tr>
<td>3.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>34</td>
<td>5.05</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>34</td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>34</td>
<td>8.08</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>34</td>
<td>11.89</td>
</tr>
<tr>
<td>3.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>38</td>
<td>5.05</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>38</td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>38</td>
<td>8.08</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>38</td>
<td>11.89</td>
</tr>
<tr>
<td>4.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>216</td>
<td>5.05</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>216</td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>216</td>
<td>8.08</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>216</td>
<td>11.89</td>
</tr>
<tr>
<td>4.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>40</td>
<td>5.05</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>40</td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>40</td>
<td>8.08</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>40</td>
<td>11.89</td>
</tr>
</tbody>
</table>
Figure 2: Solution 1 - AND - Impact Errors of False Predictions

Table 4: Solution 2 - OR

<table>
<thead>
<tr>
<th>Cut-off points for ‘Number of occurrence’</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>100</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of Subroutes</td>
<td>2,620</td>
<td>2,076</td>
<td>1,906</td>
<td>1,823</td>
<td>1,781</td>
<td>1,766</td>
<td>1,742</td>
<td>1,742</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>88</td>
<td>57</td>
<td>45</td>
<td>41</td>
<td>39</td>
<td>39</td>
<td>39</td>
<td>40</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>6.15</td>
<td>4.67</td>
<td>3.72</td>
<td>3.26</td>
<td>2.98</td>
<td>2.98</td>
<td>3.22</td>
<td>3.22</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>2,349</td>
<td>1,664</td>
<td>1,391</td>
<td>1,249</td>
<td>1,163</td>
<td>1,121</td>
<td>1,061</td>
<td>1,035</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>80</td>
<td>47</td>
<td>34</td>
<td>30</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.95</td>
<td>4.38</td>
<td>3.34</td>
<td>2.89</td>
<td>2.32</td>
<td>2.32</td>
<td>2.52</td>
<td>2.91</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>2,275</td>
<td>1,569</td>
<td>1,275</td>
<td>1,109</td>
<td>1,003</td>
<td>950</td>
<td>863</td>
<td>825</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>76</td>
<td>42</td>
<td>29</td>
<td>23</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.93</td>
<td>4.29</td>
<td>3.26</td>
<td>2.57</td>
<td><strong>2.01</strong></td>
<td><strong>1.99</strong></td>
<td><strong>2.39</strong></td>
<td>2.72</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>2,247</td>
<td>1,526</td>
<td>1,224</td>
<td>1,050</td>
<td>938</td>
<td>882</td>
<td>777</td>
<td>735</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>75</td>
<td>42</td>
<td>29</td>
<td>23</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.91</td>
<td>4.31</td>
<td>3.28</td>
<td>2.59</td>
<td><strong>1.88</strong></td>
<td><strong>2.07</strong></td>
<td><strong>2.44</strong></td>
<td>2.87</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>2,229</td>
<td>1,500</td>
<td>1,198</td>
<td>1,019</td>
<td>903</td>
<td>841</td>
<td>721</td>
<td>672</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>74</td>
<td>41</td>
<td>28</td>
<td>22</td>
<td>17</td>
<td>18</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.88</td>
<td>4.28</td>
<td>3.25</td>
<td>2.56</td>
<td><strong>1.86</strong></td>
<td><strong>2.04</strong></td>
<td><strong>2.48</strong></td>
<td>2.87</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>2,203</td>
<td>1,467</td>
<td>1,150</td>
<td>968</td>
<td>849</td>
<td>782</td>
<td>633</td>
<td>564</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>75</td>
<td>43</td>
<td>29</td>
<td>23</td>
<td>17</td>
<td>18</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.90</td>
<td>4.35</td>
<td>3.25</td>
<td>2.56</td>
<td><strong>1.71</strong></td>
<td><strong>1.90</strong></td>
<td><strong>2.47</strong></td>
<td>2.85</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>2,193</td>
<td>1,455</td>
<td>1,130</td>
<td>950</td>
<td>830</td>
<td>760</td>
<td>601</td>
<td>524</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>74</td>
<td>42</td>
<td>28</td>
<td>24</td>
<td>18</td>
<td>19</td>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.90</td>
<td>4.35</td>
<td>3.25</td>
<td>2.67</td>
<td><strong>1.81</strong></td>
<td><strong>2.00</strong></td>
<td><strong>2.90</strong></td>
<td>4.05</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>2,193</td>
<td>1,455</td>
<td>1,130</td>
<td>950</td>
<td>830</td>
<td>760</td>
<td>601</td>
<td>524</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>74</td>
<td>42</td>
<td>28</td>
<td>24</td>
<td>18</td>
<td>19</td>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>5.90</td>
<td>4.35</td>
<td>3.25</td>
<td>2.67</td>
<td><strong>1.81</strong></td>
<td><strong>1.99</strong></td>
<td><strong>2.89</strong></td>
<td>4.05</td>
</tr>
</tbody>
</table>
Table 5: Solution 2 - AND

<table>
<thead>
<tr>
<th>Cut-off points for ‘Number of occurrence’</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>100</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1,236</td>
<td>1,017</td>
<td>808</td>
<td>738</td>
<td>651</td>
<td>589</td>
<td>425</td>
<td>320</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>25</td>
<td>24</td>
<td>23</td>
<td>22</td>
<td>19</td>
<td>21</td>
<td>28</td>
<td>33</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>2.96</td>
<td>2.9</td>
<td>2.82</td>
<td>2.69</td>
<td>2.27</td>
<td>2.65</td>
<td>4.41</td>
<td>6.45</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>798</td>
<td>720</td>
<td>654</td>
<td>603</td>
<td>560</td>
<td>525</td>
<td>397</td>
<td>318</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>19</td>
<td>20</td>
<td>20</td>
<td>19</td>
<td>18</td>
<td>20</td>
<td>27</td>
<td>33</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>1.98</td>
<td>2.02</td>
<td>1.89</td>
<td>1.75</td>
<td>2.13</td>
<td>3.93</td>
<td>6.45</td>
<td></td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>655</td>
<td>598</td>
<td>553</td>
<td>526</td>
<td>503</td>
<td>479</td>
<td>378</td>
<td>311</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>18</td>
<td>20</td>
<td>20</td>
<td>21</td>
<td>20</td>
<td>21</td>
<td>28</td>
<td>34</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>2.52</td>
<td>2.61</td>
<td>2.62</td>
<td>2.73</td>
<td>2.59</td>
<td>2.78</td>
<td>4.58</td>
<td>7.1</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>585</td>
<td>543</td>
<td>506</td>
<td>487</td>
<td>470</td>
<td>449</td>
<td>366</td>
<td>303</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>18</td>
<td>19</td>
<td>19</td>
<td>20</td>
<td>20</td>
<td>21</td>
<td>28</td>
<td>34</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>2.39</td>
<td>2.44</td>
<td>2.45</td>
<td>2.55</td>
<td>2.55</td>
<td>2.75</td>
<td>4.58</td>
<td>7.1</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>534</td>
<td>500</td>
<td>463</td>
<td>449</td>
<td>436</td>
<td>421</td>
<td>353</td>
<td>297</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>22</td>
<td>23</td>
<td>23</td>
<td>24</td>
<td>24</td>
<td>25</td>
<td>31</td>
<td>35</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>4.58</td>
<td>4.63</td>
<td>4.64</td>
<td>4.74</td>
<td>4.75</td>
<td>4.94</td>
<td>6.5</td>
<td>8.16</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>490</td>
<td>458</td>
<td>430</td>
<td>419</td>
<td>408</td>
<td>396</td>
<td>343</td>
<td>296</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>24</td>
<td>24</td>
<td>25</td>
<td>31</td>
<td>35</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>4.57</td>
<td>4.57</td>
<td>4.65</td>
<td>4.75</td>
<td>4.75</td>
<td>4.95</td>
<td>6.51</td>
<td>8.17</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>440</td>
<td>413</td>
<td>391</td>
<td>380</td>
<td>370</td>
<td>360</td>
<td>321</td>
<td>285</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>26</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>35</td>
<td>37</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>7.18</td>
<td>7.18</td>
<td>7.26</td>
<td>7.36</td>
<td>7.51</td>
<td>7.72</td>
<td>9.04</td>
<td>9.93</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>418</td>
<td>392</td>
<td>373</td>
<td>364</td>
<td>364</td>
<td>346</td>
<td>313</td>
<td>281</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>28</td>
<td>28</td>
<td>29</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td>35</td>
<td>37</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>7.52</td>
<td>7.52</td>
<td>7.60</td>
<td>7.60</td>
<td>7.75</td>
<td>7.95</td>
<td>9.05</td>
<td>9.94</td>
</tr>
<tr>
<td>No of Subroutes</td>
<td>396</td>
<td>371</td>
<td>354</td>
<td>345</td>
<td>336</td>
<td>329</td>
<td>300</td>
<td>272</td>
</tr>
<tr>
<td>Number of Errors</td>
<td>29</td>
<td>29</td>
<td>30</td>
<td>30</td>
<td>31</td>
<td>32</td>
<td>36</td>
<td>38</td>
</tr>
<tr>
<td>Relative Errors</td>
<td>8.02</td>
<td>8.02</td>
<td>8.10</td>
<td>8.10</td>
<td>8.25</td>
<td>8.45</td>
<td>9.55</td>
<td>10.44</td>
</tr>
</tbody>
</table>
Appendix D

OD Algorithm Flow Charts

This section describes how the entire algorithm is structured. The algorithm is programmed in C++ which inputs/outputs from and to various text files. The aim was to include expert knowledge in the form of rules into the algorithm. The following steps are carried out by the algorithm:

- Initially all necessary parameters have to be set. This includes the following:
  - Run substitutional route identification algorithm (Yes/No);
  - Recalculate substitutional route matrix with new cut-off values (Yes/No);
  - Set value for cut-off value for number of occurrences;
  - Set value for cut-off value for calculated weights;
  - Run OD identification for single journeys (Yes/No);
  - Run OD identification for transfer journeys (Yes/No);
  - Output Single OD pair in 1 line (Yes/No);
  - Output Single OD pair in 2 lines - One for each boarding record (Yes/No);
  - Output Transfer OD pair in 1 line (Yes/No);
  - Output Transfer OD pair in 4 lines - One for each boarding record (Yes/No);
  - Output Single OD pairs directly into the Oracle database (Yes/No);
  - Output Transfer OD pairs directly into the Oracle database (Yes/No).

- If the 'Run substitutional route identification algorithm' is selected then the algorithm initially compiles a new list of substitutional routes. This procedure was described in Section 6.5;
If ‘Recalculate substitutional route matrix with new cut-off values’ is selected the algorithm re-calculates the substitutional route list using the new values of the cut-off points;

The first data file is opened;

All boarding records of one passenger are extracted;

The next part of the algorithm compares boarding records of single journeys. This contains the comparison of all single journey boarding records of this passenger. The comparison is done in pairs. A valid OD pair can only be extracted when the two boarding records were recorded on the same day. The following procedure is carried out for each comparison:

- **Route** - if it is the same route then the route identifier value is 1. In the case where the routes are different it has to be tested whether they are substitutional routes. This is done by accessing the substitutional route file and comparing the two routes from the potential OD pair to the list of substitutional routes stored in the file. If the route combination pair is considered as a substitutional route then the value is 2 otherwise the route identifier is 3;

- **Direction** - Same direction of the two boardings of the potential OD pair results in a 1 and different direction in a 2 as values for the direction identifier;

- **Repeated** - At this stage a positive OD pair has been identified and the algorithm’s aim is to find same OD pairs of the same passenger throughout the validity period of the ticket. In the case where an OD pair is repeatedly found the repeat identifier changes. 1 for repeats on the same day, 2 for repeats during the ticket duration and 3 for not repeated are assigned. The substitutional route identification algorithm is also applied throughout the search for repeated occurrences of the OD pair. A further output is the actual number of repeats that were found;

- **Order** - This part of the algorithm checks whether the boarding records which build the OD pair were in consecutive sequential order or not. 1 is assigned to consecutive order and 2 for non-consecutive boarding records.

This part of the algorithm compares boarding records of transfer journeys. The characteristics of transfer OD pairs are different as 4 boarding records are required to form a valid OD pair. A valid OD pair of a transfer journey can only be extracted if all four boardings were recorded on the same day. After identifying 4 boarding records that potentially could form an OD pair its characteristics have to be identified. The following procedure is carried out for each comparison:

- **Route A/D** - This compares the routes of boarding record A and D. The route has to be the same or substitutional. The Route A/D identifier is 1 for same routes, 2 for substitutional routes and 3 for different routes;
Route B/C - This compares the routes of boarding record B and C. The route has to be the same or substitutional. The Route B/C identifier is 1 for same routes, 2 for substitutional routes and 3 for different routes;

Direction A/D - This compares the direction of boarding record A and D. Same direction is a 1 and different direction is a 2;

Direction B/C - This compares the direction of boarding record B and C. Same direction is a 1 and different direction is a 2;

Repeated - At this stage a positive OD pair has been identified and the algorithm’s aim is to find same OD pairs of the same passenger throughout the validity period of the ticket. In case an OD pair is found repeatedly the repeat identifier changes. 1 for repeats on the same day, 2 for repeats during the ticket duration and 3 for not repeated are assigned. The substitutional route identification algorithm is also applied throughout the search for repeated occurrences of the OD pair. A further output is the actual number of repeats that were found;

Order - This part of the algorithm checks whether all four boarding records, which build the OD pair, were in consecutive sequential order or not. 1 is assigned to consecutive order and 2 for non-consecutive boarding records.

- The result of the above described algorithm is a set of numbers that describe the characteristics found of the OD pair which now have to be matched to the rule base. Depending on the journey type (single or transfer) the algorithm accesses the rule base and attempts to match the extracted characteristics of the potential OD pair with one of the pre-defined rules. Once a positive match is identified the algorithm outputs the boarding records as verified OD pair using the OD scenario assigned to the rule;

- If there are further boarding records in the file then the next set of boarding records of one particular passenger are extracted and the above described procedure is carried out again in the attempt to find new OD pairs. If no boarding records are left to analyse the next file is opened and the algorithm extracts the sets of boarding records from there.

The following figures (see Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, Figure 11, Figure 12) represent the previous section in the form of process diagrams. The squares with the double lined borders indicate a sub procedure.
Figure 5: Main Procedure of the OD Extraction Algorithm
Start (Receives passenger data array OD (x, y); x=0; y=0)

Read boarding record x

No (x++)

Single Journey

Yes (y=x++)

Read boarding record y

No (y++)

Yes (set y++)

Compare Dates

Same Day

Yes

Compare direction and set identifier

Same direction

No

Check for order of records and set identifier

Check for OD pair repeats

Check for match in rule base (single journeys)

Positive Match

Yes

Output OD pair

Return

No

Sub-Route

Same route

Yes (pass route identifier)

Set route identifier

Check for substitutional routes

No

End of Array

Figure 6: Sub Procedure: Check for Single Journey OD Pairs
Figure 7: Sub Procedure: Checking for Substitutional Routes
Start (Receives set of identifiers; x=0)

Open rule base (single journeys)

Read line x

Match found

No (x++)

Yes

Extract OD pair type

End (Return OD pair type)

Figure 8: Sub Procedure: Matching OD Pairs with the Single Journey Rule Base
Figure 9: Sub Procedure: Checking for Single Journey Repeats
D. OD ALGORITHM FLOW CHARTS

217

Figure 10: Sub Procedure: Transfer Journey OD Pair Identification
Start (Receives set of identifiers; x=0)

Open rule base (transfer Journeys)

Read line x

No

Match found

Yes

Extract OD pair types

End (Return OD pair type)

Figure 11: Sub Procedure: Matching OD Pairs with the Transfer Journey Rule Base
D. OD ALGORITHM FLOW CHARTS

Start (Receives OD information; x=n; y=n; k=0; k'=

No (x>n)

No (if consecutive then y=n++ else y=y++)

Same Day

Read boarding record x

Transfer Journey

Yes (x=n++)

No (x=n++)

Compare direction of x/l

Same Day

Compare dates of records x,y,l and k

Compare OD pair

Compare Dates

Same Direction

Same Day

Transfer Journey

Read boarding record l

Yes (kx=x++)

Yes (kx+y++)

Same Direction

Compare direction of y/l

Yes (k++++)

Read boarding record y

Transfer Journey

No (y=n++)

Read boarding record y

Transfer Journey

Check for substitutional route

Same route

Sub route

Yes

No

Set route identifier

Sub route

No

Return

No (l++)

Count OD pair as repeated

Yes

End of Array

Positive Match

Compare OD pair with received information

Same route

Sub route

Yes

No

Set route identifier

Sub route

Figure 12: Sub Procedure: Checking for Transfer Journey Repeats