The development and application of an urban link travel time model using data derived from inductive loop detectors

Steve Robinson

A thesis submitted for the degree of Doctor of Philosophy of the University of London and Diploma of the Membership of Imperial College London

Centre for Transport Studies
Department of Civil and Environmental Engineering
Imperial College London, United Kingdom

August 2005
Abstract

The need to measure urban link travel time (ULTT) is becoming increasingly important for the purposes both of network management and traveller information provision. This thesis develops a methodology by which data from single inductive loop detectors (ILDs) can be used to derive estimates of ULTT. Research is then undertaken to assess how effectively travel time data from this ULTT model can be used to model travel time variability (TTV).

This research first looks at the issue of data quality. The Overtaking Rule method is developed to clean travel time data derived from ANPR cameras in London. A way of quantifying the effectiveness of ILD data cleaning treatments is proposed and the Daily Statistics Algorithm is identified as the best treatment.

There are various limitations in existing ULTT models. To address these limitations, an alternative k nearest neighbors (k-NN) based method is proposed for use as a ULTT model. The key design parameters are identified and a 5-step process to optimise this model proposed. The k-NN method is found to outperform all other ULTT models on two links in central London. It is also found to be robust to changes in network characteristics and does not necessarily have to be optimised for each link.

Data from the k-NN ULTT model allows various components of TTV to be modelled. In particular, it allows the true patterns of day-to-day variation to be identified. It also correctly models the distribution of travel time in approximately 50% of 15-minute time periods.

As an application, the k-NN method could be used to aggregate GPS travel time records from different times but the same underlying travel time distribution together to enable more accurate estimates of ULTT.
Acknowledgements

I would first of all like to thank my supervisor, Professor John Polak, for all the many discussions and critique of my work throughout my time as his PhD student. I have John to thank for constantly challenging me and advising me on areas of my thesis which could be improved and made more rigorous. However I have also appreciated the friendship that has developed over the past 4 years, even if John was both a Statistician and a Chelsea fan!

I am very grateful to Dr John Tough of Transport for London (TfL), who provided me with the ILD data that was used within this thesis. I must also thank John for helping me obtain my current job with London Buses at TfL.

I would also like to thank Charles Buckingham of TfL for providing me with the ANPR travel time data.

There are many other people I would like to thank for helping me to get to this stage of my PhD. However, for space concerns I shall name only those who have had an input in my academic life, and have thus had an indirect input into this thesis:

- Dr Bob Noland for the many useful discussions in the corridor outside room 613 and the coffee machine - i.e. the ‘kettle and big tub of Nescafe’. At my darkest moments of data deprivation Bob always encouraged me - thanks!
- Dr Washington Ochieng for discussions on navigation related topics as well as drilling into me the importance of data quality!
- Knut Sauer for helping me when I first started at Imperial and for being a successful PhD candidate to emulate!
- Mohammed Quddus for his patience in explaining various matters of a statistical nature to me. Thanks also for being a great office mate as well as a room mate at two TRB conferences in Washington D.C.
- Stephane Hess for helping me quickly learn \LaTeX when Microsoft Word so badly let me down! Thanks also for challenging me in my own work, and for the many evenings spent in the Zetland Arms and Southside.
• Christophe Begon for assisting me with the validation of the ANPR travel time data.

• Jackie Sime for being a super sectional secretary and for providing an outlet to moan at when my PhD life and my supervisor were getting me down!
Contents

Abstract 2

Acknowledgements 3

Table of Contents 4

List of Figures 14

List of Tables 19

1 Introduction 23

1.1 Background ........................................ 23

1.1.1 Applications of travel time data .................. 23

1.1.2 Measurement of Travel Time ....................... 24

1.1.3 Estimation of Travel Time ........................ 24

1.1.4 Cleaning of travel time and ILD data .............. 25

1.2 Aims and Objectives ................................. 25

1.2.1 Objectives of part A - Data for a ULTT model .... 26

1.2.2 Objectives of part B - Modelling ULTT .......... 26

1.2.3 Objectives of part C - Application of a ULTT model in study on TTV 27

1.3 Thesis Outline ...................................... 27

1.3.1 Part A ............................................. 27

1.3.2 Part B ............................................. 28

1.3.3 Part C ............................................. 28

Part A 29

2 Measuring Actual Link Travel Time using ANPR data 30

2.1 Definition of a link .................................. 30

2.1.1 Urban Links ...................................... 30

2.1.2 Proposed nomenclature ........................... 31

2.2 Definition of link travel time ....................... 32

2.2.1 Definition of link travel time: literature review .... 32

2.2.2 Categories of link travel time distributions ....... 33
2.11.1 Experiment #1 - single-lane road ........................................ 58
2.11.2 Experiment #2 - multi-lane road ........................................ 60
2.12 ANPR data from TfL congestion charging cameras ..................... 61
2.13 Validation of ANPR data ....................................................... 62
  2.13.1 Methodology ................................................................. 64
  2.13.2 Comparison of ANPR data cleaning methods ......................... 65
  2.13.3 Results of the Kolmogorov-Smirnov test ............................ 65
  2.13.4 Conclusions ................................................................. 66
2.14 Caveat: Link-based or path based travel time models .................. 66
2.15 Conclusions ........................................................................ 67

3 Data from Inductive Loop Detectors ........................................ 69
  3.1 Inductive Loop Detectors ....................................................... 69
  3.2 Basic Traffic Parameters obtained from ILDs .......................... 70
    3.2.1 Volume and Flow ......................................................... 70
    3.2.2 Occupancy ................................................................. 71
    3.2.3 Density ...................................................................... 71
    3.2.4 Average Loop Occupancy Time Per Vehicle (ALOTPV) ........ 71
    3.2.5 Average Gap Time Between Vehicles (AGTBV) .................. 71
    3.2.6 Headway .................................................................. 72
  3.3 Errors associated with ILD data ............................................ 72
    3.3.1 Errors collecting the raw ILD data .................................. 72
    3.3.2 Errors in calculating the basic traffic parameters ............... 74
  3.4 Existing quality control methods for ILD data .......................... 75
    3.4.1 Univariate Tests .......................................................... 76
    3.4.2 Daily Statistics Algorithm ............................................. 77
    3.4.3 Multivariate Tests ....................................................... 79
    3.4.4 Quality control statistics .............................................. 81
  3.5 Adjacent Detector Test ......................................................... 81
    3.5.1 Introduction ................................................................. 81
    3.5.2 Relationship between flows at adjacent detectors ............... 82
    3.5.3 Test based on relationship between flows at adjacent detectors 85
  3.6 Quantifying the performance of different ILD data cleaning treatments .. 85
  3.7 Proposed Quality Control Methodology for ILD data ................. 86
  3.8 Conclusion ......................................................................... 87

4 ILD data used in the research ................................................ 88
  4.1 London’s SCOOT system ....................................................... 88
    4.1.1 Introduction to SCOOT .................................................. 88
    4.1.2 Characteristics of ILDs in London .................................. 89
  4.2 Traffic archive systems: literature review ................................ 90
4.2.1 PeMS ............................. 90
4.2.2 ADUS project, INFORM and SMIS ............................. 92
4.2.3 Instrumented City ............................. 92
4.2.4 ROMANSE ............................. 92
4.2.5 Highways Agency Regional Control Centres ............................. 93
4.3 LSAD ............................. 93
4.3.1 Architecture of LSAD ............................. 93
4.3.2 LSAD Stage 1 - Collection and temporary storage of raw ILD data ........................................ 95
4.3.3 LSAD Stage 2 - Transfer of data from Transport for London to Imperial College London ............................. 95
4.3.4 LSAD Stage 3 - Processing and archiving of raw ILD data ........................................ 95
4.3.5 LSAD Stage 4 - Retrieval and Analysis of data from the Saturn machine ............................. 96
4.3.6 Validation of data stored in LSAD ............................. 96
4.4 LSAD, ASTRID, and COMET ............................. 97
4.5 ASTRID ILD data used in this thesis ............................. 98
4.6 ILD data quality ............................. 100
4.7 Primary data set used to train ULTT models: Russell Square 2003 ........................................ 101
4.8 Alternative data sets ............................. 102
4.8.1 Tottenham Court Road - 2003 ........................................ 102
4.8.2 Russell Square - 2004 ........................................ 103
4.8.3 Tottenham Court Road - 2004 ........................................ 105
4.9 Conclusions ............................. 105

Part B ............................. 106

5 Review of Link Travel Time Literature ............................. 107
5.1 Definition of ULTT literature ............................. 108
5.2 Components of ULTT ............................. 108
5.3 Characterising the performance of a ULTT model ........................................ 109
5.3.1 Desired properties of a ULTT model ........................................ 109
5.3.2 Quantifying Accuracy ........................................ 110
5.4 Spot speed LTT models ........................................ 114
5.4.1 Spot-speed, time-mean-speed, and space-mean-speed ........................................ 115
5.4.2 Dual loop spot speed measures ........................................ 116
5.4.3 Basic single loop measures ........................................ 117
5.4.4 Biased estimators ........................................ 118
5.4.5 Violation of assumption of homogenous SMS ........................................ 119
5.5 Correlation LTT models ........................................ 119
5.5.1 Basic principles of cross correlation ........................................ 119
5.5.2 Signature matching ........................................ 120
<table>
<thead>
<tr>
<th>Section</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5.3</td>
<td>Platoon Matching</td>
</tr>
<tr>
<td>5.6</td>
<td>Queuing theory based LTT models</td>
</tr>
<tr>
<td>5.6.1</td>
<td>SCOOT and TRANSYT models</td>
</tr>
<tr>
<td>5.6.2</td>
<td>Other Queuing models</td>
</tr>
<tr>
<td>5.6.3</td>
<td>Sensitivity to saturation flow</td>
</tr>
<tr>
<td>5.6.4</td>
<td>Difficulty of collecting data inputs</td>
</tr>
<tr>
<td>5.7</td>
<td>Regression based LTT models</td>
</tr>
<tr>
<td>5.8</td>
<td>Artificial Neural Network Models</td>
</tr>
<tr>
<td>5.9</td>
<td>Emerging ULTT models</td>
</tr>
<tr>
<td>5.9.1</td>
<td>Fusion Techniques</td>
</tr>
<tr>
<td>5.9.2</td>
<td>k-Nearest Neighbors techniques</td>
</tr>
<tr>
<td>5.10</td>
<td>Summary of Existing ULTT models</td>
</tr>
<tr>
<td>5.11</td>
<td>Modelling turning movements</td>
</tr>
<tr>
<td>5.12</td>
<td>Modelling Mid-Link Delay</td>
</tr>
<tr>
<td>5.12.1</td>
<td>Stopping and Turning vehicles</td>
</tr>
<tr>
<td>5.12.2</td>
<td>Incident Delay</td>
</tr>
<tr>
<td>5.12.3</td>
<td>INCIBEN and INCA</td>
</tr>
<tr>
<td>5.12.4</td>
<td>Conclusion</td>
</tr>
<tr>
<td>5.13</td>
<td>Required data input into ULTT models</td>
</tr>
<tr>
<td>5.14</td>
<td>Limitations of using data solely from ILDs</td>
</tr>
<tr>
<td>5.14.1</td>
<td>Example #1 - Data from ILDs other than flow and occupancy may be required.</td>
</tr>
<tr>
<td>5.14.2</td>
<td>Example #2 - Importance of the location of an ILD on the link in detecting mid-link delay</td>
</tr>
<tr>
<td>5.14.3</td>
<td>Eliciting non-traffic variables that influence travel time</td>
</tr>
<tr>
<td>5.14.4</td>
<td>The use of temporal metrics as a proxy of underlying factors</td>
</tr>
<tr>
<td>5.15</td>
<td>Conclusions</td>
</tr>
<tr>
<td>6</td>
<td>The k-Nearest Neighbors Method</td>
</tr>
<tr>
<td>6.1</td>
<td>Basic Concept</td>
</tr>
<tr>
<td>6.2</td>
<td>Theoretical basis of the k-NN method</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Historical background of the k-NN method</td>
</tr>
<tr>
<td>6.3</td>
<td>Assumptions of the k-NN method</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Similar a posteriori distributions</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Large historical database</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Random distribution of the k-nearest neighbors about the input feature vector</td>
</tr>
<tr>
<td>6.4</td>
<td>Key parameters of the k-NN method</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Value of k</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Distance Metric</td>
</tr>
<tr>
<td>6.4.3</td>
<td>Local estimation method</td>
</tr>
</tbody>
</table>
6.4.4 Attributes to include in the feature vector ........................................ 164
6.5 Computational Issues ................................................................................ 165
  6.5.1 Univariate Screening ............................................................................. 166
  6.5.2 Problems with univariate screening ....................................................... 166
6.6 k-NN Transport Applications: Literature Review ........................................ 167
  6.6.1 Application To Traffic Flow Estimation ............................................... 168
  6.6.2 Application To Travel Time Estimation ............................................... 169
  6.6.3 Critique of existing usage of the k-NN method within transportation 171
6.7 Conclusions ................................................................................................. 172

7 Characteristics of the k-NN method when used as a ULTT model .......................... 173
  7.1 Introduction ............................................................................................... 173
  7.2 Using the k-NN as a ULTT model .............................................................. 173
    7.2.1 Collection of historical data .............................................................. 174
    7.2.2 Specification of k-NN model parameters ......................................... 174
    7.2.3 Collection of data for input feature vector ....................................... 175
    7.2.4 Determination of the k closest neighbors ........................................ 175
    7.2.5 Estimation of travel time from the k closest neighbors .................. 176
    7.2.6 Conclusions ......................................................................................... 176
  7.3 Testing Basic Assumptions ........................................................................ 177
    7.3.1 Data .................................................................................................... 177
    7.3.2 Methodology ...................................................................................... 178
    7.3.3 Test of equivalent distributions ....................................................... 178
    7.3.4 Results ............................................................................................... 179
    7.3.5 Conclusion .......................................................................................... 180
  7.4 Determining the optimal k-NN settings ....................................................... 181
    7.4.1 Methodology ...................................................................................... 181
    7.4.2 Data .................................................................................................... 183
    7.4.3 Performance Evaluation .................................................................... 183
  7.5 Stage #1: Factors considered independently of one-another ...................... 183
    7.5.1 Methodology and Data ...................................................................... 184
    7.5.2 Results ............................................................................................... 184
    7.5.3 Conclusions ....................................................................................... 185
  7.6 Stage #2: Full factorial grid search ............................................................ 185
    7.6.1 Methodology and Data ...................................................................... 185
    7.6.2 Results ............................................................................................... 186
    7.6.3 Conclusions from the full factorial approach ................................... 188
  7.7 Stage #3a: Modelling the error ................................................................ 189
    7.7.1 Fitting the theoretical MSE curve .................................................... 189
    7.7.2 Cubic spline fitting ............................................................................ 194
  7.8 Stage #3b: Line Search ............................................................................. 194
7.9 Stage #4: Sensitivity Analysis .................................................. 197
  7.9.1 Methodology and Data ..................................................... 197
  7.9.2 Results ........................................................................... 198
7.10 Recommended Approach to determining optimal parameters for a k-NN
      ULTT model ........................................................................ 201
7.11 Optimal Settings ................................................................... 202
  7.11.1 Optimal ULTT model settings ........................................... 202
  7.11.2 Optimal MAPE settings ................................................... 203
  7.11.3 Optimal RMSE settings .................................................... 205
7.12 Effect of the size of the underlying database on the performance of the k-NN
      model ................................................................................... 205
  7.12.1 Methodology and Data ..................................................... 206
  7.12.2 Results ........................................................................... 207
7.13 Conclusions ........................................................................... 208

8 Comparison of the k-NN method against other ULTT models ......... 211
  8.1 Other ULTT models .................................................................. 211
    8.1.1 Basic Regression ............................................................. 212
    8.1.2 Gault and Taylor .............................................................. 212
    8.1.3 Artificial Neural Network ............................................... 213
    8.1.4 Naïve Mean .................................................................. 213
    8.1.5 Naïve Median ................................................................. 213
    8.1.6 Time: day of week .......................................................... 213
    8.1.7 Time: period 15 ............................................................... 213
    8.1.8 Time: day of week & period 15 ........................................ 213
  8.2 Russell Square Link - 2003 ..................................................... 214
    8.2.1 Data and Methodology ................................................... 214
    8.2.2 Results ........................................................................... 215
  8.3 Tottenham Court Road Link - 2003 .......................................... 221
    8.3.1 Data and Methodology ................................................... 222
    8.3.2 Results ........................................................................... 222
  8.4 Russell Square Link - 2004 ..................................................... 224
    8.4.1 Data and Methodology ................................................... 224
    8.4.2 Results ........................................................................... 225
  8.5 Effect of the ILD data cleaning method .................................... 225
    8.5.1 Data for experiment ....................................................... 226
    8.5.2 Experiment Methodology ............................................... 226
    8.5.3 Results ........................................................................... 227
  8.6 Potential application of the k-NN method to GPS data ............... 234
  8.7 Conclusions ........................................................................... 234
Part C

9 Travel Time Variability: Literature Review

9.1 What causes travel time to vary

9.1.1 Drivers of Travel Time

9.1.2 Demand

9.1.3 Capacity

9.1.4 Vehicle Performance

9.1.5 Control

9.1.6 Interrelationship between the 4 generic drivers

9.2 Disaggregation of travel time variability

9.2.1 Disaggregation by travel time driver

9.2.2 Disaggregation by predictability

9.2.3 Disaggregation by location

9.2.4 Disaggregation by temporal scale

9.2.5 Caveat of disaggregating by temporal scale

9.3 Quantifying TTV

9.4 Empirical research into TTV on road networks

9.4.1 Temporal Scale TTV

9.4.2 Other empirical work on TTV

9.5 Other aspects of TTV

9.5.1 Other Modes of transport

9.5.2 Traveller behaviour towards travel time variability

9.5.3 Perception and Predictability

9.6 Limitations of existing research

10 Application of the k-NN method to the characterisation of TTV

10.1 Apportioning the sources of TTV

10.1.1 Disaggregating TTV into various temporal scales

10.1.2 Experiment and Methodology to decompose TTV into its various components

10.1.3 Results - ANPR data

10.1.4 Results - k-NN

10.1.5 Conclusion: Ability of the k-NN method to disaggregate TTV

10.2 Modelling day-to-day TTV using the k-NN method

10.2.1 Identification of similar days-of-the-week

10.2.2 Variation of d2d TTV over the course of the day

10.2.3 Relationship between d2d TTV and the mean travel time

10.2.4 Conclusion: Ability of the k-NN method to model d2d TTV

10.3 Modelling vehicle-to-vehicle TTV using the k-NN method

10.3.1 Characterising v2v TTV
Contents

13

10.3.2 Effect of the sample size on the estimation of the v2v metrics . . . . 271
10.3.3 Estimating the vehicle to vehicle TTV using the k-NN method . . . 274
10.3.4 Ability of the k-NN method to model the v2v distribution . . . . . . 274
10.3.5 Ability of the k-NN method to estimate the v2v metrics . . . . . . . 277
10.3.6 Variation of v2v TTV with mean travel time . . . . . . . . . . . . . 281
10.3.7 Conclusion: Ability of the k-NN method to model v2v TTV . . . . . 285
10.4 Conclusion: Ability of the k-NN method to model TTV . . . . . . . . . . . 286
11 Conclusions and Recommendations for Further Research

287

11.1 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 287
11.1.1 Define clearly the concept of an urban link and the travel time on
an urban link. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 287
11.1.2 Develop a methodology to clean matched ANPR data . . . . . . . . 288
11.1.3 Develop a methodology to clean data from ILDs. . . . . . . . . . . . 288
11.1.4 Create an archive to store ILD data obtained from the London
SCOOT system. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 288
11.1.5 Undertake a comprehensive literature review and critique of ULTT
models. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 289
11.1.6 Develop and optimise a ULTT model based on the k-NN method . . 289
11.1.7 Propose a set of criteria and performance indicators by which ULTT
models can be evaluated . . . . . . . . . . . . . . . . . . . . . . . . . 290
11.1.8 Evaluate the performance of various ULTT models . . . . . . . . . . 290
11.1.9 Undertake a comprehensive literature review into TTV. . . . . . . . 290
11.1.10 Identify ways to characterise and quantify TTV. . . . . . . . . . . . 290
11.1.11 Determine whether travel time estimates obtained from the ULTT
model developed on part B allows the TTV to be characterised accurately. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 291
11.2 Summary of the methodology for turning data from ILD into estimates of
ULTT . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 291
11.3 Recommendations for future research . . . . . . . . . . . . . . . . . . . . . . 293
11.3.1 Data Cleaning: future research . . . . . . . . . . . . . . . . . . . . . 293
11.3.2 ULTT models: future research . . . . . . . . . . . . . . . . . . . . . 294
11.3.3 Travel Time Variability: future research . . . . . . . . . . . . . . . . 295
11.4 Real world application . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 295
Appendices

296

A Basic Traffic Theory

297

A.1 Traffic Flow Theory . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 297
A.1.1 Flow . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 297
A.1.2 Density . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 298
A.1.3 Speed . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 299


List of Figures

2.1 Nomenclature for defining links ............................................. 32
2.2 Example of an S-S and a D-E link ............................................. 32
2.3 Vehicles travelling within the same short time interval (e.g. 15 minute period) on a link A-B, may have a range of travel times due to the random chance that they may get delayed either by traffic lights or some other factor mid-link. The distribution of travel times in any short interval can thus be described by a random variable. ............................................. 36
2.4 Difficulty of identifying outliers in non-stationary data ............... 47
2.5 Graph showing the probability of the target vehicle being valid given the difference in travel times of the two vehicles. It assumes that $TDA_{ij} = 0$ and $C = 45$. Top graph shows function 1. Bottom graph shows function 2. 52
2.6 Urban link created in VISSIM microsimulation .......................... 54
2.7 Map of the Tottenham Court Road and Russell Square links .......... 62
2.8 Scatterplot of individual vehicle travel times by time-of-day for the Tottenham Court Road link on Wednesday 9th April 2003. Data obtained from ANPR cameras ............................................. 63
2.9 Scatterplot of individual vehicle travel times by time-of-day for the Russell Square link on Wednesday 9th April 2003. Data obtained from ANPR cameras ............................................. 63
2.10 Link based methods should ensure that delay experienced solely by one turning movement is not included in the travel time of another turning movement. ............................................. 67
3.2 Log-normal distribution of the ratio of the flows of the two ILDs n02/253e1 and n02/055e1. ............................................. 83
3.3 Historical relationship between the two ILDs n02/253e1 and n02/055e1. .. 84
4.1 ILDs in London often span two lanes - ILD n02/253e1 on the Russell Square link ............................................. 91
4.2 Architecture of the London SCOOT Archive Database, LSAD .......... 94
List of Figures

4.3 There are four key components in LSAD stage #3 .......................... 95
4.4 Map of the location of ILDs on the Tottenham Court Road Link ........ 99
4.5 Map of the location of ILDs on the Russell Square Link .................. 100
4.6 Data collected from the Russell Square link in central London in 2003. .. 103
4.7 Data collected from the Tottenham Court Road in 2003. ................. 104
4.8 Data collected from the Russell Square link in central London in 2004. .. 105

5.1 Dual loop ILDs are able to measure vehicle spot-speed and vehicle length
directly. .................................................. 116
5.2 Basic Cross Correlation. ........................................... 120
5.3 The hybrid ULTT model proposed by Xie et al. (2004) uses an ANN to fuse
estimates of ULTT from a regression based model and travel times from
probe vehicles. .................................................. 134
5.4 Data from ILDs other than flow and occupancy may be required. .......... 146
5.5 Location of the ILD may determine its ability to indicate mid-link delay .. 146

6.1 Example of the use of the k-NN method when $k = 5$. ....................... 151
6.2 The k-NN method assumes that the historical observation come from the
same class of distribution. ...................................... 152
6.3 As $k$ increases the variance error decreases ............................... 155
6.4 As $k$ increases the bias error increases ................................ 156
6.5 Example of the use of the regression LEM .................................. 163
6.6 Univariate screening employed undertake to speed up search .......... 166
6.7 Univariate screening, if applied rashly, may eliminate observations that are
amongst the k-nearest neighbors. .................................. 167

7.1 Relationship between the probability that the validation and historical
records have the same underlying travel time distribution with the rank
of the distance between the two. .................................. 180
7.2 Methodology for optimising the k-NN ULTT model ........................ 181
7.3 Variation of MAPE with value of $k$ for each LEM and distance metric
combination. .................................................... 190
7.4 Variation of RMSE with value of $k$ for each LEM and distance metric com-
bination. ....................................................... 191
7.5 Fitted error curve for the setting LEM = mean, DM = SE-VAR. ....... 192
7.6 Fitted error curve for the setting LEM = median, DM = Mahalanobis ... 193
7.7 Fitted error curve for the setting LEM = median, DM = Mahalanobis ... 195
7.8 Actual versus Estimated travel times for the overall best k-NN ULTT model203
7.9 Actual versus Estimated travel times for the model which minimised MAPE 204
7.10 Actual versus Estimated travel times for the model which minimised RMSE 206
7.11 Effect of the size of the database and the value of $k$ on the performance of
the k-NN model. LEM = median, DM = SE-VAR .................... 208
7.12 Effect of the size of the database and the value of $k$ on the performance of the k-NN model. LEM = Lowess, DM = SE-VAR .......................... 209

8.1 Performance of the various ULTT models for the Russell Square 2003 data.
a. Actual vs Estimated Travel Time, b. Actual vs Residuals ................. 216

8.2 Performance of the k-NN Lowess based ULTT model against other ULTT models over the whole range of operation on the Russell Square Link - 2003 dataset. ................................................................. 220

8.3 Performance of the k-NN TCR model against other ULTT models over the whole range of operation on the Tottenham Court Road link - 2003 dataset. 223

8.4 pdf of the error (estimated - actual travel time) for all ILD treatments with the Basic Regression ULTT model .................................................. 229

8.5 Top graph: Plot of the pdf of $\Delta_i$. Bottom graph: LHS of top diagram (red line) projected onto RHS of top diagram .......................... 231

8.6 Difference between the pairwise residuals from the estimates using the DSA and Univariate data set. 12,500 validation records were used in the analysis 232

9.1 Scatterplot of individual vehicle travel times by time of day. The v2v TTV accounts for the ‘width’ of the plot. .............................. 245

10.1 Variation of mean travel time with the period of the day (Sunday through to Wednesday). Both ANPR and k-NN results are shown for the Russell Square 2004 dataset. ............................................. 262

10.2 Variation of mean travel time with the period of the day (Thursday through to Saturday). Both ANPR and k-NN results are shown for the Russell Square 2004 dataset. ............................................. 263

10.3 Variation of day-to-day TTV with the period of the day (Sunday through to Wednesday). Both ANPR and k-NN results are shown for the Russell Square 2004 dataset. ............................................. 265

10.4 Variation of day-to-day TTV with the period of the day (Thursday through to Saturday). Both ANPR and k-NN results are shown for the Russell Square 2004 dataset. ............................................. 266

10.5 Relationship between the congestion index and the d2d coefficient of variation. Equation 10.18 was fit to the data. The ANPR dataset of Russell Square 2004 was used ............................................. 268

10.6 Relationship between the congestion index and the d2d coefficient of variation. Equation 10.18 was fit to the data. The k-NN dataset of Russell Square 2004 was used ............................................. 269

10.7 Kolmogorov-Smirnov test to determine if the distribution of travel times from the k-NN method is the same as the actual distribution. ............ 276
10.8 Effect of the number of nearest neighbors used on the ability of the k-NN method to estimate the v2v metric. The Russell Square 2004 dataset was used. .................................................. 278
10.9 Relationship between the mean travel time and the v2v standard deviation of travel times. ANPR data from the Russell Square 2004 dataset was used. 283
10.10 Relationship between the mean travel time and the v2v standard deviation of travel times. k-NN data from the Russell Square 2004 dataset was used. 284

11.1 Five-step methodology for turning data from ILDs into estimates of ULTT 292

A.1 Vehicle trajectories and the intersection with a region of time-space. Adapted from Daganzo (1997). ................................................................. 298
A.2 Lighthill, Whitham and Richards, LWR model. K = density, V = velocity, Q = flow. ................................................................. 300
A.3 Delay at a junction at very low flow ......................................................... 302
A.4 Derivation of Webster’s formula for delay in under-saturated traffic ................................. 303
A.5 Derivation of formula for delay in over saturated traffic. R = red time, G = green time. ................................. 304
A.6 Derivation of Kimber and Hollis model ......................................................... 305
A.7 General representation of a queuing system ......................................................... 306

B.1 Effect of a bottleneck on merging adjacent platoons ........................................ 308
B.2 Vehicles which make a mid-link movement are delayed for \( m_t \) seconds when making that movement. ......................................................... 310

C.1 Comparison of the estimates of ULTT using the actual ANPR measurements and the k-NN derived estimates. Data from the Russell Square 2004 dataset, weeks 1 - 4. ......................................................... 312
C.2 Comparison of the estimates of ULTT using the actual ANPR measurements and the k-NN derived estimates. Data from the Russell Square 2004 dataset, weeks 5 - 8. ......................................................... 313
C.3 Comparison of the estimates of ULTT using the actual ANPR measurements and the k-NN derived estimates. Data from the Russell Square 2004 dataset, weeks 9 - 12. ......................................................... 314
C.4 Comparison of the estimates of ULTT using the actual ANPR measurements and the k-NN derived estimates. Data from the Russell Square 2004 dataset, weeks 13 - 14. ......................................................... 315

D.1 cdf of travel times - Russell Square 2003 dataset - see section 4.7 ................................. 317
D.2 Travel Time by time-of-day: Russell Square 2003 dataset ................................. 318
D.3 cdf of travel times - Russell Square 2003 dataset - see section 4.7 ................................. 320
D.4 No of records by time-of-day: - Russell Square 2003 dataset - see section 4.7 ................................. 321
D.5 cdf of travel times - Russell Square 2004 dataset - see section 4.7 ................................. 322
List of Figures

D.6 Travel Time by time-of-day: Russell Square 2004 dataset .......... 323
D.7 cdf of travel times - Russell Square 2004 dataset - see section 4.7 ..... 325
D.8 No of records by time-of-day: - Russell Square 2004 dataset - see section 4.7 326
D.9 cdf of travel times - Tottenham Court Road 2003 dataset - see section 4.7 . 327
D.10 Travel Time by time-of-day: Tottenham Court Road 2003 ............. 328
D.11 cdf of travel times - Tottenham Court Road 2003 dataset - part a ........ 329
D.12 cdf of travel times - Tottenham Court Road 2003 dataset - part b .... 330
D.13 No of records by time-of-day: Tottenham Court Road 2003 dataset .... 331
List of Tables

2.1 Comparison of existing technologies for measuring the actual travel time . . 42
2.2 Other parameter settings in the simulation ................................. 56
2.3 Overtaking rule settings used in the experiment of section 2.11 ............ 57
2.4 Ability of each method to identify outliers in the matched ANPR data, on a single-lane road link. Emergency vehicles and buses filtered from the dataset. 58
2.5 Accuracy of the estimates of $\mu^*$ and $\sigma^*$ from each of the data cleaning methods. Data from single-lane road link. Emergency vehicles and buses filtered from the dataset. ................................................................. 58
2.6 Ability of each method to identify outliers in the matched ANPR data, on a single-lane road link. Emergency vehicles and buses NOT filtered from the dataset. ................................................................. 59
2.7 Accuracy of the estimates of $\mu^*$ and $\sigma^*$ from each of the data cleaning methods. Data from single-lane road link. Emergency vehicles and buses NOT filtered from the dataset. ................................................................. 59
2.8 Accuracy of the estimates of $\mu^*$ and $\sigma^*$ from each of the data cleaning methods. Data from multi-lane road link. Emergency vehicles and buses NOT filtered from the dataset. ................................................................. 61
2.9 Characteristics of the Tottenham Court Road and Russell Square links ................................. 62
2.10 Timetable of ANPR validation survey ........................................... 65
2.11 Validation of the overtaking rule using real world data (Begon 2004) ....... 65
2.12 Two-sample Kolmogorov tests to determine whether the distribution of travel times from the ANPR data differs significantly from the actual distribution of travel times (Begon 2004). ................................................................. 66

3.1 Errors collecting the raw ILD data ................................................ 74
3.2 Errors in calculating the basic traffic parameters ............................ 75
3.3 Quality Control checks currently undertaken on traffic data by TMCs .... 76
3.4 Possible status settings given to ILDs managed by PeMS (Chen 2004) ... 81
3.5 Thresholds used for the DSA test. .................................................. 86
3.6 Schema of database table used to store daily ILD quality control information 87

4.1 DSA performance of the ILDs on the Tottenham Court Road link ........... 101
4.2 DSA performance of the ILDs on the Russell Square link .......................... 101
4.3 Overtaking rule settings used on Tottenham Court Road Link .................. 104

5.1 Desired Properties of an urban link travel time model ............................ 111
5.2 Effect of various traffic and road topology characteristics on the average
journey speed in a study undertaken by Freeman Fox and Associates (1972). 127
5.3 Summary of ULTT models ................................................................. 137
5.4 List of possible inputs to a ULTT model ............................................. 144
5.5 Additional factors that may affect link travel time yet not affect the output
from ILDs ............................................. 148

6.1 Common Kernel Functions (Härdle 1990, p.138) .................................. 161

7.1 Historical data required to use the k-NN method as a ULTT model ......... 174
7.2 The 100 closest historical travel time records to the input feature vector,
sorted by distance ................................................................. 176
7.3 Performance of each distance metric for use in estimating ULTT using the
k-NN method ................................................................. 184
7.4 Best settings when performance is measured by MAPE ......................... 187
7.5 Best settings when performance is measured by RMSE ......................... 187
7.6 Best settings. PI = rank MAPE + rank RMSE + (rank time / 5) ............. 188
7.7 Modelling the RMSE - fitting equation 7.6 to the data ......................... 191
7.8 Identification the optimal value of k to reduce MAPE. LEM = lowess, DM
= SE-VAR ................................................................. 195
7.9 Identification the optimal value of k to reduce RMSE. LEM = lowess, DM
= SE-VAR ................................................................. 196
7.10 Regression model of the MAPE (MN = mean, MD = median, REG =
regression, LOW = Lowess) ............................................... 199
7.11 Regression model of the RMSE (MN = mean, MD = median, REG =
regression, LOW = Lowess) ............................................... 201

8.1 Comparison of 10 ULTT models on the Russell Square Link - 2003 dataset. 215
8.2 Comparison of 10 ULTT models on the Tottenham Court Road Link - 2003
dataset. ................................................................. 222
8.3 Comparison of 10 ULTT models on the Russell Square Link - 2004 dataset. 225
8.4 Number of valid records after filtering by the various methods ............... 226
8.5 MAPE for each combination of ULTT model and ILD data cleaning treatment. 227
8.6 Performance of DSA method against other methods for Gault & Taylor ULTT 228
8.7 Performance of DSA method against other methods for basic regression ULTT 228
8.8 Performance of DSA method against other methods for k-NN ULTT ........ 228
8.9 Performance of DSA method against other methods for ANN ULTT ........ 230
8.10 RMSE for each combination of ULTT model and ILD filtering method. .... 233
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>Travel Time Factors affecting the demand driver</td>
<td>238</td>
</tr>
<tr>
<td>9.2</td>
<td>Travel Time Factors affecting the capacity driver</td>
<td>239</td>
</tr>
<tr>
<td>9.3</td>
<td>Travel Time Factors affecting the vehicle behaviour driver</td>
<td>240</td>
</tr>
<tr>
<td>9.4</td>
<td>Travel Time Factors affecting the control driver</td>
<td>241</td>
</tr>
<tr>
<td>9.5</td>
<td>Disaggregation of TTV by time period. Bold terms in brackets are the</td>
<td>244</td>
</tr>
<tr>
<td></td>
<td>nomenclature used in the rest of this thesis</td>
<td></td>
</tr>
<tr>
<td>9.6</td>
<td>TTV sources using Nested ANOVA (adapted from Park et al. (1999))</td>
<td>250</td>
</tr>
<tr>
<td>10.1</td>
<td>Sources of TTV by temporal scale. ANPR data used</td>
<td>259</td>
</tr>
<tr>
<td>10.2</td>
<td>Sources of TTV by temporal scale. k-NN estimates used</td>
<td>260</td>
</tr>
<tr>
<td>10.3</td>
<td>Correlation between mean travel times from different days of the week.</td>
<td>264</td>
</tr>
<tr>
<td></td>
<td>ANPR data from the Russell Square 2004 dataset was used</td>
<td></td>
</tr>
<tr>
<td>10.4</td>
<td>Correlation between mean travel times from different days of the week.</td>
<td>264</td>
</tr>
<tr>
<td></td>
<td>k-NN data from the Russell Square 2004 dataset was used</td>
<td></td>
</tr>
<tr>
<td>10.5</td>
<td>Relationship between the congestion index and the d2d coefficient of variation</td>
<td>267</td>
</tr>
<tr>
<td>10.6</td>
<td>Estimators used to calculate the v2v metrics</td>
<td>271</td>
</tr>
<tr>
<td>10.7</td>
<td>Effect of the sample size on the MAPE of the estimates of the characteristics of the vehicle-to-vehicle travel time distribution. The ANPR data from the Russell Square 2004 dataset was used</td>
<td>272</td>
</tr>
<tr>
<td>10.8</td>
<td>Effect of the sample size on the MPE of the estimates of the characteristics of the vehicle-to-vehicle travel time distribution. The ANPR data from the Russell Square 2004 dataset was used</td>
<td>273</td>
</tr>
<tr>
<td>10.9</td>
<td>Effect of the number of nearest neighbours chosen on the ability of the k-NN model to estimate the characteristics of the vehicle-to-vehicle distribution. Accuracy measured by MAPE. The Russell Square 2004 dataset was used</td>
<td>279</td>
</tr>
<tr>
<td>10.10</td>
<td>Effect of the number of nearest neighbors chosen on the ability of the k-NN model to estimate the characteristics of the vehicle-to-vehicle distribution. Accuracy measured by MPE. The Russell Square 2004 dataset was used</td>
<td>280</td>
</tr>
<tr>
<td>D.1</td>
<td>No records for each day-of-week, Russell Square 2003 dataset</td>
<td>318</td>
</tr>
<tr>
<td>D.2</td>
<td>No records for each day-of-week, Russell Square 2004 dataset</td>
<td>323</td>
</tr>
<tr>
<td>D.3</td>
<td>No records for each day-of-week, Tottenham Court Road 2003 dataset</td>
<td>328</td>
</tr>
<tr>
<td>E.1</td>
<td>List of acronyms</td>
<td>333</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Background

1.1.1 Applications of travel time data

In recent years there has been increased interest amongst Transport Engineers to measure or estimate the travel time of vehicles over a small section of road-network or over a whole journey. Travel time is of interest to both the individual traveller and to the network authorities who manage the road network.

For most travellers, travel-time is a significant cost to making a journey (Noland & Small 1995, Noland & Polak 2002) and hence travellers will aim to reduce their overall cost by reducing the travel time of their journey.

This desire of travellers to reduce their travel time over a road network has led to a necessity for transport authorities to reduce the overall travel-time and amount of congestion on their road network. A common metric to quantify congestion is to subtract the free-flow travel time (i.e. when there is no congestion) from the current measure of travel time (Lomax et al. 1997). There are modifications which normalise this metric, such as the use of the travel-rate (TfL 2004). These methods of quantifying congestion all rely on the ability to measure or estimate the travel time of vehicles on a road.

In addition to travel time, the variability of travel time is also an important phenomenon to understand and model. This variability in travel times results in travellers allowing additional time for their journey in case they are delayed on that journey. This additional time is commonly called the safety margin (Turner et al. 1998). It was referred to as far back as 1968, when Thomson (1968) noted that although the traveller often leaves a safety margin, ‘on the majority of occasions he will arrive early and his time will often be wasted just as much as if he were sitting in a traffic jam.’ Such travel time variability is not only wasteful for individual car drivers, but also for other users of the road network such as bus passengers and operators. Thomson (1968) continues to remark that ‘there can be no doubt that in London the unpredictability of journey times is an important source of time losses. In addition it plays such havoc with bus schedules that it is one of the greatest problems affecting the management of London bus services.’
Three important uses of travel time data have thus been identified. It is thus important to be able to measure or estimate the travel time of vehicles on the road network. Indeed the modeling of travel time variability requires extensive measurements of vehicle travel times.

### 1.1.2 Measurement of Travel Time

There are various technologies available to measure the travel time of a vehicle directly. These include Moving Car Observers (MCO), whereby a passenger will record the time when a vehicle passes certain points on the road. The travel time between those points can then be determined. One criticism of this technique is that it is expensive and hence costly to obtain multiple measurements in any one time period (say 15-minute period). The limitation of a small sample size of travel times will affect the accuracy of the estimation of the typical travel time for vehicles in that time period. This is because the travel time of vehicles varies, even in a short time period. This effect is more pronounced in urban areas than on highways. In urban areas there are such features as traffic lights, parked vehicles, and turning vehicles which delay some vehicles but not others, causing the travel time between vehicles to vary, even in a very short period of time. Travel time in a short time period is thus a random variable. An MCO will provide only one travel time record, and hence only one realisation of the random variable. This one realisation of the random variable may not be representative of other vehicle travel times in the same time period, and it is thus imperative to measure the travel time of multiple vehicles in the same time-period to obtain an accurate estimate of the typical travel time in the period.

More recently technologies such as Global Positioning System (GPS) devices have allowed the automation of the MCO survey vehicle. Survey vehicles based upon GPS are able to measure the travel time of the survey vehicle over most links very accurately (Zhao 1997). However, most vehicles fitted with commercial GPS devices measure their location at poor temporal resolutions, resulting in an error in measuring the travel time over the desired link. Travel time measurement using GPS also suffers from low sample sizes.

Automatic Number Plate Recognition (ANPR) systems read the licence plate at one location, and then again at a second location. By matching those license plates read at both camera sites, the travel time between those camera sites can be measured. However, ANPR systems are expensive to install, and are limited to links between camera locations.

Both GPS and ANPR travel time measurement technologies also have issues of privacy concerning them (Faithful 2002).

### 1.1.3 Estimation of Travel Time

The previous section has identified problems with measuring travel time in any time period directly. An alternative approach is to estimate the travel time indirectly, including through the use of data derived from Inductive Loop Detectors (ILDs). ILDs are wire loops laid in the road, which measure flow and occupancy. One characteristic of ILDs
is that they measure the flow and occupancy from the whole vehicle population - i.e. they are able to count all vehicles passing over them. However, ILDs output only basic traffic parameter data such as flow and occupancy at a single point in space, and not the travel time of vehicles over the link. Nevertheless, there are many ILDs currently installed in cities in the UK, as they provide key input to Urban Traffic Control systems such as SCOOT. In London for example, there are approximately 5270 SCOOT ILDs located throughout the city. The challenge thus becomes that of merging high-quantity low quality ILD data, with low-quantity high quality travel-time measurement data, to obtain high-quality high-quantity travel-time estimates.

Various methodologies have been proposed to convert data from ILDs into estimates of link travel time. These include models based upon spot-speed measurements (Turner et al. 1998; Dailey 1999), correlation techniques (Dailey 1993; Petty et al. 1998), queuing theory (Hunt et al. 1981; Takaba et al. 1991), regression analysis (Gault & Taylor 1981; Zhang & He 1998), and Artificial Neural Networks (Cherrett et al. 2001). One limitation of the existing literature is a failure to compare different ULTT models against one another using the same dataset. It has thus been difficult to explicitly identify the most appropriate ULTT model to use.

1.1.4 Cleaning of travel time and ILD data

An important issue concerning ULTT models, and one which has tended to be overlooked is the importance of data-quality. Research by Turochy & Smith (2000a) has suggested that the failure to screen raw ILD data could explain the poor performance of many ILD based incident detection systems. Various ILD treatments have been proposed. These primarily concern univariate tests which for example check that flow and occupancy fall within certain ranges (Chen & May 1987). Other data cleaning methods use multivariate tests, ensuring for example that the relationship between the measured values of flow, occupancy, and speed is reasonable (Turochy & Smith 2000a). More elaborate tests such as the Daily Statistics Algorithm (Chen et al. 2003) have also been proposed. One common limitation of much of the existing literature is its failure to quantify the performance of the ILD cleaning treatments.

The literature to clean travel time data is smaller than that of the literature of cleaning ILD data. Travel time cleaning methods revolve primarily around finding outlier records in a certain period of time (Clark et al. 2002). As such they ignore the temporal structure of the data (i.e. the order and sequence of travel-times). Similarly there tends to be a failure in the literature to quantify the performance of the travel-time cleaning methods.

1.2 Aims and Objectives

Section 1.1 has provided the background to this PhD. There are two aims within this PhD. The primary aim is to develop a methodology by which data from ILDs can be
used to estimate the travel time on an urban link. This methodology will include the identification of suitable data cleaning methods for both ILD and travel time data, and the development of an Urban Link Travel Time (ULTT) model. The secondary aim of this PhD is to determine how effectively travel-time estimates derived using this methodology can be used to model travel-time variability.

This thesis has been divided into three primary parts:

- **Part A - Data**: This part looks at the data requirements of any urban link travel time (ULTT) model. In particular there are two sources of data that are needed to successfully build a ULTT model. Firstly an accurate measure of actual urban link travel time is required. Such data is needed to calibrate and train any ULTT model. Secondly data derived from ILDs such as flow and occupancy is needed as input to the ULTT model. This part describes how this data is both obtained, and subsequently treated to filter out erroneous data.

- **Part B - Model**: This part develops a ULTT model that can be used to estimate the link travel time given input from inductive loop detectors.

- **Part C - Application**: This section demonstrates the importance and relevance of developing ULTT models. As an application of the model, the variability of link travel time is studied.

The objectives for each of these parts is listed below:

### 1.2.1 Objectives of part A - Data for a ULTT model

There are 4 primary tasks in part A of this thesis:

- Define clearly the concept of an urban link and the travel time on an urban link.
- Develop a methodology to clean matched ANPR data.
- Develop a methodology to clean data from ILDs.
- Create an archive to store ILD data obtained from the London SCOOT system. This archive is called the London SCOOT Archive Database, LSAD.

### 1.2.2 Objectives of part B - Modelling ULTT

There are 4 primary tasks in part B of this thesis:

- Undertake a comprehensive literature review and critique of ULTT models. Identify the limitations of these models and suggest modifications and alternative models that could be used to better model travel time variability.
1.3. Thesis Outline

- Propose a methodology based on the k-Nearest Neighbours approach for calculating travel-time through a link using data from single loop ILDs. Provide guidelines for selecting the key design parameters of the k-NN method.

- Propose a set of criteria and performance indicators by which ULTT models can be compared to one another.

- Perform an experiment to compare various ULTT models.

1.2.3 Objectives of part C - Application of a ULTT model in study on TTV

There are 3 primary tasks in part C of this thesis:

- Undertake a comprehensive literature review into TTV.

- Identify ways to characterise and quantify TTV.

- Determine whether travel time estimates obtained from the ULTT model developed on part B allows the travel time variability to be characterised accurately.

This thesis will describe how each of these 11 tasks have been successfully completed.

1.3 Thesis Outline

To provide the reader a ‘road-map’ of the thesis, a brief summary of the main objectives and results of each chapter is given below:

Chapter 1 provides the background and context to the PhD. It also presents the aims and objectives of the research.

1.3.1 Part A

Chapter 2 discusses how the link travel time can be measured. It firstly defines what a link is and what link travel time is. It identifies various technologies which can be used to measure link travel time, and elicits the problems associated with each of these technologies. A review of existing travel time data cleaning methods is presented, and a new method, the overtaking rule which uses the temporal-structure of the data, is presented. An experiment is then undertaken to quantify the performance of each travel time cleaning method. The final part of the chapter describes the ANPR data used within this PhD.

Chapter 3 reviews data from Inductive Loop Detectors (ILDs). The errors associated with ILDs are listed. Various ILD data cleaning treatments are reviewed and a new one called the adjacent detectors test proposed. An experiment to quantify the performance
of the ILD data cleaning treatments is then proposed.

Chapter 4 describes the ILD data used in this thesis. It describes the data archive, LSAD, that has been created as part of this PhD. It then describes the three main datasets that will be used throughout the rest of the thesis.

1.3.2 Part B

Chapter 5 provides a comprehensive review of the Link Travel Time (LTT) literature. The desired properties of a ULTT model are identified, and a review and summary of existing models provided. Specific areas of modelling ULTT that are often not incorporated in existing models are identified. This includes the mid-link delay and specific turning-movement delay. Finally the limitations of using solely ILD derived data is considered.

Chapter 6 introduces the k-Nearest Neighbors method. This is a non-parametric estimation method. The theoretical basis of this method is outlined, and the assumptions upon which this method relies are stated. The key design parameters are elicited, and a way to overcome the computational requirements of the method proposed. The existing use of the k-NN method in the transport literature is then reviewed.

Chapter 7 builds upon chapter 6 by identifying a methodology by which the k-NN method can be optimised for use as a ULTT model. The effect of the size of the underlying database is then studied.

Chapter 8 undertakes a series of experiments to compare the performance of various ULTT models with the three datasets. It identifies the best ULTT model for use on the London road network. An experiment is also carried out which quantifies the performance of the various ILD data cleaning treatments.

1.3.3 Part C

Chapter 9 provides a comprehensive introduction to the topic of travel time variability (TTV). It identifies how TTV can be disaggregated and quantified. it provides a review of the empirical research into TTV.

Chapter 10 examines how effective the k-NN method is at modelling TTV. In particular the ability of the k-NN method to model day-to-day and vehicle-to-vehicle TTV is studied.

Chapter 11 presents the conclusions of the research undertaken in this thesis. It also presents recommendations for future work.
Part A - Data

The first part of this thesis looks at the data used within the research. Chapter 2 outlines the technologies that can be used to collect the urban link travel time of individual vehicles. It discusses problems with collecting such data, and proposes a filter that can be used to improve the quality of such data. This chapter closes by introducing the travel time data that will be used in the rest of the research. Chapter 3 provides a comprehensive introduction to inductive loop detectors. It briefly outlines the data they produce and identifies various problems with such data. Various treatments to improve the quality of inductive loop detector data are proposed. Chapter 4 outlines the datasets that will be used in parts B and C of this thesis.
Chapter 2

Measuring Actual Link Travel Time using ANPR data

The primary goal of the PhD is to design a model to estimate link travel time using data from ILDs. In order to calibrate and validate any urban link travel time (ULTT) model, it is first necessary to measure the actual travel time over the link. This task is more complex than it may first appear. This section begins by first explicitly defining what a link is and proposes a nomenclature for defining different types of links. The next section then defines what link travel time is, outlining all the errors associated with its measurement. This is followed by a brief review of various sources that could provide the actual travel time. For reasons explained in section 2.4.6, it is chosen to use Automatic Number Plate Recognition (ANPR) data. The rest of the chapter looks specifically at the problems of using ANPR data. It reviews the existing literature used to clean matched license plate data, and then proposes a new cleaning method called the Overtaking Rule. The performance of the various cleaning methods are then compared to one another using data from a microsimulation. The chapter concludes by stating the further work needed to enhance the use of ANPR data for the estimation of actual link travel time. ANPR travel time data cleaned using the Overtaking Rule will be used for calibration and validation of the ULTT models in chapters 7 and 8.

2.1 Definition of a link

The primary objective of this research is to develop a methodology for estimating ULTT. This section clarifies what an urban link is. It firstly identifies various types of link found in urban networks. It then proposes a general framework for describing and classifying links. Finally the definition of an urban link within this thesis is described.

2.1.1 Urban Links

A link denotes a stretch of road between two clearly identifiable points. Within urban road networks, there are many possible types of link. Several examples are given below:
• **road** - a road is an urban link, commonly demarcated between a single dashed start line, and an end stop line (either a double dash or a continuous line drawn in the road).

• **turning movement** - a turning movement can be described as a link between the stop-line of one road and the start line of another road.

• **SCOOT link** - A SCOOT link describes the section of road between an inductive loop detector and the stop line at an intersection (see section 4.1.1)

• **bus stop to bus stop** - the stretch of road between two bus-stops can be considered as a link on the bus route.

These different definitions of links disaggregate the network in different ways, and may not allow the network to be completely covered. For example, since a SCOOT link only describes the stretch of road between the detector and stop-line, the stretch of road between the start of the road and detector will not be included on a SCOOT link map of the network. Ideally a consistent and generic way is needed to characterise these links. This is considered in the next section.

### 2.1.2 Proposed nomenclature

This section proposes a nomenclature for classifying each type of link described in the previous section. The start and end of a link will often be demarcated by one of the following features:

• **Start of a road** - immediately after an intersection, and often demarcated by a single dashed line.

• **End of a road** - immediately before an intersection, and often demarcated by a double dashed line or single continuous line.

• **A detector at a point of the road.** This may be an ILD or a camera.

However, occasionally the start or end of a link is simply an arbitrary position on a road between two junctions.

A nomenclature to describe what type of link is being used is proposed. The link type can be described by two letters. The first letter describes how the start of the link is defined. The second link describes how the end of the link is defined. This methodology will be used within this thesis to describe what type of link is being studied. The following codes are used:

• **S** = Start of a road

• **E** = End of a road
2.2. Definition of link travel time

The previous section has identified that the type of link being modelled within this thesis is a D-D link. This section clarifies what is meant by the term ‘link travel time’.

2.2.1 Definition of link travel time: literature review

Surprisingly, much of the existing literature on modelling link travel time fails to explicitly define what link travel time is. If it is considered, it is often given scant attention. As an example, Quiroga (1998) states that the representative travel time can be given by the arithmetic mean of all individual vehicle travel times recorded. He correctly notes that the use of the arithmetic mean is sensitive to outlying high travel times from individual vehicles and thus proposes the use of the median value of all individual vehicle travel times. Although the use of a robust method is advised, Quiroga (1998) fails to identify the sources of error that could lead to erroneous measures of travel time. Importantly he
fails to consider in detail what subset of all vehicle travel times should be used. This is particularly important in the urban environment. Fowkes (1983) alludes to this problem:

"It will be assumed here that we are trying to determine the distribution of travel times between points A and B, of vehicles of a particular type (e.g. car) passing point A at a particular time and having no ‘goal’ prior to reaching point B. By this definition it is wished to exclude vehicles which stop en route between A and B, or take a meandering route for purposes other than to reach point B more quickly."

The next section explicitly identifies various categories of travel time records and suggests the appropriate category of travel time records that should be used in the thesis.

### 2.2.2 Categories of link travel time distributions

The problem of using travel time records from individual vehicles to determine the true value of the average link travel time from point A to point B, is thus in filtering out erroneous records. In effect each vehicle travel time record can be classified as belonging to a certain category of link travel time, according to what characteristics are associated with that vehicle record. Each category of travel times has its own distribution. Vehicles can be classified into one of the following five categories, each with its own travel time distribution:

a. Vehicles which took an alternative route from A to B.

b. Vehicles which stopped en-route (e.g. to drop off a passenger, allow the driver to pop into a local shop, fill-up with petrol, etc).

c. Vehicles which are not restricted to normal traffic regulations - e.g. emergency vehicles are able to travel over the speed-limit, travel on the ‘wrong’ side of the road, and travel through red-lights at intersections. Buses may have a lane reserved for their sole use.

d. Vehicles which did not attempt to travel from A to B in a prompt fashion, but rather travelled in an irrational manner - e.g. did not attempt to move into a faster moving lane.

e. Vehicles which did attempt to travel from A to B in a prompt fashion.

The problem that the practitioner faces is in selecting travel time records that come from vehicles in the correct category (i.e. category e). This distribution shall be referred to in this thesis as the *actual link travel time distribution*. If travel time records from other distributions are not filtered out, then this is likely to lead to an error in the estimation of any characteristic of the actual link travel time distribution.

Perhaps the hardest criterion to determine is whether a vehicle was driven irrationally. For example, a road may contain 3 lanes. Two of these lanes may be moving significantly faster than a third lane. It is assumed that any rational driver would attempt to force
themselves into the faster lane. However, this is very subjective and very difficult to measure.

### 2.2.3 Definition of link travel time

Following the discussion in the previous section, the link travel time in a time interval $\tau$, can be defined as the arithmetic mean travel time of the population of all valid vehicles that pass the entrance of the link during the time interval $\tau$. Within this thesis the true value of the average link travel time of the valid vehicles will be defined by equation 2.1.

$$\mu^* = \frac{1}{N} \sum_{i=1}^{N} TT_i$$  \hspace{1cm} (2.1)

$\mu^*$ = True value of the average link travel time of the valid vehicles.

$N$ = Total number of valid vehicles passing over point A in the time period $\tau$.

$TT_i$ = Travel time of valid vehicle $i$.  

A valid vehicle belongs to category $e$ outlined in section 2.2.2.

In addition the population standard deviation of travel time of all valid vehicles ($\sigma^*$), is a measure of how much variability there is within the time interval. Within this thesis it will be defined by equation 2.2.

$$\sigma^* = \left[ \frac{1}{N} \sum_{i=1}^{N} (TT_i - \mu^*)^2 \right]^\frac{1}{2}$$  \hspace{1cm} (2.2)

These metrics will be used in section 2.11 to assess the performance of ANPR data cleaning treatments.

### 2.3 Problems with measuring link travel time

There are various problems in measuring characteristics of the actual link travel time distribution, such as those given by equations 2.1 and 2.2. This section outlines these.

Section 2.2.2 indicated that the set of all travel times from vehicles passing between two points A and B, comes from a variety of different travel time distributions. Each of these travel time distributions can be modelled as a random variable. For example, let $f_a(x, \theta_a)$ denote the travel time distribution coming from vehicles which took an alternative route from A to B. $\theta_a$ denotes a set of dependent variables. Similarly let $f_e(x, \theta_e)$ denote the travel time distribution of vehicles which did attempt to travel from A to B in a prompt fashion, where $\theta_e$ denotes a set of dependent variables. In a similar fashion the travel time distributions for categories b, c, and d can be given by $f_b(x, \theta_b)$, $f_c(x, \theta_c)$, and $f_d(x, \theta_d)$.

If $g(x, \theta)$ denotes the travel time distribution of all travel times from vehicles passing between two points A and B, then the following relationship between the random variable
2.3. Problems with measuring link travel time

\( g(x, \theta) \) and the other random variables exists. This is shown in equation 2.3.

\[
g(x, \theta) = f_a(x, \theta_a) + f_b(x, \theta_b) + f_c(x, \theta_c) + f_d(x, \theta_d) + f_e(x, \theta_e)
\]  

(2.3)

The travel time of an individual vehicle measured between A and B is a realisation of the random variable \( g(x, \theta) \), \( X_g^* \). Associated with this realisation is a measurement error such that the actual time the vehicle took to travel between A and B, \( X_g^* \), may differ from that measured, \( X_g^{**} \). This is shown in equation 2.4.

\[
X_g^{**} = X_g^* + e
\]

(2.4)

\( e \) = term characterising the measurement error.

The aim of this exercise is to use a sample of measurements, \( X_g^{**} \), to estimate the travel time distribution of valid vehicles, \( f_e(x, \theta_e) \). However, since realisation \( X_g^* \) will have come from one of the travel time distributions a to e, it is necessary to filter all travel time records so that only those from realisations which originated from travel time distribution e are used to estimate \( f_e(x, \theta_e) \). From the above discussion, it is clear that there are three sources of error:

- Measurement Error - Time the vehicles entered and exited the link.
- Excluding invalid travel time records.
- Sample size.

The last two errors are sampling errors. The first error is a non-sampling error. The errors are elaborated in further detail below.

### 2.3.1 Time the vehicles entered and exited the link

The first requirement for measuring \( \mu^* \) is to know the exact times at which all valid vehicles entered and exited the link. The link travel time will then be calculated from these two values. These data needs necessitate the following requirements:

- The recording device should operate at a good temporal resolution so that there is only a small time lag between the vehicle passing the link boundary and the detector recording this event.
- The clock measuring the time the vehicle entered the link is synchronised with the clock measuring the time the vehicle exited the link.
- It is possible to identify the position of the vehicle on the road accurately.
2.3. Problems with measuring link travel time

The distribution of travel times in any short interval can thus be described by a random variable.

2.3.2 Excluding invalid travel time records

The second requirement for measuring $\mu^*$ is to identify all travel time records that came from valid vehicles. A valid vehicle belongs to the travel time distribution $\mu$ outlined in section 2.2.2, i.e. vehicles that travelled in a prompt fashion over the link.

2.3.3 Sample size

Another important criterion in measuring ULTT is that a sufficient and representative number of travel time records from valid vehicles are included. Ideally all travel time records should be recorded. Link travel time is likely to vary from one valid vehicle to another valid vehicle, even over a very small time period. This is due to such factors as traffic lights, buses stopping, vehicles turning mid-link delaying following vehicles, etc. This is illustrated by figure 2.3 which shows distance-velocity (x-v) plots for a variety of vehicles in a short time interval over a typical urban link. It then shows the expected pdf of link travel time of a valid vehicle in this short time interval.

For example, vehicles which are not delayed mid-link (path 1 on the left hand diagram) and are not delayed by the traffic lights (path 2 on the left hand diagram) will have a low travel time, of around $t_1$; This is mode 1. Vehicles which are delayed mid-link (path 3 on the left hand diagram) and delayed by the traffic lights (path 4 on the left hand diagram) will have the longest travel time, greater than time $t_2$.

The important thing to note is that due to random factors, link travel time within a short time period (say 15 minutes) can be thought of as a random variable. A travel time record obtained from a vehicle is simply one realisation of this random variable. It is thus important to take several samples from this random variable in order to estimate characteristics of this distribution.

Srinivasan & Jovanis (1996) formulated an expression for predicting the number of
2.3. Problems with measuring link travel time

probe vehicles, \( n \), required in any time period to allow for an estimate of travel time known to be accurate to a stated level. This formulation does assume that travel times in a short period of time (say 15 minutes) are distributed normally.

\[
n \geq \left( \frac{\phi^{-1} \cdot \left( \frac{1+r}{2} \right)}{\varepsilon_{\text{max}} \cdot \left( \frac{\mu}{\sigma} \right)} \right)^2 \tag{2.5}
\]

\( \phi \) = Cumulative distribution function of the \( N(0, 1) \) distribution
\( r \) = minimum probability that the absolute error is less than \( \varepsilon_{\text{max}} \).
\( \varepsilon_{\text{max}} \) = Maximum allowable error
\( \mu \) = mean travel time of all vehicles in time period.
\( \sigma \) = standard deviation of travel time of all vehicles in time period.

Importantly it shows that as the amount of variability in the travel time distribution increases (i.e. the standard deviation increases), so the number of travel time records required to estimate the link travel time to a certain level of accuracy also increases.

The above formula assumes that the \( n \) records chosen to estimate \( \mu^* \) will be chosen randomly. If this is not the case, then the estimate of \( \mu^* \) may be biased. Of course, if all travel time records from valid vehicles are measured then \( \mu^* \) can be calculated directly, rather than having to be estimated.

2.3.4 Summary of data requirements to measure link travel time

From the above three sections, the following requirements of any data source that will be used to measure \( \mu^* \) can be stipulated:

- Ability to measure accurately and precisely the time each vehicle entered and exited the link.
- Ability to ensure that any travel time record came from a valid vehicle - i.e. a vehicle which travelled in a prompt fashion over the link (i.e. belongs to travel time distribution outlined in section 2.2.2).
- Ability to measure the travel time of a sufficient and representative number (ideally all) of valid vehicles.

2.3.5 Estimating \( \mu^* \) with travel time data.

Assuming that it is not possible to collect data that meets all the criteria laid out above, a way of estimating the population mean travel time of valid vehicles (\( \mu^* \) as defined in equation 2.1) needs to be defined using the travel time records that have been obtained. Two obvious estimators would be the mean and median of a sample of travel time records collected. The performance of these estimators is considered in greater detail in section 2.10.
2.4 Data Collection Methods

This section gives a brief review of various sources of travel time data. Their ability to meet the various criteria laid out in section 2.3.4 will be considered.

2.4.1 Moving Car Observer

The oldest method of collecting travel time data is known as the moving car observer (MCO) technique, or floating vehicle (Taylor et al. 2000). Essentially a vehicle is driven around a set route, and a passenger writes down the time when the vehicle passes certain points. This method thus allows the times when the vehicle entered and exited each link to be determined within an accuracy of a couple of seconds, and allows the passenger to verify that the vehicle travelled ‘reasonably from A to B.’

The main errors associated with the MCO technique are sampling errors. Perhaps the biggest problem with using floating vehicle data to estimate the actual link travel time concerns the number of travel times samples obtained. As stated in section 2.3.3 due to the inherent variability of the system, travel times from several floating vehicles are required in each time interval in order to calculate an estimate of travel time to a certain statistical confidence. Since floating vehicles are very expensive to operate, there will only be limited data from such vehicles.

An additional sampling error with the use of MCO travel time records has been raised by Hellinga & Fu (1999). They noted that due to traffic signals at the downstream section of the link, the travel time along a link would in part depend on the time the vehicle entered the link. If traffic signals in a region of the urban network have a common cycle time, then vehicles entering the link from a certain approach will tend to arrive at the downstream traffic lights at the same stage (i.e. red lights or green lights). Thus if only travel time records from vehicles which come from the same approach are used, then the estimate of ULTT is likely to be biased. In many MCO surveys, the same route tends to be followed. Thus MCO surveys are likely to give biased estimates of travel time. Further work by Hellinga & Fu (2002), showed that in order to obtain unbiased results from MCO surveys, travel time records from MCO vehicles entering the link from all approaches, weighted by the total number of vehicles which enter the link from that approach, should be used.

2.4.2 GPS

Global Positioning System, GPS, uses trilateration from at least 4 satellites to help locate the position of a user on earth. It is described in detail in Zhao (1997). In the past 10 years much research has been undertaken by the transport community into using GPS derived data for travel time estimation. A comprehensive methodology for using GPS in this field is given by Quiroga & Bullock (1998). Most research has concentrated on the application of GPS on highways, although some work has been carried out on its application in urban
areas (Srinivasan & Jovanis 1996; and Ochieng & Sauer 2002).

GPS systems are theoretically capable of accurately providing travel time records from individual vehicles meeting the 3 criteria stated in section 2.3.4. However, the accuracy of the GPS information depends on 3 aspects; GPS hardware used, analysis software, and the position of the satellites relative to the target vehicles.

Concerning the first requirement, ability to measure the exact time that the vehicle entered and exited the link, GPS systems do not output their position on a road, but instead output their estimated longitude and latitude. It is thus necessary to use these position coordinates to estimate where this vehicle is on the road. This process is called map-matching. One of the most successful map-matching algorithms has been proposed by Quddus et al. (2003). However, like other map-matching methods, its approach is dependent on obtaining position coordinates at a frequency of 1Hz (i.e. once a second). In reality only survey vehicles (i.e. automated MCO survey vehicles) will capture data at this frequency. Due to data transmission costs and products that utilise GPS not needing such high temporal resolutions, commercial GPS solutions rarely record the position of the vehicle at temporal resolutions better than 30 seconds. To the author’s knowledge, little work has been undertaken at map-matching such low-temporal resolution data to the road network. This lack of temporal resolution would thus make it difficult to determine the precise times the vehicle entered and exited the link. This fact is overlooked by the vast majority of literature that proposes to use GPS data to obtain travel time data.

Another problem of using GPS in the urban environment is the ‘urban canyon effect.’ A GPS system needs to have line-of-sight with at least 4 satellites to estimate its location. In urban areas with tall buildings, overhanging trees, and tunnels, this requirement may not be met and the GPS system will not be able to estimate its position. For central London this effect was seen to exist on 9% of all roads (Ochieng et al. 2003). This problem will be specific to certain locations of the urban network, and thus it will not be possible to obtain link travel time estimates for these areas. This will put constraints on how the urban network is divided up into links. Nevertheless, the use of GPS probe vehicles to measure ULTT is spatially robust, since they can be used to record travel time for any feasible geographical link, provided of course that the start and end of those links do not fall within an urban canyon. This is one major advantage GPS has over ANPR.

GPS systems also suffer from sampling errors. Firstly the low temporal resolution of the data provided by many commercial GPS systems may make it difficult to determine whether a vehicle has ‘not reasonably’ travelled between A and B.

Perhaps the biggest problem with using GPS data to estimate the true link travel time concerns the sample of travel times obtained. As stated in equation 2.5, due to the inherent variability of the system, travel times from several probe vehicles are required in each temporal period in order to calculate an estimate of travel time to a certain statistical confidence. Since current GPS vehicle penetration is low, there are unlikely to be sufficient GPS probe vehicles to enable accurate estimates of travel time to be made. This problem is accentuated by the fact that those vehicles which do have GPS systems installed in them
may not be representative of the total vehicle population. Private communication with GPS vendors has indicated that lorries tend to be over represented within any sample of GPS travel time records. Since lorries tend to be slower than private automobiles, any estimate of link travel time is likely to be an overestimate.

### 2.4.3 ANPR

Automatic Number Plate Recognition systems (ANPR) have also been widely used in the literature to obtain estimates of link travel time. ANPR systems rely on cameras mounted above the road recording a picture of the license plate of all passing vehicles and then using software to identify the license number. By matching licence plates recorded at 2 camera sites, the travel time of vehicles between those 2 camera sites can be calculated. This method is an extension of manual license plate surveys. Examples of studies which obtained the actual link travel time information from manual license plate surveys include those undertaken by Cherrett et al. (1995) and Anderson & Bell (1997). For detailed information about ANPR refer to Klein (2001).

If ANPR data is used to estimate the link travel time, then the start and end of links must be defined at camera location sites. However this restriction helps to eliminate the uncertainty of the time when a vehicle actually entered and exited the link. This latter point is the first major advantage that ANPR has over GPS.

Concerning the second data requirement listed in section 2.3.4, unless there are cameras located on the link itself, it is not possible to explicitly check that a vehicle has ‘reasonably travelled’ between A and B without making stops en-route or been delayed due to mechanical failure. In fact there may be ambiguity over the route taken by the vehicle between the 2 ANPR cameras. This is more likely to be a problem in the urban context. Little work has been done in this area with the notable exception of van der Zijpp (1997). Nevertheless, it may be possible to infer whether a vehicle has travelled ‘unreasonably’ from A to B by comparing the travel time of each vehicle with those immediately ahead of and behind it at camera A. A data cleaning method based on this is proposed in section 2.7.

The second major advantage that ANPR has over GPS concerns reduced sampling errors. There are two characteristics of sampling error. The recognition rate describes the ability of a single camera to accurately read the license plate of all vehicles passing on the road beneath it. The matching rate describes the ability to accurately read the license plate of all vehicles passing on the road beneath both camera sites, and then to match all the vehicles successfully. ANPR is able to capture a far higher proportion of vehicle travel times from A to B than GPS based systems. Typically ANPR has a recognition rate of 50-90% of all vehicles at each camera location. (van der Zijpp 1997). Wiggins (1999) reports a recognition rate of 86% during a 2 hour survey in the UK. Since the probability of a successful reading of a number plate depends primarily on the vehicle characteristics, ANPR system used, quality of installation, and weather condition, then a vehicle that has
its license plate successfully read at an upstream point will most probably be successfully
detected at a downstream point, and the matching rate is likely to be very similar to the
recognition rate.

However, the recognition rate is likely to be lower in an urban environment since the
separation between vehicles is likely to be lower, and hence the license plate of vehicles
may be obscured by larger vehicles such as double-decker buses. The latter phenomenon
is known as occlusion. Occlusion is not dependent on the vehicle characteristics, and thus
a vehicle occluded at one camera site will not necessarily be occluded at another camera
site. This will result in the matching rate in the urban environment being less than the
recognition rate. Further research is needed to verify the recognition and matching rate
of ANPR in an urban environment. Nevertheless it is reasonable to assume that ANPR
technology is able to capture a far more representative sample than GPS probe vehicles.
Sources of error from ANPR data are discussed further in section 2.5.

2.4.4 Other Technologies

There are other technologies that can be used to measure the location and /or travel time
of a vehicle. These are briefly reviewed below.

Various technologies have been proposed to continuously record an image of the whole
link and then analyse these images to determine various traffic parameters including travel
time. Angel et al. (2002) used images taken from a helicopter to derive travel times of a
platoon of vehicles. He presented a 10-stage framework showing how this can be completed.
Although only limited data was available, the results suggest that such images could be
used to provide accurate measures of travel time. Manned flying vehicles are however
subject to many aviation rules, and are susceptible to cloud cover and poor weather. To
overcome the issue of cloud cover, the use of unmanned aerial vehicles (UAV), which are
permitted to fly lower than manned vehicles and hence can fly below clouds, has been
studied by Coifman & Yang (2004). They reported that they were unable to identify each
individual vehicle, so if the vehicle left the field of view then they would not be able to
relocate the vehicle. They suggested that higher resolution cameras may in future allow
characteristics on each individual vehicle to be identified.

Other work has used remote sensing; using images taken from satellites to measure
traffic parameters, notably annual average daily traffic (AADT). Satellite images are ca-
pable of producing images with a resolution of 1m (Coifman et al. 2004). To this author’s
knowledge, research explicitly measuring travel time from satellite data has not yet been
undertaken. Neither has work been produced using electro-magnetic frequencies outside
of the optical range (e.g. infrared or microwave).

Other land-based technologies, which could be used to track vehicle position, include
the use of GSM mobile-phone technology and low-frequency (LF) waves. Both these
technologies do not suffer from the urban canyon effect experienced by GPS probe vehicles.
Research by Alger et al. (2003) attempted to derive the travel times of vehicles on a
2.4. Data Collection Methods

motorway by tracking the time they left each GSM base station. However, the travel time estimates produced were often highly inaccurate. However with the introduction of both E911 and E112 legislations in the USA and European Union respectively, the use of GSM data in locating vehicles may improve. This legislation requires that in future emergency services must be able to track down each mobile telephone to within 100m (Warrior et al. 2003).

Another tracking technology is based on trilateration using low-frequency waves. This technology has been commercialised by Siemens (Banks 1989). Research by Scorer & Last (1995) reported a test of the system in Milton Keynes, UK, where the technology located the test vehicle a mean distance of 2m from the centreline of the road with a standard deviation of 10.5m.

2.4.5 Comparison of ‘truth’ data collection techniques

Table 2.1 summarises the existing performance of various systems in providing the information required to estimate the actual link travel time. It should be noted that other technologies exist which can measure spot-speed, such as infra-red and acoustic sensors. Only those technologies which can explicitly measure link travel time are included.

<table>
<thead>
<tr>
<th>Technology</th>
<th>Non Sampling Error</th>
<th>Sampling Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time vehicle entered and exited link</td>
<td>Verification that vehicle travelled 'reasonably' between A and B</td>
<td>Ability to capture the whole vehicle population.</td>
</tr>
<tr>
<td>Moving Car Observer</td>
<td>Moderate - depends on performance of passenger recording the details of the journey.</td>
<td>Good - any unusual events can be logged.</td>
</tr>
<tr>
<td>GPS - survey</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>GPS - commercial</td>
<td>Poor to moderate. Dependent on temporal resolution and map-matching software.</td>
<td>Moderate - although this may need to be inferred.</td>
</tr>
<tr>
<td>ANPR</td>
<td>Good</td>
<td>Poor - needs to be inferred.</td>
</tr>
<tr>
<td>Remote sensing and aerial images</td>
<td>Moderate to good - dependent on resolution of image and image processing software.</td>
<td>Good - the image shows the vehicle travelling over the whole link.</td>
</tr>
<tr>
<td>GSM</td>
<td>Poor - greater accuracy may be available when 3-G technology is implemented.</td>
<td>Poor - greater accuracy may be available when 3-G technology is implemented.</td>
</tr>
<tr>
<td>Low Frequency</td>
<td>Moderate</td>
<td>Moderate - although this may need to be inferred.</td>
</tr>
</tbody>
</table>

Table 2.1: Comparison of existing technologies for measuring the actual travel time
2.5 Sources of error in ANPR data

This section identifies different sources of error when using travel time data derived from ANPR systems. The first error is a measurement error, whilst the last two are sampling errors.

2.5.1 Errors in matching the data

Until recently, it was not possible to read license plates automatically, and all license plates had to be recorded manually (Kryger & Ottesen 1956; Fowkes 1983). To enable all vehicles to be recorded, and also to allow for quicker transcription, only a partial license plate was recorded. In the UK for example, it was common to record only the 3 digits and one year-letter. The other 3 letters were omitted. Discarding these 3 letters resulted in the license plate not being unique, allowing for the possibility that a partial license plate recorded downstream and matched with a license plate recorded upstream referred to two separate vehicles. Such matches were called ‘spurious matches.’

With modern ANPR systems this is unlikely to be a problem since the whole license plate is read. However spurious matching could theoretically occur if the ANPR software misread a license plate.

2.5.2 Errors from unreasonable vehicles

The first source of error in estimating urban link travel time is in using matched license plate data from ‘invalid’ vehicles. An invalid vehicle has been defined in section 2.2.2. It is a vehicle that did not travel in a prompt fashion over the link.

2.5.3 Sampling Errors

The less than perfect recognition rate of ANPR cameras has already been highlighted in section 2.4.3. Further research on the recognition rate of ANPR systems is required.
In particular, it is necessary to know if certain classes of vehicles are not successfully recognised. These classes may include:

- Foreign vehicles
- Vehicles immediately behind large vehicles

Consider the latter of these two classes. Large vehicles tend to be slower than private automobiles. Thus, if the ANPR system is not recognising those vehicles delayed behind larger vehicles, then the ANPR estimation of travel time is likely to be an underestimate.

A further sampling error could arise if a camera site has multiple lanes, with one camera monitoring each lane. If one of these cameras were to malfunction, or to have a poorer recognition rate than the other cameras, then the resulting sample from the camera site would contain an over-representation of vehicles in one lane, and an under-representation of vehicles from the lane with the faulty camera. Work by Hellinga & Fu (1999) has showed that depending on the topology of the network and signal settings of any intersections on the link, this could lead to a biased estimate of the population mean being calculated.

### 2.5.4 ANPR data - Conclusions

This section has identified various errors associated with the use of ANPR data. It has indicated that it is necessary to filter raw ANPR data before it can be used to characterise the urban link travel time distribution. Thus, the rest of this chapter is primarily concerned with developing a methodology to clean ANPR data. The next section presents a literature review of ANPR cleaning treatments. Following this a new ANPR cleaning treatment, the Overtaking Rule, is presented. Variations of this treatment are also proposed. Following this an experiment is undertaken which identifies the Overtaking Rule as the best ANPR data cleaning treatment to use.

### 2.6 Cleaning matched licence plate data: Literature Review

As far back as 1956 (Kryger & Ottesen 1956), the potential for using matched license plate data to determine link travel time has been recognised. Until recently the collection of such data was not automated, and had to be undertaken manually. Fowkes (1983) describes this procedure; two observers manually reading the license plate into a tape recorder, and time stamps being noted frequently. In order to save time in both recording and transcribing the data it was common to record only parts of the licence plate. This meant it was possible that matched license plates did not come from the same vehicle. Much of the earliest research in matched license plate data was in correcting for such spurious matches. See for example Hauer (1979), and Watling & Maher (1988). Beyond this, relatively little research in cleaning matched ANPR data has been undertaken. The literature that does exist tends to take a very statistical approach to cleaning matched
license plate data, overlooking more heuristic approaches of looking at the reasonableness of the data.

2.6.1 Statistical approaches

Clark et al. (2002) summarise three statistical approaches for identifying outliers in matched license plate data. The most basic of these is the percentile test. The percentile test defines outliers to be all values in the time period (usually 5 or 15 minutes) which do not fall within two chosen percentiles; Clark et al. (2002) use the 10th and 90th percentiles. This is a rather arbitrary test, since in all cases 20% of all values are filtered out. In addition one would intuitively expect that there would be more outliers at the high travel-time tail of the distribution rather than at the low travel-time tail of the distribution, and thus it seems inappropriate to remove an equal percentage of records from either end. One reason for this is that outliers on the lower end are bounded whereas at the higher end they are not.

The second test Clark et al. (2002) mentions is the mean absolute deviation test. This test will be referred to as the MAD test. This test screens out all individual travel times, \( TT_i \), which are further than a critical distance from the median \( TT_{median} \) of the \( n \) travel times in the time period. The critical distance is defined to be 3 times the mean absolute deviation, MAD. The MAD is given by equation 2.6. A record must then meet the criteria of equation 2.7 if it is to be considered valid. Otherwise the record will be treated as an outlier.

\[
MAD = \frac{\sum_{i=1}^{n} |TT_i - TT_{median}|}{n}
\]  

(2.6)

\( MAD \) = Mean Absolute Deviation

\( TT_i \) = Travel Time of vehicle \( V_i \) from A to B.

\( TT_{median} \) = Median travel time of the \( n \) vehicles in the time period \( t \).

\( n \) = Number of vehicles passing over point A in the time period \( t \).

\[
TT_{median} - (3 \cdot MAD) \leq TT_i \leq TT_{median} + (3 \cdot MAD)
\]  

(2.7)

Because the MAD is sensitive to the presence of outliers, it may be preferable to use a more robust measure of range such as the interquartile range. This is proposed in the final test mentioned by Clark et al. (2002). The test was first proposed by Fowkes (1983), and will be referred to as the Fowkes test. Similar to the MAD test, it involves removing all values that are beyond a certain range from the sample median. The range, \( R \), is presented in equation 2.8. The first bracket represents a robust estimate of the population standard deviation. The second bracket uses the t-distribution to establish the 95% confidence limits for a valid travel time record. Because the calculated sample standard deviation is not used it is necessary to use fewer degrees of freedom when determining the appropriate t-statistic (i.e. \( n^* \leq n - 1 \)). \( F_1 \) is a correction term due to the use of the median rather
2.6. Cleaning matched licence plate data: Literature Review

than the mean (Clark et al. 2002). A record must then meet the criteria of equation 2.10 if it is to be considered valid.

\[ R = \left( \frac{Q_3 - Q_1}{1.35} \right) (t_{0.025,n^*}) \cdot F_1 \]  

(2.8)

\( F_1 = \) Correction factor for an interval around the median rather than the mean. Given by equation 2.9 (Clark et al. 2002)

\( Q_1 = \) Lower quartile of the sample

\( Q_3 = \) Upper quartile of the sample

\( n^* = \) Modified degrees of freedom for the Fowkes test

\( F(n) = \sqrt{\frac{1}{\frac{2}{\pi} + \frac{1}{n} \left( \frac{n}{\pi} - 1 \right)}} \)  

\( n \) even

\( F(n) = \sqrt{\frac{1}{\frac{2}{\pi} + \frac{1}{n} \left( \frac{4}{\pi} - 1 \right)}} \)  

\( n \) odd

(2.9)

\[ TT_{median} - R \leq TT_i \leq TT_{median} + R \]  

(2.10)

A third method for cleaning matched license plate data is reported by Papagianni (2003). This method will be referred to as the iteration-mean method. This method is comprised of the following five steps:

- Remove all records in the time period which imply a mean average speed of under 2km/h. The remaining dataset is called \( DS_1 \).

- Calculate the mean and standard deviation of the dataset \( DS_1 \). Remove all records which are a distance of more than 4 times the standard deviation from the mean. The remaining dataset is called \( DS_2 \).

- Calculate the mean and standard deviation of the dataset \( DS_2 \). Remove all records which are a distance of more than 3 times the standard deviation from the mean. The remaining dataset is called \( DS_3 \).

- Calculate the mean and standard deviation of the dataset \( DS_3 \). Remove all records which are a distance of more than 2 times the standard deviation from the mean. The remaining dataset is called \( DS_4 \).

- The estimated population mean and standard deviation are then calculated from the dataset, \( DS_4 \).

2.6.2 Critique of the existing literature

One detailed criticism of the iteration-mean method is its use of the arithmetic mean to find a central value. The median is a more robust estimator than the median and thus
2.7 New technique - Overtaking Rule approach

2.7.1 Assumptions

A new method for cleaning matched ANPR data, known as the Overtaking Rule approach is proposed. Unlike other methods it specifically uses the serial structure of the data set to help detect erroneous or inappropriate data records. It is based on the following assumptions:

- On a single-lane road, in the urban environment, it would not be expected for valid vehicles to be overtaken by following vehicles.

- On a multi-lane road, in the urban environment, valid vehicles may be overtaken. However, it would be expected that the difference between the travel time of the overtaken slower vehicle and the faster vehicle would not be great, due to the small

![Figure 2.4: Difficulty of identifying outliers in non-stationary data](image)
2.7. New technique - Overtaking Rule approach

link length and slow speed limits in urban areas. In congested conditions where only one lane is blocked, it would be expected for a driver in this lane to move into the unblocked lane. The major cause of any large difference in travel times would be due to the slower vehicle, having been overtaken, then being stopped by a red signal, whilst the faster vehicle managed to get through the junction on the green or amber light.

2.7.2 Formulation

This new technique identifies a matched ANPR travel time record on the link AB, obtained from the target vehicle, as an outlier if the following conditions, or rules, are met:

- The target vehicle has been overtaken by another vehicle.
- The target vehicle arrives at end point B more than a certain time (the tolerance time) after any vehicle which has overtaken it.

Mathematically, the travel time record corresponding to vehicle $i$, is identified as valid if its own travel time, $TT_i$, meets the criteria given in equation 2.11 when tested against the set $F(i)$ of immediately following vehicles. A following vehicle is defined as a vehicle which passes the start point A after the target vehicle.

$$TT_i \leq TT_j + TDA_{ij} + C \quad \forall j \in F(i)$$ (2.11)

The first term on the RHS, $TT_j$, represents the travel time of the following vehicle. The second variable, $TDA_{ij}$, is the time between the target and following vehicle passing point A. It allows the filter to take into account the fact that the target vehicle might be stopped by a traffic light on the link AB, enabling the following vehicle to catch it up. The third variable on the RHS is the tolerance time C. This allows for the target vehicle to be overtaken by a certain amount before it is classified as invalid.

Only following vehicles are included in the comparison tests. The reason why leading vehicles are not included, is that it is possible that there may be an incident (i.e. accident) that occurs somewhere on the link between the location of a leading vehicle and the location of the target vehicle. This would delay the target vehicle but not the leading vehicle, hence resulting in it being identified as an invalid vehicle, even if it was a valid vehicle. By including only following vehicles, any incident will delay both the target vehicle and the following vehicles.

The Overtaking Rule method has two key design parameters:

- **Number of following vehicles** - The Overtaking Rule method compares the travel time of the current vehicle against the travel time of a certain number of immediately following vehicles. The number of vehicles chosen is an important design parameter.
If a target vehicle under study is not valid, and too few following vehicles are chosen, it may be possible that all these following vehicles are also invalid, and thus the travel time of the target vehicle will look valid when compared against these other vehicles. There is no disadvantage to selecting too many vehicles to compare against apart from the increased computational effort to do so. Although selecting a large number will lead to vehicles from different periods of traffic state being included in the tests, the rule based model factors into its calculations, the actual time difference between the vehicle under test and the following vehicles passing the first camera point. Thus if a following vehicle, passing point A 15 minutes after the target vehicle, is included in the test, the value of $TDA_{ij}$ will be 900 seconds, and it will thus be unlikely that this vehicle will invalidate the target vehicle.

- **Tolerance Time** - The second key design parameter is the tolerance time, $C$, allowed. The tolerance time is the amount of time that the target vehicle can arrive at the second camera point B after any vehicle which has overtaken it en-route. Of course, on a single lane road, it would not be expected for a valid vehicle to be overtaken, and thus the tolerance time would be set to zero. On a multi-lane road, an appropriate value for the tolerance time could thus be the red-time at the signal plus a small amount. This takes into account that an overtaking vehicle may just travel through the traffic lights on a green signal, whereas the overtaken vehicle just gets stopped by a red signal and must thus wait for the green signal. The small amount allows for the fact that a valid vehicle may drive at different speeds on the link. As the tolerance time increases, the number of vehicles identified as valid but which are in fact invalid will increase. At the same time, as the tolerance time increases, the number of vehicles identified as invalid, but which are in fact valid will decrease. There is thus a trade-off when setting the tolerance time. In the experiments described later the tolerance time was set to 0 seconds for the single-lane site. On the multi-lane site, the tolerance time was set at 45 seconds in light traffic and 600 seconds in congested traffic.

### 2.7.3 Caveat - quick vehicles

The above Overtaking Rule approach is sensitive to following vehicles whose travel time is significantly quicker than other valid vehicles. In particular if a target vehicle is followed by an emergency vehicle which is able to travel at speeds in excess of the speed limit, and is able to go through red-lights, then the target vehicle may be incorrectly identified as an outlier by this technique. Links which contain dedicated bus-lanes may also lead to valid vehicles being classified as invalid. Depending on what vehicular travel time distribution is desired, it may be preferable to filter out all emergency vehicles, and buses. It may be possible to link ANPR systems to government vehicle registration databases to obtain the vehicle type. Nevertheless, empirical evidence presented by Papagianni (2003) suggests that there are many more outliers with too high a travel time than outliers with too small
a travel time. Thus there may only be a few rare occasions when the estimates from the Overtaking Rule are significantly affected by fast vehicles.

2.7.4 Overtaking Rule with pre-filtering

The previous section noted that the Overtaking Rule may be susceptible to following vehicles whose travel time is significantly quicker than other valid vehicles. This section proposes the use of pre-filtering of the travel times of the following vehicles to overcome this. The pre-filtering involves selecting a certain percentage (say 5%) of the quickest following vehicles. These vehicles are then disallowed from invalidating the target vehicle. Removing the effect of the quickest vehicles is likely to reduce the problem caused by following vehicles whose travel time is significantly quicker than other valid vehicles, although the removal of these vehicles may also prevent invalid vehicles from being identified as such. This relationship is explored in the experiment of section 2.11.

2.7.5 Overtaking Rule at low flows

A further criticism of the Overtaking Rule concerns its possible susceptibility to poor performance during times of low flow. At very low flows the time between the target vehicle and the following vehicle passing point A, $TDA_{ij}$ in equation 2.11, is likely to be large. Therefore the travel time of an invalid target vehicle $TT_i$ could be far larger than the travel time of the following vehicle $TT_j$, yet still be deemed to be a valid record. This situation could also arise if an incident occurred on the link immediately upstream of the target vehicle. This failing of the Overtaking Rule at low flows could explain why the 95th percentile travel time on Russell Square, shown in figures D.2 and D.6, is far higher between 02:00 and 04:00 than during the rest of the day, even though the median travel time during these hours is less than the rest of the day.

2.8 Probabilistic Overtaking Rule

The Overtaking Rule described in section 2.7 is deterministic. That is the target vehicle is identified as valid or invalid by each following vehicle. If any following vehicle identifies the target vehicle as invalid then that vehicle is invalidated. This deterministic characteristic of the Overtaking Rule can result in one single very quick vehicle (i.e. an emergency vehicle) invalidating several valid vehicles. An approach to overcome this has been outlined in the previous section. An alternative approach is to move away from a deterministic model to a probabilistic one. That is each of the $n$ following vehicles calculates a probability, $p_j$, that the target vehicle is invalid. The overall probability of the target vehicle being valid, $PTargetValid$, is then given by equation 2.12.

$$PTargetValid = \prod_{n} p_j$$ (2.12)
The vehicle is said to be valid if \( P_{TargetValid} \) is greater than a certain threshold, \( P_{threshold} \). This is stated formally in equation 2.13.

\[
P_{TargetValid} > P_{threshold}
\]  

(2.13)

The challenge is to find a function to describe \( p_j \). Two such functions are proposed below.

**2.8.1 Function 1 - Logistic model**

The first function uses the logistic model and is given by equation 2.14.

\[
p_j = \frac{1}{1 + \exp(-\beta \cdot \Delta(i,j))}
\]  

(2.14)

\( \beta \) = parameter to be chosen.

Where

\[
\Delta(i, j) = TT_j + TDA_{ij} + C - TT_i
\]  

(2.15)

Figure 2.5 shows the response curve of this function for varying values of \( TT_j - TT_i \) having fixed \( C \) as 45 and assumed that both vehicles pass the starting point at the same time (i.e. \( TDA_{ij} = 0 \)).

**2.8.2 Function 2 - Truncated logistic model**

One problem with the form of the logistic function of equation 2.14 is that even when the travel time of the target vehicle is less than the travel time of the following vehicle the probability that the target vehicle is not valid is less than one. Since the overall validity of the target vehicle is given by the product of the probabilities calculated for each individual vehicle (equation 2.12) it is possible for a vehicle which is clearly valid to be identified as invalid. To overcome this short fall the truncated logistic model is suggested. It is given by equation 2.16 and shown as the bottom graph in figure 2.2.

\[
p_j = \frac{1}{1 + \exp(-\beta \cdot \Delta(i,j))} \quad \Delta(i, j) < 0
\]

\[
p_j = 1 \quad \Delta(i, j) \geq 0
\]  

(2.16)

The performance of these two probabilistic Overtaking Rule methods compared to the deterministic Overtaking Rule will be explored in section 2.11. Before this can be done, it is first necessary to define the measures of performance.
2.9 Measures of Performance

The previous three sections have identified existing ANPR data cleaning treatments, and also proposed a new treatment called the Overtaking Rule. The next three sections present the methodology and results of an experiment to quantify the performance of each of these ANPR data cleaning methods. One of the criticisms of the existing literature is that they fail to explicitly measure the performance of the data cleaning methods. This first section presents two ways of measuring the performance of the models.

2.9.1 Ability to Identify Outliers

The first performance measure was how successful each outlier test identified outliers. In particular, three metrics were calculated:

- Percentage correctly classified - the percentage of all records that are correctly classified as being either valid or invalid.
- False Positives - Number of vehicles identified as having valid travel time records, when in fact, they had invalid records.
- False Negatives - Number of vehicles identified as having invalid travel time records, when in fact, they had valid records.
2.9. Measures of Performance

2.9.2 Accuracy in characterising the distribution

The end goal of any ANPR data exercise is to be able to characterise the distribution of the actual travel time (i.e. the true travel time of valid vehicles) in any selected time interval. The motivation for identifying and excluding outliers from the data sample is that these outliers may affect any estimator used to characterise this actual travel time distribution. Two obvious choices of metric that can characterise the distribution are the population mean travel time of all valid vehicles ($\mu^*$) and the population standard deviation of the travel time of all valid vehicles ($\sigma^*$). Ideally each cleaning method should remove all invalid vehicles allowing $\mu^*$ and $\sigma^*$ to be accurately estimated.

The two sample based estimators used to estimate $\mu^*$ using the cleaned data in each time interval are:

- Sample mean - arithmetic mean of all remaining cleaned data.
- Sample median - median of all remaining cleaned data.

The two sample based estimators used to estimate $\sigma^*$ will be:

- Sample standard deviation - standard deviation of all remaining cleaned data.
- Percentile Standard Deviation - This is defined as the difference between the 84th and the 16th percentile divided by two. It is based upon the fact that in a normal distribution 68% of all values are within one standard deviation of the mean. This metric has been widely used by geologists as a robust measure of standard deviation (Helsel & Hirsch 2002). It is robust because any extreme values will lie between the 0 and 16th or between the 84th and 100th percentile, and thus have limited influence on the measure.

The actual and estimated $\mu^*$ and $\sigma^*$ can then be compared to one another using the mean absolute percentage error (MAPE) or the root mean square error (RMSE). These are given in equations 2.17 and 2.18 below, where $x_i$ refers to the population mean or population standard deviation, and $\hat{x}_i$ refers to the estimated value of this using a relevant estimator outlined above. $L$ refers to the number of test runs or data records the accuracy is measured over.

\[
MAPE = \left( \frac{1}{L} \sum_{i=1}^{L} \left| \frac{x_i - \hat{x}_i}{x_i} \right| \right) \cdot 100 \tag{2.17}
\]

\[
RMSE = \sqrt{\frac{1}{L} \sum_{i=1}^{L} (x_i - \hat{x}_i)^2} \tag{2.18}
\]

The estimators and measures of accuracy will be used in the experiment outlined in the next section. Another metric of accuracy used later in the thesis is the Mean Absolute
2.10. Comparison of cleaning methods - experiment methodology

2.10.1 Data Requirements

As a first step in the investigation of the performance of each outlier test, an experiment was set up using data from the VISSIM microsimulation (Fellendorf & Vortisch 2001; PTV 2001). VISSIM’s performance has been widely evaluated and validated (e.g. Bloomberg & Dale 2000; Park & Schneeberger 2003). VISSIM allows any road network to be created, and the characteristics of the vehicle fleet (vehicle types, and desired speed) to be set. It is possible to record the time and vehicle ID of all vehicles passing a desired location. This effectively simulates the operation of the ANPR cameras. By locating such ‘data collection points’ over the parking areas, bus-stop, and alternative route, VISSIM is able to identify which of the 5 types of travel time distributions outlined above each travel time record came from. It is thus possible to explicitly identify all valid vehicles enabling $\mu^*$ and $\sigma^*$ to be calculated. This is because VISSIM allows the whole vehicle population, and not just a sample of vehicles, to be monitored. These two metrics were thus calculated for each time interval of 5 minutes.

Error (MAE), and is given by equation 2.19.

\[
MAE = \left( \frac{1}{L} \sum_{i=1}^{L} | x_i - \hat{x}_i | \right)
\]  

(2.19)
2.10.2 Network Layout

The road network was constructed in the VISSIM microsimulation. It is shown in figure 2.6. The link included the following features:

- **Alternative Route** - this feature simulates vehicles that may pass the camera site at point A when travelling on a journey that does not include the link AB. Later in the day, the vehicle may pass over the camera site B, on another journey that does not include the link AB. However, the ANPR record at camera sites A and B will be matched. The resulting calculated travel time is likely to be very large, often in the order of hours.

- **Medium term parking area** - this feature simulates vehicles that park on the link AB for a medium amount of time (e.g. to fill up with petrol, long wait for vehicle passengers, etc)

- **Short term parking area** - this feature simulates vehicles that park on the link AB for a short amount of time. (e.g. to drop off a passenger, to buy something from a newsagent, etc)

- **Bus-stop** - this feature simulates buses stopping mid-link in a bus bay.

- **Emergency Vehicle Lane** - emergency vehicles can bypass the traffic lights at junction 2 as well as any queuing traffic caused by this junction.

- **ANPR camera sites** - various detectors are able to determine which travel time distribution the vehicle belongs to. These sites record all the vehicles that pass them. In VISSIM each vehicle is assigned a unique number, so can be tracked throughout the network.

- **Bottleneck at junction 3** - The width of the road at junction 3 is only one lane wide. In addition, vehicles are prevented from turning right by oncoming traffic. Since the road is only one lane wide, vehicles are forced to queue behind the turning vehicle. The effect of the bottleneck can be regulated by changing the amount of oncoming flow $q_{BA}$. It is this bottleneck that causes a variation in link travel time over the duration of the simulation.

The network thus allowed 4 of the 5 travel time distributions to be incorporated into the data. Placing detectors on the features enabled the category of travel time distribution of each vehicle to be known. No possible way seems available within VISSIM to explicitly identify irrational drivers - e.g. those drivers which do not change lane when it would be very beneficial to do so. Since it was not possible to explicitly identify irrational drivers, the VISSIM network was designed to limit the possibility of irrational drivers being present. VISSIM allows various driving parameters to be changed to achieve this. Changing lanes was encouraged, by increasing the average length between stopped vehicles and reducing
the gap required for vehicles to attempt an overtaking manoeuvre. Similar number of obstructions were placed on either side of the carriageway in order to reduce the difference in link travel times experienced by different lanes of traffic. A visual inspection of the running of the simulation suggested that the vehicles were behaving in a rational and realistic fashion.

2.10.3 Simulation Settings

The settings used within the simulation are presented in table 2.2. An unusually large proportion of emergency vehicles was used to allow the effect of very quick vehicles on the performance of the Overtaking Rule to be noticeable. It is assumed that the ANPR cameras capture the licence plates of the whole vehicle population, although it has already been noted that in real life recognition rates tend to be only of the order of 50 - 90%. To allow for random variation in the microsimulation, the results were averaged over 10 runs, each run using a different random seed.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Single Lane Setting</th>
<th>Multi Lane Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic Signals</td>
<td>Cycle Time</td>
<td></td>
</tr>
<tr>
<td>A-B</td>
<td>60 secs</td>
<td>60 secs</td>
</tr>
<tr>
<td></td>
<td>Green Time</td>
<td>40 secs</td>
</tr>
<tr>
<td>Veh. composition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cars</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>Buses</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>HGV</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>Emergency Veh.</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>Desired veh. speed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergency Veh.</td>
<td>45 - 60 kph</td>
<td>45 - 60 kph</td>
</tr>
<tr>
<td>Buses</td>
<td>25 - 30 kph</td>
<td>25 - 30 kph</td>
</tr>
<tr>
<td>Other Vehicles</td>
<td>30 - 40 kph</td>
<td>30 - 40 kph</td>
</tr>
<tr>
<td>Demand flow q_{AB}</td>
<td>0 - 2000 secs</td>
<td>700 vph</td>
</tr>
<tr>
<td></td>
<td>2001 - 4000 secs</td>
<td>900 vph</td>
</tr>
<tr>
<td></td>
<td>4001 - 6000 secs</td>
<td>1500 vph</td>
</tr>
<tr>
<td></td>
<td>6001 - 12000 secs</td>
<td>900 vph</td>
</tr>
<tr>
<td></td>
<td>12001 - 16000 secs</td>
<td>700 vph</td>
</tr>
<tr>
<td>Demand flow q_{BA}</td>
<td>0 - 5000 secs</td>
<td>800 vph</td>
</tr>
<tr>
<td></td>
<td>5001 - 8000 secs</td>
<td>300 vph</td>
</tr>
<tr>
<td></td>
<td>8001 - 16000 secs</td>
<td>100 vph</td>
</tr>
</tbody>
</table>

Table 2.2: Other parameter settings in the simulation

2.10.4 Data Collection

For each vehicle travelling from A to B a record in the database was created. This record included the time the vehicle passed point A, the travel time, vehicle type, and the travel time distribution the vehicle belonged to (i.e. whether the vehicle stopped en-route, took an alternative route, etc). For each 5 minute period μ* and σ* were calculated using all records from valid vehicles which passed over point A in the 5-minute period. In addition, an estimate of these values was obtained from each of the ANPR data cleaning methods using travel time records from all vehicles within the same 5-minute period. These values
were used to calculate the accuracy of the methods.

### 2.10.5 Overtaking rule settings

Sections 2.7 and 2.8 proposed three new ILD data cleaning treatments: the deterministic Overtaking Rule (OR), deterministic Overtaking Rule with pre-filtering (OR + pre-filtering), and the probabilistic Overtaking Rule. The first two treatments were included in the comparison of cleaning methods. Initial trials suggested there was a small amount of difficulty in determining settings for the multi-lane link. The performance of this Overtaking Rule method was seen to improve if two tolerance times were given: one for uncongested periods and the other for congested periods. The road was said to be uncongested if, in the set of travel times of the following vehicles, there was at least one travel time below a certain value (150 seconds was used). For the single-lane road no tolerance time was given. This follows the justification given in section 2.7.2. The Overtaking Rule settings are outlined in Table 2.3.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Single-lane link</th>
<th>Multi-lane link</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of following vehicles to include.</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Tolerance Time: uncongested, secs</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>Tolerance Time: congested, secs</td>
<td>0</td>
<td>600</td>
</tr>
<tr>
<td>Threshold congested / uncongested. Minimum travel time in set, secs.</td>
<td>n/a</td>
<td>150</td>
</tr>
<tr>
<td>Pre-filtering: percentage of fastest vehicles which are not allowed to exclude others.</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

* Due to the averaging process, these values are not integer.

Table 2.3: Overtaking rule settings used in the experiment of section 2.11.

The probabilistic Overtaking Rule is not included in the experiment. Whereas the deterministic Overtaking Rule has two key design parameters (number of following vehicles, and the tolerance time), the probabilistic Overtaking Rule has an additional three (truncated or non-truncated model, $\beta$, and $p_{\text{threshold}}$). A comprehensive search through combinations of these settings failed to find a combination that gave better results than the deterministic Overtaking Rule. Due to its poor performance and increased complexity it was thus decided to exclude this treatment from the experiment described in the next section.
### 2.11 Comparison of cleaning methods - results

#### 2.11.1 Experiment #1 - single-lane road

The first site used to test the performance of the various matched license plate cleaning methods was a single-lane road. Such roads are not uncommon in urban road networks in the UK. The value of $\mu^*$, aggregated over a 5 minute period, was seen to vary from 94 to 323 seconds over the course of the simulations, and $\sigma^*$ from 6 to 72 seconds depending on the flow settings. The methods were thus tested in both congested and uncongested conditions.

Three measures of performance are proposed; the ability of each method to identify outliers in the data, and the accuracy of each method in estimating $\mu^*$ and $\sigma^*$. Table 2.4 shows the performance of each method in identifying outliers when the data is pre-screened to exclude all emergency vehicles and buses. After this screening an average of 30.2% of the total travel time records still came from invalid vehicles.

Table 2.5 shows the performance of all methods in estimating $\mu^*$ and $\sigma^*$ when the data is pre-screened to exclude all emergency vehicles and buses. Table 2.6 shows the...
2.11. Comparison of cleaning methods - results

<table>
<thead>
<tr>
<th>Method name</th>
<th>No of records*</th>
<th>% correctly classified</th>
<th>no. false positive*</th>
<th>no. false negative*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overtaking Rule</td>
<td>3496.0</td>
<td>80.0</td>
<td>186.7</td>
<td>514.1</td>
</tr>
<tr>
<td>OR + pre-filtering</td>
<td>3496.0</td>
<td>88.1</td>
<td>212.4</td>
<td>204.7</td>
</tr>
<tr>
<td>Percentile</td>
<td>3496.0</td>
<td>62.9</td>
<td>546.6</td>
<td>750.0</td>
</tr>
<tr>
<td>MAD</td>
<td>3496.0</td>
<td>67.9</td>
<td>1085.9</td>
<td>37.7</td>
</tr>
<tr>
<td>Fowkes</td>
<td>3496.0</td>
<td>71.8</td>
<td>946.6</td>
<td>40.2</td>
</tr>
<tr>
<td>Iteration - mean</td>
<td>3496.0</td>
<td>65.6</td>
<td>1162.9</td>
<td>39.4</td>
</tr>
<tr>
<td>Iteration - median</td>
<td>3496.0</td>
<td>68.7</td>
<td>1053.1</td>
<td>39.6</td>
</tr>
</tbody>
</table>

* Due to the averaging process, these values are not integer

Table 2.6: Ability of each method to identify outliers in the matched ANPR data, on a single-lane road link. Emergency vehicles and buses NOT filtered from the dataset.

The performance of all methods in identifying outliers when the data is not pre-screened, and includes data from all vehicle types. On average 44.5% of the total travel time records came from invalid vehicles. Table 2.7 shows the performance of all methods in estimating \( \mu^* \) and \( \sigma^* \) when the data is not pre-screened, and includes data from all vehicle types.

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Estimate of ( \mu^* )</th>
<th>Estimate of ( \sigma^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sample Mean</td>
<td>Sample Median</td>
</tr>
<tr>
<td>Overtaking Rule</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR + pre-filtering</td>
<td>4.26</td>
<td>8.22</td>
</tr>
<tr>
<td>Percentile</td>
<td>2.68</td>
<td>5.30</td>
</tr>
<tr>
<td>MAD</td>
<td>13.65</td>
<td>19.05</td>
</tr>
<tr>
<td>Fowkes</td>
<td>22.53</td>
<td>32.30</td>
</tr>
<tr>
<td>Iteration - mean</td>
<td>11.75</td>
<td>23.66</td>
</tr>
<tr>
<td>Iteration - median</td>
<td>40.10</td>
<td>65.49</td>
</tr>
<tr>
<td>No cleaning</td>
<td>21.25</td>
<td>32.58</td>
</tr>
</tbody>
</table>

Table 2.7: Accuracy of the estimates of \( \mu^* \) and \( \sigma^* \) from each of the data cleaning methods. Data from single-lane road link. Emergency vehicles and buses NOT filtered from the dataset.

From these results the following conclusions can be drawn:

- The overtaking (OR) method with pre-filtering produces the most accurate estimates of \( \mu^* \) and \( \sigma^* \) in both filtered and unfiltered data.
- The OR without pre-filtering produces more accurate estimates of \( \mu^* \) and \( \sigma^* \) in both filtered and unfiltered data than the other non-OR methods
- The best estimator of \( \mu^* \) for all methods bar the OR treatments was the median travel time of all remaining cleaned records.
- With the exception of the OR treatments, the best estimator of \( \sigma^* \) is the percentile standard deviation.
• The OR method with pre-filtering was most accurate in identifying the validity of records.

• As predicted the iteration-median method provides more accurate results than the iteration-mean method.

• All cleaning methods output more accurate results than if no cleaning is undertaken.

As predicted the OR method without pre-filtering appeared to be sensitive to the inclusion of travel time records from emergency vehicles and buses. The percentage of vehicles correctly classified fell from over 97% to just 80% when pre-screening of the data was not performed. Similarly the MAPE increased from 0.31% to 4.26% when filtering was not performed. The OR rule with pre-filtering was less sensitive to the lack of pre-screening. The percentage of vehicles correctly classified fell from 98% to 88%, whilst the MAPE increased from 0.22% to 2.68%. Other statistical methods are also less sensitive to the failure to pre-screen the data. Nevertheless, the OR method outperformed all other methods regardless of whether screening was performed.

2.11.2 Experiment #2 - multi-lane road

The second experiment to test the performance of the various outlier tests, is on a link with multiple-lanes. Many arterial routes in urban environments are dual or even triple-lane carriageways, so it is important to verify the success of the models on these types of links too. The network layout was the same as for the single-lane of figure 2.6. This time however, the main link between A and B was 3 lanes wide. The value of \( \mu^* \) was seen to vary from 85 to 755 seconds, and \( \sigma^* \) from 9 to 192 seconds depending on the flow settings.

The pattern of results for the multi-lane link followed that of the single lane link. Thus due to space considerations only table 2.8 is presented. This table shows the performance of all methods in estimating \( \mu^* \) and \( \sigma^* \) when the data is not pre-screened, and includes data from all vehicle types.

Similar conclusions can be drawn from the multi-lane site as were drawn from the single lane site. However, in this case the OR without pre-filtering was seen to outperform the OR with pre-filtering when used with both the pre-screened and non-screened data sets. The OR method without pre-filtering provides the most accurate estimates for both \( \mu^* \) and \( \sigma^* \) in all bar one case; when accuracy is measured solely by the RMSE. This result suggests that the performance of the Overtaking Rule may deteriorate in congested periods when travel times are large. The performance of the Overtaking Rule was seen to improve if two different tolerance values in equation 2.11 were used; one for uncongested, and one for congested traffic. For all methods, bar that of the Overtaking Rule when estimating \( \mu^* \), the robust estimators are found to provide more accurate estimates of both \( \mu^* \) and \( \sigma^* \).
2.12. ANPR data from TfL congestion charging cameras

On 17th February 2003, congestion charging was implemented in central London. Drivers must pay to use their cars within central London between 7:00 and 18:30. In order to enforce the scheme, a total of 254 colour and 434 mono ANPR cameras were installed at 203 different fixed sites, located primarily on the boundary of the charge zone, but also within the zone (TfL 2003). These ANPR cameras record the license plates of all vehicles passing the sites and check that the vehicle has been registered as having paid the congestion charge. For each vehicle recognised by an ANPR camera, the following information is stored in a record:

- Unique anonymous vehicle id
- Timestamp that vehicle passed the camera site
- Type of vehicle (private car, lorry, etc). This is obtained from the Driver and Vehicle Licensing Agency (DVLA) database

This record is then archived taking into account data protection principles. By calculating the difference between the times that a vehicle passed one camera site and the time it passed a second camera site, it is possible to calculate the travel time between the two camera sites. Since most ANPR cameras are located on the outer boundary of the congestion charging zone there are few opportunities to create D-D links of a length under 2km (using the parlance of section 2.1). However, there are several camera sites located within the congestion charging boundary and from these camera sites it was possible to identify two candidate links which also contained several ILDs on them. These were the Tottenham Court Road link (running from Bedford Avenue to Grafton Way) and the Russell Square Link (running from Upper Woburn Place to the junction of Southampton Row and Theobald’s road). The first of these links was a multi-lane road, whilst the second

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Estimate of $\mu^*$</th>
<th>Estimate of $\sigma^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sample Mean</td>
<td>Sample Median</td>
</tr>
<tr>
<td></td>
<td>MAPE</td>
<td>RMSE</td>
</tr>
<tr>
<td>Overtaking Rule</td>
<td>3.68</td>
<td>16.10</td>
</tr>
<tr>
<td>OR + pre-filtering</td>
<td>3.82</td>
<td>14.82</td>
</tr>
<tr>
<td>Percentile</td>
<td>11.15</td>
<td>22.46</td>
</tr>
<tr>
<td>MAD</td>
<td>21.89</td>
<td>36.24</td>
</tr>
<tr>
<td>Fowkes</td>
<td>8.01</td>
<td>17.35</td>
</tr>
<tr>
<td>Iteration - mean</td>
<td>32.45</td>
<td>61.26</td>
</tr>
<tr>
<td>Iteration - median</td>
<td>18.34</td>
<td>34.03</td>
</tr>
<tr>
<td>No filtering</td>
<td>130.95</td>
<td>198.28</td>
</tr>
</tbody>
</table>

Table 2.8: Accuracy of the estimates of $\mu^*$ and $\sigma^*$ from each of the data cleaning methods. Data from multi-lane road link. Emergency vehicles and buses NOT filtered from the dataset.
2.13 Validation of ANPR data

To ensure that the travel time data obtained from the ANPR system, and used in the development of the ULTT models in part B, was a good representation of the actual travel time, some validation work was performed. This validation work was carried out with the link contained only a single lane. The characteristics of these two links are shown in table 2.9. A map showing their relative location in London is shown in Figure 2.7.

<table>
<thead>
<tr>
<th>Link Name</th>
<th>Tottenham Court Road</th>
<th>Russell Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera site #1</td>
<td>201 (Bedford Avenue)</td>
<td>65 (Upper Woburn Place)</td>
</tr>
<tr>
<td>Camera site #2</td>
<td>73 (Grafton Way)</td>
<td>202 (Southampton Row)</td>
</tr>
<tr>
<td>Length, m</td>
<td>680</td>
<td>1060</td>
</tr>
<tr>
<td>No bus routes</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Number of lanes (total)</td>
<td>2 to 3</td>
<td>1</td>
</tr>
<tr>
<td>Bus lane present?</td>
<td>No</td>
<td>Partly</td>
</tr>
<tr>
<td>On street parking allowed</td>
<td>Yes with restrictions</td>
<td>Yes with restrictions</td>
</tr>
<tr>
<td>no traffic lights on route</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>no other major junctions</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2.9: Characteristics of the Tottenham Court Road and Russell Square links

Figures 2.8 and 2.9 show scatterplots of the typical travel times by time-of-day for the Tottenham Court Road link and Russell Square link respectively.

2.13 Validation of ANPR data

To ensure that the travel time data obtained from the ANPR system, and used in the development of the ULTT models in part B, was a good representation of the actual travel time, some validation work was performed. This validation work was carried out with the link contained only a single lane. The characteristics of these two links are shown in table 2.9. A map showing their relative location in London is shown in Figure 2.7.

<table>
<thead>
<tr>
<th>Link Name</th>
<th>Tottenham Court Road</th>
<th>Russell Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera site #1</td>
<td>201 (Bedford Avenue)</td>
<td>65 (Upper Woburn Place)</td>
</tr>
<tr>
<td>Camera site #2</td>
<td>73 (Grafton Way)</td>
<td>202 (Southampton Row)</td>
</tr>
<tr>
<td>Length, m</td>
<td>680</td>
<td>1060</td>
</tr>
<tr>
<td>No bus routes</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Number of lanes (total)</td>
<td>2 to 3</td>
<td>1</td>
</tr>
<tr>
<td>Bus lane present?</td>
<td>No</td>
<td>Partly</td>
</tr>
<tr>
<td>On street parking allowed</td>
<td>Yes with restrictions</td>
<td>Yes with restrictions</td>
</tr>
<tr>
<td>no traffic lights on route</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>no other major junctions</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2.9: Characteristics of the Tottenham Court Road and Russell Square links

Figures 2.8 and 2.9 show scatterplots of the typical travel times by time-of-day for the Tottenham Court Road link and Russell Square link respectively.
2.13. Validation of ANPR data

Figure 2.8: Scatterplot of individual vehicle travel times by time-of-day for the Tottenham Court Road link on Wednesday 9th April 2003. Data obtained from ANPR cameras

Figure 2.9: Scatterplot of individual vehicle travel times by time-of-day for the Russell Square link on Wednesday 9th April 2003. Data obtained from ANPR cameras
assistance of an MSc student, Christophe Begon, although the preparation and management of the validation project was undertaken by this author. This section describes the validation work that was undertaken and presents the results of this validation work.

2.13.1 Methodology

This section outlines an experiment undertaken to determine how similar the characteristics of ULTT (such as the mean and standard deviation of travel time) derived using ANPR data were to the actual characteristics of ULTT. This experiment primarily involved the collection of two data sets; an ANPR data set, and a data set representing the actual ULTT. The requirements for measuring the actual ULTT have been outlined in section 2.3.4. Thus the following approach was used to obtain the dataset of actual ULTT:

- Manually record the license plates, time of passage, and vehicle type (car, bus, taxi, etc.), of all vehicles passing both ANPR camera sites on the study link.

- Identify all those vehicles which did not travel ‘in a reasonable fashion’ between the two camera sites. For example, it was desired to note the license plate of all vehicles exiting or entering the link mid-link, and note down those vehicles which parked mid-link.

A survey was conducted on Tuesday 29th June 2004 at both the Tottenham Court Road and Russell Square links described in section 2.12. Pilot studies suggested that both a video camera and a person could only accurately note down all vehicles in a single lane. This could be overcome to some extent by focussing the video to ensure that it captured all vehicles passing over one lane, and then reading the license plate into the video of the traffic passing over the other lane. However camera sites 201 and 202 both had three lanes of traffic. Since only two video cameras were available it was thus necessary to have two people and one video camera at these two sites. Due to limited resources the survey team consisted of only 4 persons. This meant that only one person was able to note down vehicles not travelling in a reasonable fashion. Since there was no obvious ‘rat-run’ over these two links, it was felt that any vehicle entering or exiting mid-link would not be taking short cuts but would be on a totally different journey. If by coincidence it did pass over two camera sites in sequence it would be likely that the time between passing over both cameras sites would be so high as to identify it clearly as an outlier. Thus this observer was used to note down those vehicles which parked for short periods of time (e.g. delivery vans, taxis, etc.).

To ensure that both sites were surveyed under peak and off-peak conditions the following timetable was used for the survey.

In addition to travel time records collected from the manual survey, travel time records derived from the ANPR cameras on the links were also made available by TfL. Begon
2.13. Validation of ANPR data

Table 2.10: Timetable of ANPR validation survey

<table>
<thead>
<tr>
<th>Dataset no</th>
<th>Link name</th>
<th>Traffic State</th>
<th>Start Time</th>
<th>End Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Russell Square</td>
<td>Peak</td>
<td>08:15</td>
<td>09:15</td>
</tr>
<tr>
<td>2</td>
<td>Tottenham Court Road</td>
<td>Off-peak</td>
<td>09:45</td>
<td>10:45</td>
</tr>
<tr>
<td>3</td>
<td>Russell Square</td>
<td>Off-peak</td>
<td>15:00</td>
<td>16:00</td>
</tr>
<tr>
<td>4</td>
<td>Tottenham Court Road</td>
<td>Peak</td>
<td>16:30</td>
<td>17:30</td>
</tr>
</tbody>
</table>

(2004) undertook two tests. In the first test he used solely the survey data (which represents that actual travel time distribution) to determine which of the ANPR data cleaning methods correctly identified invalid vehicles (i.e. those which stopped en-route). In the second method he then verified that the travel time records from the ANPR data-set came from the same underlying travel time distribution as the actual travel time records.

2.13.2 Comparison of ANPR data cleaning methods

The first test undertaken by Begon (2004) was to use solely the survey data to see which ANPR cleaning method could correctly identify invalid vehicles and enable the correct calculation of the basic characteristics of the underlying travel time distribution (i.e. the mean and standard deviation of vehicle-to-vehicle travel times). This work was carried out using dataset #4 of Table 2.10 since this dataset contained the most records. Table 2.11 shows that the estimate of the mean travel time and standard deviation made by the Overtaking Rule were closest to the actual survey results and outperform the other ANPR cleaning methods. However, this study did not include the Overtaking Rule with pre-filtering.

Table 2.11: Validation of the overtaking rule using real world data (Begon 2004).

<table>
<thead>
<tr>
<th></th>
<th>Survey (truth)</th>
<th>Percentile</th>
<th>MAD</th>
<th>Fowkes</th>
<th>Iteration Mean</th>
<th>Iteration Median</th>
<th>Overtaking rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean TT, secs</td>
<td>193.6</td>
<td>195.8</td>
<td>198.0</td>
<td>191.5</td>
<td>192.2</td>
<td>191.5</td>
<td>192.7</td>
</tr>
<tr>
<td>Standard Error, secs</td>
<td>2.0</td>
<td>1.7</td>
<td>1.8</td>
<td>1.9</td>
<td>1.9</td>
<td>1.9</td>
<td>2.0</td>
</tr>
<tr>
<td>Standard deviation TT removed</td>
<td>41.5</td>
<td>31.2</td>
<td>33.7</td>
<td>38.4</td>
<td>37.5</td>
<td>38.4</td>
<td>40.8</td>
</tr>
<tr>
<td>Valid TT removed</td>
<td>-</td>
<td>58</td>
<td>0</td>
<td>7</td>
<td>10</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Invalid TT nor removed</td>
<td>-</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
was undertaken for each of the four datasets of table 2.10 using a Kolmogorov-Smirnov test. The results of this are shown in Table 2.12.

<table>
<thead>
<tr>
<th>Dataset no (see table 2.10)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n^* )</td>
<td>36</td>
<td>79</td>
<td>71</td>
<td>95</td>
</tr>
<tr>
<td>Test statistic = ( D \cdot n^*^{\frac{1}{2}} )</td>
<td>0.565</td>
<td>0.598</td>
<td>0.812</td>
<td>1.517</td>
</tr>
<tr>
<td>Significance of the test. ( p = 0.05 ) (critical value = 1.36)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Significance of the test. ( p = 0.01 ) (critical value = 1.63)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 2.12: Two-sample Kolmogorov tests to determine whether the distribution of travel times from the ANPR data differs significantly from the actual distribution of travel times (Begon 2004).

### 2.13.4 Conclusions

The first test of Begon (2004) showed that the Overtaking Rule was the most effective ANPR cleaning filter, of all those tested, for use with real-world travel time records. The second part of Begon’s work suggests that the estimates of the distribution properties of travel time obtained from the filtered ANPR data are not significantly different from the actual values of travel times. Importantly the mean and standard deviation of travel time from the ANPR Russell Square link dataset were found not to be significantly different from those of the actual travel time distribution. These results justify the use of the ANPR derived data as a source of actual travel times. The ANPR data filtered using the Overtaking Rule will thus be used to calibrate the ULTT models developed in part B of this thesis.

### 2.14 Caveat: Link-based or path based travel time models

This chapter has discussed how the actual travel time over a link can be accurately and reliably measured. Accurate estimates of actual link travel time are used for the correct calibration of any ULTT model. However, there is some debate in the literature over whether link based or path based travel time models are more appropriate. In the former, the travel time for a journey from point A to point B will be calculated by first determining which links connect these two points and then summing up the travel times of these individual links. This approach is sometime called the Link-Junction (LJ) model (Dale et al. 1996). In the latter case, the link is defined as the roadway from point A to point B.

Chen & Chien (2001) propose the use of the latter model as they argue reasonably that depending on the definition of the link, the link travel time may include travel time from all vehicles regardless of their turning movements. However, it would be expected that the travel time of a vehicle turning against oncoming traffic at an intersection will be
Figure 2.10: Link based methods should ensure that delay experienced solely by one turning movement is not included in the travel time of another turning movement.

However, this is not a fault of the link-based method, but rather of the way the links are defined. If the LJ model is used, then the network must be divided up into continuous links, ideally so that any delay experienced solely by one turning movement is not included in the travel time of another turning movement. This suggests further that if a link-based travel time model is used then D-D links should be used to define links on the road network. This will ensure that the correct intersection delays are modelled, thus providing the link-based model with one of the key benefits of the path-based travel time model. This is illustrated in figure 2.10.

However, even using D-D links may not eliminate all such errors. There may be more than one junction or traffic light on any one link. Often the traffic lights on straight links are coordinated so that vehicles get a ‘green-wave’ and do not have to stop. However, vehicles entering the link are likely to get stopped by the first traffic signal until they too enter the ‘green wave’. However the link travel time will include the travel times of these vehicles which have just entered the link and thus bias upwards any estimate of travel time for vehicles already in the ‘green wave.’ This problem has been recognised by Hellinga & Fu (1999).

However, the major advantage that the link-based model has over the path-based model is the sample size. As stated in section 2.3.3, a travel time record is only one realisation of the underlying distribution over the link (or path). To obtain an accurate measure of $\mu^*$ several measurements of travel time are needed as stipulated in equation 2.5. Since a link may belong to several different paths, then a link will get more travel time readings in the same time interval than a path. It is for this reason that link-based travel time models are preferred in the literature.

2.15 Conclusions

This chapter has presented a comprehensive review of the urban link travel time data that will be used in the rest of this thesis. Both the terms ‘link’ and ‘urban link travel time’
were explicitly defined. A nomenclature for describing the type of link was proposed. The links used in this thesis are D-D links.

Three difficulties in measuring urban link travel time were then discussed. These were: determining the time each vehicle entered and exited the link, ensuring that only travel times from vehicles which travelled in a ‘reasonable’ fashion over the link were used, and finally in ensuring that a representative sample of travel times in the time-period of interest was obtained. Various technologies for collecting travel time data were discussed and GPS and ANPR technologies seen to be the most favourable. Since an ANPR data source was made available by TfL, ANPR data was used.

A literature review of methods used for cleaning ANPR data was given. Limitations in the existing methods were identified. In particular, existing methods fail to use the temporal-structure of the travel time dataset. Thus a new cleaning method, the Overtaking Rule was proposed. This method identifies erroneous travel time records by determining if a vehicle which provided a travel time record was overtaken on the link and arrived at the end of the link a certain time after the vehicle which overtook it. This method thus uses the temporal-structure of the travel time dataset. A novel experiment was designed using simulated data to quantify the performance of the various ANPR cleaning methods. This experiment identified the Overtaking Rule as the best cleaning method. The Overtaking Rule is therefore used to clean ANPR travel time data used in the rest of this thesis.

The last part of the chapter describes the ANPR data used in this thesis. Data comes from two links: multi-lane Tottenham Court Road, and single-lane Russell Square. An experiment was carried out using data collected from field surveys which suggested that ANPR travel time records came from the same distribution as the actual travel time records. This experiment also confirmed the Overtaking Rule as the best cleaning method to use.

This section has considered the ‘output’ data from the ULTT models that will be developed in part B - i.e. travel time. The next chapter considers the primary input data to the ULTT model - data from Inductive Loop Detectors.
Chapter 3

Data from Inductive Loop Detectors

The primary objective of this thesis is to propose a methodology to model Urban Link Travel Time (ULTT) using data from Inductive Loop Detectors (ILDs). This chapter provides a comprehensive introduction to ILDs, and describes how they can provide data of good quality for use in a ULTT model. This chapter thus builds on the previous chapter which proposed a filtering method to obtain measures of urban link travel time using data from ANPR cameras. The chapter begins with a brief explanation of what an ILD is. Section 3.2 then outlines how basic traffic parameters such as flow and occupancy can be measured using this sensor. ILDs are prone to outputting erroneous data. The causes of this erroneous data are discussed in section 3.3. Section 3.4 provides a literature review identifying existing methods undertaken to identify erroneous ILD data. Section 3.5 then proposes a new ILD data filtering method called the Adjacent Detector Test. Section 3.6 proposes a novel way to quantify the performance of each ILD data cleaning treatment. This leads to the identification of a proposed quality control methodology for ILD data which is presented in section 3.7. This methodology will be used in the rest of the thesis to filter ILD data used in the ULTT models developed in chapters 7 and 8, and in the research into travel time variability presented in chapter 10.

3.1 Inductive Loop Detectors

Single loop ILDs are widely used to collect flow and occupancy data (dual loop ILDs are two single loop ILDs spaced several metres from one another, and are capable of directly measuring the speed of a vehicle). ILDs are wires laid in 2m long by 1.5m wide loops, several centimetres under the road. They are connected to a power source, which applies an oscillating voltage across the wires. The oscillating current that is produced generates a magnetic field around the ILD. If a metallic object is placed above the ILD then eddy currents are created in this metallic object. This causes the amplitude of the current within the ILD to decrease. If the current decreases beyond a certain threshold then the
presence of a metallic object is signalled by the output of a binary ‘1’ bit. If no vehicle is detected then a binary ‘0’ bit is output. Most ILDs in the UK operate at a polling frequency of 4Hz (Walmsley 1982) - that is they check for the presence of a vehicle every 250 milliseconds. In the USA ILDs tend to operate at a polling frequency of 60Hz (Hu et al. 2001). It should be noted that there are three types of frequency associated with an Inductive Loop Detector:

- **supply frequency** - the frequency of the mains power supply. This is approximately 50Hz in Europe and 60Hz in North America.

- **loop frequency** - in order to generate a magnetic field which will interact with the metal in the vehicles, it is necessary to apply an oscillating voltage across the ILD. The frequency of this oscillation is known as the loop or oscillating frequency. This frequency is typically of the order 10 kHz to 200 kHz (Klein & Kelley 1996).

- **polling frequency** - this is the number of times per second that the loop checks for the presence of a vehicle.

Care should be taken not to confuse the three. In this thesis, the frequency referred to is the polling frequency.

### 3.2 Basic Traffic Parameters obtained from ILDs

This section outlines how the basic traffic parameters (density, flow, and speed) are obtained from single ILDs (Cherrett et al. 2000).

#### 3.2.1 Volume and Flow

Every time a vehicle (i.e. a large metallic object) is situated above the ILD, the detector returns a 1-bit. When there is no vehicle over the ILD, a 0-bit is returned. Thus ILDs can easily measure volume. This is done by counting the number of 0-bit to 1-bit transitions in the series of data. Flow is simply the volume per unit time. It is formalised by equation 3.1.

\[
q = \frac{D_{01}}{T}
\]

\[q = \text{flow}\]

\[D_{01} = \text{Number of 0-bit to 1-bit transitions measured by the detector - i.e. the volume}\]

\[T = \text{Duration of time, in hours, that the count took place.}\]

A special case is the capacity flow and saturation flow. Capacity flow is the maximum possible flow on a road such that any queue at the downstream intersection will not grow over time. Saturation flow refers to the maximum possible flow should any traffic lights
3.2. Basic Traffic Parameters obtained from ILDs

on the link be permanently set to green. Saturation flow is thus greater or the same as capacity flow.

### 3.2.2 Occupancy

The occupancy gives the fraction of time that a vehicle is covering the ILD. This is formalised by equation 3.2.

$$o = \frac{D_1}{D_0 + D_1}$$

\(o\) = occupancy

\(D_0\) = Number of 0-bits recorded by the detector

\(D_1\) = Number of 1-bits recorded by the detector

### 3.2.3 Density

The occupancy measurement outlined above gives a proxy measure of density. Density is given approximately by equation 3.3. It assumes that the flow is stationary in time (Daganzo 1997). The definition of density in periods where the flow is not stationary in time is discussed in further detail in appendix A.1.

$$k \simeq \frac{o}{l}$$

\(k\) = density

\(o\) = occupancy

\(l\) = average effective length of vehicles in the traffic stream.

### 3.2.4 Average Loop Occupancy Time Per Vehicle (ALOTPV)

The average loop occupancy time per vehicle (ALOTPV), gives a measure of the combined effects of speed and length of the vehicle. It gives the average time (in bits) that a vehicle spends over the detector. It is found by dividing the number of ‘1-bits’ by the volume. This can be formalised by equation 3.4.

$$ALOTPV = \frac{D_1}{D_{01}}$$

### 3.2.5 Average Gap Time Between Vehicles (AGTBV)

The average gap time between vehicles (AGTBV), gives the average time gap between the back of the leading and front of the following vehicle. In the literature this metric has also
been referred to as the average headway time between vehicles (AHTBV) (Cherrett et al. 2000).

\[ AGTBV = \frac{D_0}{D_{01}} \]  

(3.5)

3.2.6 Headway

The average headway is the average time duration between the same points of successive vehicles. (i.e. from the front of one vehicle to the front of the following vehicle). This is subtly different from the AGTBV presented in section 3.2.5. It can be formalised by equation 3.6.

\[ AH = \frac{1}{n-1} \left[ \sum_{j=0}^{n-1} (f_{j+1} - f_j) \right] \]  

(3.6)

AH = Average headway between vehicles

\( f_{j+1} \) = Time that the front of the second vehicle passes the rear end of the ILD

\( f_j \) = Time that the front of the first vehicle passes the rear end of the ILD

\( n \) = number of vehicles in the sample - there are \( n-1 \) headways

It should be noted that \( f_{j+1} \) and \( f_j \) can collectively be replaced with the time that the back of the vehicle passes the front of the ILD.

3.3 Errors associated with ILD data

There are many possible sources of error that may lead to the calculation of erroneous values of the basic traffic parameters outlined above. It is important to identify these sources of error so that suitable error detection algorithms can be developed. This section identifies the errors involved, firstly, in collecting the raw ILD data, and secondly, in converting the raw ILD data into the basic traffic parameters. The next section will then present a literature review of the algorithms used to identify these errors.

3.3.1 Errors collecting the raw ILD data

If operating correctly, the ILD should output to the traffic management centre (TMC) a 1-bit when a vehicle is over the detector, and a 0-bit otherwise. However, this may not always be the case resulting in erroneous values of occupancy and flow being calculated.

In a study in London, Head (1982) found that less than 50% of all vehicle actuated ILDs were operating correctly. The main causes of failure were given as:

- Insulation damage
- Damage to cables crossing the kerb and in the footway of verges
3.3. Errors associated with ILD data

- Joint failures
- Detectors out of tune - 29% of all errors were found to be caused by this.
- Detector electronic failures
- Roadworks

This work led to new guidelines in installing and maintaining ILDs which has seen their accuracy improve. If ILDs are maintained properly they are the most accurate commonly available sensors to measure flow. Klein & Kelley (1996) report research that compared the flow measured by an ILD with that calculated from manual processing of a video tape. The ILD was shown to have a count accuracy of 99.4%. Nevertheless, ILDs still output erroneous data. Table 3.1 highlights the errors and the characteristics of these errors that may result in the TMC receiving faulty raw data.

<table>
<thead>
<tr>
<th>Error Name</th>
<th>Description and Characteristics of error</th>
<th>Test for Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILD broken cable</td>
<td>ILD not outputting any data. All zero output recorded</td>
<td>All zero output.</td>
</tr>
<tr>
<td>ILD chattering</td>
<td>An ILD incorporates an analogue to digital converter. If the analogue signal varies slightly around the threshold, the ILD will rapidly change from an ‘on’ to an ‘off’ state. This is often caused by crosstalk.</td>
<td>(1) Minimum on pulse time, (2) Maximum flow</td>
</tr>
<tr>
<td>ILD Crosstalk</td>
<td>Interaction between ILD and other electrical devices causes ILDs to output erroneous 1-bits.</td>
<td>Minimum on pulse time.</td>
</tr>
<tr>
<td>ILD desensitised</td>
<td>ILD gets out of tune, particularly when the wire becomes wet, reducing the threshold at which a vehicle presence is determined (Head, 1982). The occupancy recorded is more sensitive to this than the flow recorded.</td>
<td>(1) Check against historical data (preferably data following recalibration)</td>
</tr>
<tr>
<td>ILD hanging</td>
<td>Sensors can get stuck in either an ‘on’ or ‘off’ position.</td>
<td>(1) Maximum on and off pulse time, (2) Minimum entropy</td>
</tr>
<tr>
<td>ILD pulsing</td>
<td>Changes in a vehicle’s inductance over its length may lead to two pulses being recorded for the same vehicles</td>
<td>None available</td>
</tr>
<tr>
<td>ILD recovery time</td>
<td>After a period of sustained occupancy the ILD takes time to return to its original sensitivity. This recovery time may lead it to not being able to detect a separate vehicle following another vehicle.</td>
<td>None available</td>
</tr>
<tr>
<td>Communications failure</td>
<td>Communication link between ILD and TMC broken. No raw data obtained.</td>
<td>(1) All zero output.</td>
</tr>
</tbody>
</table>
3.3. Errors associated with ILD data

The UTC software is commonly rebooted (once a week). During this out-time it will not be possible to store the data.

Parked Vehicles

Vehicles parked over the detector cause the ILD to constantly output ‘1-bits.’

Software error

Software occasionally has bugs, and must be fully tested to find and rectify them.

Table 3.1: Errors collecting the raw ILD data

<table>
<thead>
<tr>
<th>Error Name</th>
<th>Description and Characteristics of error</th>
<th>Test for Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polling Frequency</td>
<td>Data is not continuous but refers to a discrete unit of time. ILDs measure presence only once every 0.25 seconds resulting in a poor precision estimate of occupancy when vehicles are travelling fast. In addition this low polling frequency may lead to an undercount of the number of vehicles. This could occur if the time gap between successive vehicles (see equation 3.5) were less than 0.25 seconds. However this is unlikely, even in congested conditions as leaving such a small time gap between vehicles is likely to lead to accidents between vehicles.</td>
<td>None available</td>
</tr>
<tr>
<td>Discretisation</td>
<td>The basic traffic parameters are aggregated to ‘x’ minutes. Due to non-stationary traffic conditions, the average value of flow and occupancy calculated may not be representative of conditions at the start or end of this period.</td>
<td>None available</td>
</tr>
</tbody>
</table>

3.3.2 Errors in calculating the basic traffic parameters

Once the raw ILD data has been received at the TMC it is a simple process to calculate the occupancy and flow. However, even in this process, errors are present which may result in the estimated traffic parameters being incorrect. Table 3.2 highlights the errors and the characteristics of these errors. Of course, if the raw ILD data is faulty in the first place, then the flows and occupancies calculated are also likely to be faulty.
3.4 Existing quality control methods for ILD data

The above section has indicated that there are many possible sources of error when using ILD data. This section gives a literature review of methods used to identify ILD data which is likely to be erroneous. This is important since outlier data could severely influence any other metric derived from using these basic traffic parameters. This section begins by first looking at univariate methods used to identify outliers. This is followed by a section on the use of multivariate techniques. The section then discusses how the literature reports the quality of the raw data to the end user.
3.4 Existing quality control methods for ILD data

3.4.1 Univariate Tests

Univariate tests of data quality look at each attribute independently of all other attributes. They are the simplest to perform, and are widely used by many traffic management centres (TMCs) to help ensure the quality of the traffic data they collect. Nevertheless, Cleghorn et al. (1991) report that in a survey undertaken, more than 1/3rd of all TMCs did not have any data checking procedure at all. Table 3.3 presents a summary of various univariate tests that can be carried out to verify the quality of the data. Many of the above tests are self-evident and do not need any explanation. However, a couple of tests mentioned, merit further explanation.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Tests on the variable</th>
</tr>
</thead>
</table>
| Flow       | (i) There must be a minimum and maximum flow within a time period. These limits may be based on historical data  
            (ii) No count within a certain time period  
            (iii) Check against historical data  
            (iv) Comparison against manual counts  
            (v) Hanging tests - i.e. ensure that same value of flow is not repeated over many time periods. |
| Occupancy  | (i) There must be a minimum and maximum occupancy within a certain time period. For example, in Minnesota the minimum occupancy in a 5-minute period is 3% and the maximum occupancy is 80%, although in Canada these figures are far more liberal at 1% and 95% respectively (Chen & May 1987).  
            (ii) Check against historical data  
            (iii) Hanging tests  
            (iv) Entropy tests |
| Pulse Length | Pulses must have a minimum and a maximum length. NB. American ILDs often run at 60 Hz so each on bit represents 16.7ms. It is unreasonable for a car of length 6m to pass over the loop in two such time periods (32.3ms). However it would not be unreasonable for a car to travel over a loop in 500ms, so it is quite likely in London, where ILDs run at 4Hz. (nb. if a single ‘1-bit’ is recorded, on a 4Hz system, then the time taken for the vehicle to pass the loop must be between 0 and 500ms) |

Table 3.3: Quality Control checks currently undertaken on traffic data by TMCs

One difficult fault to detect is hanging, i.e. when ILDs output the identical value for several continuous time periods. If this output value lies within the thresholds, then it will not be identified as an outlier. Hu et al. (2001) proposed a hanging-test to check for the reasonableness of an ILD outputting a constant non-zero flow reading over several time periods. To do this several assumptions were made. Firstly it was assumed that the flows recorded in any 30-second time interval do not vary drastically from one another within the same 5-minute period. Secondly it was assumed that the difference between the average flow and each 30-second flow during the 5-minute period were statistically independent.
From the second assumption the 30-second flow in any 5-minute time interval is modelled as coming from a Poisson distribution. The probability, $p_x$, of observing a flow of $x$ is thus given by equation 3.8.

$$p_x = \frac{e^{-\mu} \cdot \mu^x}{x!} \quad (3.8)$$

$\mu =$ mean 30-second flow in the 5-minute period.

Therefore the overall probability of obtaining $n$ continuous readings of a flow of $x$ is given by $p^n_x$. The readings are marked as erroneous if the overall probability of having obtained $n$ continuous readings falls below a certain value, $P_{min}$, i.e.

$$p^n_x \leq P_{min} \quad (3.9)$$

Hu et al. (2001) suggested a value for $P_{min}$ of 0.0005. Thus the maximum permissible number of continuous non-zero constant values $n_{max}$, before the readings are marked as erroneous is contained by equation 3.10.

$$n_{max} \leq \log\left(\frac{P_{min}}{p_x}\right) \quad (3.10)$$

The test was devised for use on highways. In urban road networks, due to vehicles being clustered into platoons the first assumption may not be satisfied. However this increased variability in flows recorded within the same 5-minute period would make it even more unreasonable to have continuous constant non-zero values. Thus in the urban context the test of Hu et al. (2001) would output an over conservative estimate of the maximum number of continuous readings permissible.

### 3.4.2 Daily Statistics Algorithm

The PeMS loop detector archive (see section 4.2.1) uses a series of univariate tests slightly different to those used by most other TMCs (Chen et al. 2003). These tests are collectively called the Daily Statistics Algorithm (DSA). Rather than test each 5-minute aggregated record of flow or occupancy individually, they analyse a whole day’s data together. Thus the DSA outputs a diagnosis of the state of the ILD only once each day rather than once every 5 minutes (once for each sample). The DSA identifies an ILD as working correctly if the data passes the following 4 tests:

- **Test #1**: The number of samples in a day that have zero occupancy must be less than a certain threshold.

- **Test #2**: The number of samples in a day that have occupancy greater than zero and flow equal to zero must be less than a certain threshold. (Strictly speaking this is a multivariate test - see the next section).
3.4. Existing quality control methods for ILD data

- Test #3: The number of samples in a day that have occupancy greater than a certain value (PeMS uses 35%) must be less than a certain threshold.

- Test #4: The entropy of occupancy samples must be greater than a certain threshold.

The last of these measures is probably the most intriguing. The entropy measures the randomness of the sample. This statistic is commonly used in information theory. It is given by equation 3.11.

\[
E = \sum_{x:p(x)>0} \log(\hat{p}(x))
\]  

(3.11)

\(\hat{p}(x)\) = Number of times in the sample where the occupancy is equal to \(x\).

The rationale for using the entropy metric is that it is expected that the flow measured by an ILD monitoring a non-deterministic traffic stream will vary (as argued by Hu et al. 2001). Even in stationary flow, the flow will fluctuate around a constant, due to small variations in speed and vehicle composition. If a sensor is broken it will often output a constant value. If the sample has many constant values then the entropy will be low. Thus an indication of a faulty ILD is low flow entropy.

The thresholds used by the PeMS system are common for all ILDs, and not based on historical data. Typical values for each traffic parameter tend to be bi-modal, with one mode for healthy detectors and another mode for faulty detectors. Very few healthy ILDs will output values close to these thresholds. Thus the system is not very sensitive to the thresholds (Chen 2004).

Chen et al. (2003) report good performance of the DSA cleaning method in identifying erroneous ILD data. A test was undertaken on the data output from 662 loops on a highway in a single day. Of the 142 loops identified by the algorithm as faulty, only 14 loops were subsequently determined to be not faulty when plots of occupancy were checked. They reported a false negative rate (ILDs incorrectly identified as good when they were in fact bad) of only 2.7% and a false positive rate of 0% (ILDs identified as bad when they are in fact good). One limitation of the work of Chen et al. (2003) is the way their method is validated. In order to identify faulty ILDs, the results of the algorithm were checked against a manual visual inspection of the time-occupancy plots. Thus both the data for the DSA, and validation test came from the same ILD. It would be better to have an independent data source for use in validating the results. For example a video image processing (VIP) system could be used to independently estimate flow and occupancy and check these against the output of the ILD. It may be better to use data generated from a microsimulation to test various algorithms. Raw data from these microsimulations could then be deliberately corrupted to mimic various ILD faults. A pragmatic approach of assessing the performance of data cleaning treatments is given in section 3.6.
3.4. Existing quality control methods for ILD data

3.4.3 Multivariate Tests

Univariate tests consider each variable independently of one another. Such tests may not identify many outliers since they do not check whether the various variables recorded are reasonable when measured in combination with one another. Various multivariate methods have thus been proposed that do check this.

A basic multivariate test is to check whether the data collected from one ILD are consistent with those from adjacent ILDs on the network. Nihan (1997) reported the results of a study undertaken by the Washington State Department of Transportation. This study defined a metric called the storage rate which measured the rate of change in the number of vehicles in a link. Due to the principle of 'conservation of vehicles,' this value should tend to zero over time, since the number of vehicles entering the link should equal the number exiting the link. However, a fault in one of the ILDs may result in this detector consistently over or underestimating the point flow. Thus over time, the storage rate would not converge to zero. Nihan (1997) suggested that this method was very useful in identifying chattering ILDs since this effect may not be so pronounced to make the ILD fail a univariate threshold test. One negative aspect of this method is that it does not explicitly determine which ILD is faulty, since it could be any of the ILDs measuring flow into or out of the link. It also requires all lanes leading into and out of the link to have working ILDs installed. On urban streets in London this criterion will not be met since some ILDs straddle two lanes, and side roads tend not to have ILDs installed on them.

Jacobson et al. (1990) recognised that there is a relationship between flow and occupancy as shown in Figure 3.1. Their test revolved around the assumption that the ratio of flow to occupancy (a proxy indication of speed) must lie within certain bounds. These bounds depend on which of 4 occupancy zones the observation lay in.

However, this procedure has several inadequacies. Firstly, the valid region is not continuous but exhibits discrete changes at the occupancy boundaries. Secondly, this method still relies on a univariate upper-threshold limit on flow to bound the valid region.

Cleghorn et al. (1991) modified this concept to allow for a continuous upper threshold of the v/o ratio (i.e. a measure of speed). Their experience showed that nearly all combinations of flow-occupancy values are possible and hence they argued that there was no point in having a lower threshold. However, like the model by Jacobson et al. (1990) this model still relies on a univariate test on the upper limit of flow. No model has been suggested to date that allows the use of a single multivariate test without the use of a univariate test on flow.

The analysis should include an additional variable, that of time-of-day, to act as a proxy of demand and vehicle composition (this will be elaborated on in section 5.14.4). Work by Coifman (2001) has demonstrated that the average length of a vehicle on the road can vary significantly by time-of-day. In the early-morning when there is a far higher proportion of HGVs on the road, low flow, and reasonably high occupancy most probably signifies free-flowing HGVs. During the day, similar flow and occupancy readings probably indicates
3.4. Existing quality control methods for ILD data

The distribution of speed in this flow-occupancy region is thus likely bi-modal. The mode at the higher speed corresponding to early morning flow of lorries, and the low-speed mode corresponding to congested cars during the day. More generally it must be noted that the patterns of flow and occupancy will not only vary over time but also by geographical location.

Cleghorn et al. (1991) also considered robust multivariate tests for paired loops. These ILDs directly measure speed as well as flow and occupancy, and thus an observation containing 3 variables is obtained. All observations were sorted into regions with similar occupancy and flow. For each flow-occupancy region, the mean and standard deviation of the speeds was calculated using historical data. The speed of any new observation was checked against a maximum and minimum speed appropriate for an observation with similar flow and occupancy and rejected if it fell outside these bounds.

Another multivariate test based upon a basic traffic flow relationship has been proposed by Turochy & Smith (2000b,a). They use the relationship between the average effective vehicle length (AEVL), speed measured by the ILD, $u$, occupancy measured at the ILD, $o$, and the flow measured by the ILD, $q$, which is stated in equation 3.12.

$$AEVL = \frac{10 \cdot u \cdot o}{q}$$ (3.12)

The AEVL can be calculated using the speed, flow, and occupancy measured by the ILD. They then verify that the AEVL estimated from this calculation is reasonable, between 2.7m and 18m. Of course, this method can only be applied when dual loop ILDs that measure speed are used. In London, the ILDs are single loop and do not measure speed.
3.5. Adjacent Detector Test

### 3.5.1 Introduction

Nihan (1997) suggested, but did not implement, the use of the storage rate to identify faulty ILD data. Such an idea would require that each lane had its own separate ILD, that vehicles did not straddle more than one ILD, and that all entrance and exit points of the link were monitored. Such conditions are unlikely to be met on motorways, and directly. Thus it will not be possible to incorporate this test into any quality procedure used in this PhD.

### 3.4.4 Quality control statistics

A study by Turner on traffic data quality (Turner 2002) recommends that an indicator of data quality should be provided along with any raw traffic data. For example, the ARTIMIS centre in Cincinnati reports the number of 30-second data samples used to compute each 15-minute summary statistic of traffic conditions measured by the ILD. The Washington State Department of Transport (WSDOT) provides a similar metric of the number of 20-second data samples used in deriving the 5-minute summary statistic. In addition they also provide a data quality flag for each record. This flag can take the value of good, bad, suspect, or disabled loop. The analyst can then use this flag to filter out any uncertain data, or replace any erroneous raw data with imputed data.

Similarly SCOOT (see section 4.1.1) gives an indicator of the quality of the raw ILD data (Hounsell & McLeod 1990). If an ILD continuously outputs a ‘1-bit’ for 3 minutes or a ‘0-bit’ for 6 minutes, then the ILD is labelled as ‘suspect’. If an ILD continuously outputs a ‘1-bit’ or ‘0-bit’ for 30 minutes, then the ILD is labelled as ‘faulty’. In all other cases, the ILD is labelled as ‘OK’. SCOOT outputs a warning W05 message when a SCOOT link is faulty.

PeMS also presents quality control statistics on its websites for its detectors. Each ILD is given a diagnosis once a day. The possible status of the ILDs are listed in Table 3.4.

<table>
<thead>
<tr>
<th>Status</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>ILD working ok</td>
</tr>
<tr>
<td>Comm Down</td>
<td>No samples</td>
</tr>
<tr>
<td>Card Off</td>
<td>Samples are all zero</td>
</tr>
<tr>
<td>High Occ</td>
<td>Occupancies are higher than normal (Chen et al. 2003)</td>
</tr>
<tr>
<td>Constant</td>
<td>Occupancy values don’t change very much over the day and takes on only a few distinct values</td>
</tr>
<tr>
<td>Intermittent</td>
<td>No description given</td>
</tr>
<tr>
<td>Other</td>
<td>All other cases</td>
</tr>
</tbody>
</table>

Table 3.4: Possible status settings given to ILDs managed by PeMS (Chen 2004)
very unlikely to be met on urban road links where a more complex topology makes the
capturing of all vehicles entering and exiting the link difficult. However, it would still be
expected that there should be a relationship between the flow at an ILD and the flow at the
adjacent upstream and downstream ILD. It would be expected that these patterns would
change as the origin-destination (OD) patterns of the vehicles on the network changed. It
is not unreasonable to assume that these OD patterns should be similar at the same time
of day (i.e. OD patterns at 8am on a Monday are likely to be similar to those at 8am
on a Tuesday). Indeed Chen et al. (2003) has used such relationships to impute missing
ILD data. This section proposes a test, which compares the ratio of the flows at adjacent
detectors with the historical ratio at the same time of day.

3.5.2 Relationship between flows at adjacent detectors

It has been suggested that there should be a relationship between the flow recorded at
adjacent ILDs. It is possible to characterise and test the strength of this relationship as
outlined below.

Suppose the flow over two adjacent ILDs is available for each 15-minute period of each
day (i.e. a total of 96 flow records for each ILD for each day). The flow measured in the
t-th 15-minute time-period of day \(d\) at the first ILD is given by \(q_{1,d,t}\) and at the second
ILD by \(q_{2,d,t}\). The ratio between these two flows, \(r_{d,t}\) is simply given by equation 3.13.

\[
r_{d,t} = \frac{q_{2,d,t}}{q_{1,d,t}}
\]  

(3.13)

The distribution of \(r_{d,t}\) is likely to be log-normal, since a ratio can never be less than 0,
and has no upper bound. Empirical analysis using a Kolmogorov-Smirnov test presented
in figure 3.2 confirms this view (see section 4.8.2 for details about the data used to derive
this). In other words the log of the ratio is likely to come from a normal distribution. The
mean historical log ratio, \(\mu_t\), between the flows at the adjacent ILDs for the t-th 15-minute
time-period over \(D\) days is given simply by equation 3.14.

\[
\mu_t = \frac{1}{D} \sum_{d=1}^{D} \ln(r_{d,t})
\]  

(3.14)

Similarly the mean log standard deviation, \(\sigma_t\), between the flows at the adjacent ILDs
for the t-th 15-minute time-period over \(D\) days is given by equation 3.15.

\[
\sigma_t^2 = \frac{1}{D-1} \sum_{d=1}^{D} (\ln(r_{d,t}) - \mu_t)^2
\]  

(3.15)

Since \(\ln(r_{d,t})\) is likely to come from a normal distribution, the standard deviation \(\sigma_t^2\)
can then be used to create confidence intervals around \(\mu_t\) (McClave et al. 2001). As an
example of this, figure 3.3 shows the relationship between two adjacent ILDs for each of
the 96 15-minute time periods in a day (see section 4.8.2 for details about the data used
Figure 3.2: Log-normal distribution of the ratio of the flows of the two ILDs n02/253e1 and n02/055e1.
3.5. Adjacent Detector Test

Figure 3.3: Historical relationship between the two ILDs n02/253e1 and n02/055e1.

to derive this). This chart shows the mean logarithm of the ratio between the adjacent flows (blue line) as well as 95% confidence limits of this value (green line and red line). It would be expected that if both ILDs were functioning correctly then 95% of all values of $\ln(r_{d,t})$ should fall within the upper and lower 95% confidence intervals. This can then be used to identify erroneous data. For example, and referring to figure 3.3, if at time A (in the 36th 15-minute time period of the day $d_A$) the value of $\ln(r_{d,t})$ is calculated as 0.08, then since this value falls within the 95% confidence interval it can be assumed that both ILDs are functioning correctly. If however, at time B (in the 36th 15-minute time period of the day $d_B$) the value of $\ln(r_{d,t})$ is calculated as 0.33, then since this value falls outside the 95% confidence interval it can be assumed that one of the ILDs is outputting erroneous data.

As a footnote to this section, it should be stressed that the ratio of the flows and not the occupancies should be used. This is because occupancy is not only affected by the number of vehicles passing over a point but also the speed at which they pass over the point.
3.5.3 Test based on relationship between flows at adjacent detectors

The previous section suggested how the historical relationship between flows at adjacent detectors could be used to identify erroneous data. This section develops a specific test to do this.

Under assumptions that $r_{d,t}$ is log-normally distributed with mean $\mu_t$, and standard deviation $\sigma_t$, then the test statistic, $z$, can be assumed to have an $N(0, 1)$ distribution and can be calculated from equation 3.16.

$$z = \frac{\mu_t - \ln(r_{d,t})}{\sigma_t} \quad (3.16)$$

Using this test statistic, the normal cumulative distribution function can then be used to calculate the probability $p(z)$ that the current ratio of the flow at the target ILD and the flow at the adjacent ILD has the same value as the historical ratio.

Assume that a certain ILD has an adjacent downstream ILD. Let the probability that the current ratio of the flow at the target ILD and the flow at the adjacent downstream ILD comes from the same distribution as historical values be $p_d(z)$. The adjacent detector test states that both ILDs are functioning correctly if the condition stated by equation 3.17 is met.

$$p_d(z) < p_{critical} \quad (3.17)$$

$p_d = \text{probability that ratio of the flow at the target ILD and the downstream ILD comes from the same distribution as historical values}$

$p_{critical} = \text{critical probability - typically 5\%}$.  

For any target ILD, the adjacent detector test can be undertaken twice, once using the adjacent downstream detector, and secondly using the adjacent upstream detector. The performance of this new treatment will be tested against the performance of existing methods in section 8.5.

3.6 Quantifying the performance of different ILD data cleaning treatments

One noticeable shortcoming in the literature of the data cleaning treatment is the failure to quantify the improvement offered by such treatments. To the author’s knowledge only the work of Chen et al. (2003) attempts to quantify the performance of the treatment. Doubts about this method of validating the DSA method have already been raised in section 3.4.2. This section proposes a pragmatic methodology by which the performance of different ILD data treatments can be compared against one another.

The raison d’etre of cleaning raw ILD data is that erroneous records will cause poor results from models which use ILD data as input. Such inaccurate results are caused...
3.7 Proposed Quality Control Methodology for ILD data

The proposed data quality methodology for the filtering of ILD data will be based on a temporal resolution of one day. That is each ILD will be given a daily status. The determination of its daily status will depend on two tests:

- **Missing Data** - If there is any missing data then the ILD will be deemed to have failed the test. The only exception to this is if the missing data refers to a known reboot of the UTC system. SCOOT machines are rebooted once a week, resulting in an outage of around 4 hours each time.

- **DSA test** - The DSA treatment outlined in section 3.4.2 will be used to test all the data which passed the first missing data test. The thresholds used by the DSA were set by looking at graphs of the four various test statistics from several detectors. All graphs were strongly bi-modal with one mode corresponding to valid data and one mode corresponding to invalid data. It was then possible to set the threshold in the middle of these two clear modes. No records were seen to lie in this middle region. Table 3.5 lists the thresholds used for each of the four tests that comprise the DSA treatment.

These quality indicators will be stored in a database table. There will be one record for each day and for each ILD. The schema is shown in Table 3.6. This approach to ILD

<table>
<thead>
<tr>
<th>Test Number</th>
<th>Description</th>
<th>Threshold test indicating failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The percentage of samples in a day that have zero occupancy</td>
<td>&gt; 10</td>
</tr>
<tr>
<td>2</td>
<td>The percentage of samples in a day that have occupancy greater than zero and flow equal to zero</td>
<td>&gt; 10</td>
</tr>
<tr>
<td>3</td>
<td>The percentage of samples in a day that have occupancy greater than a certain value</td>
<td>&gt; 60</td>
</tr>
<tr>
<td>4</td>
<td>The entropy of occupancy samples</td>
<td>&lt; 1</td>
</tr>
</tbody>
</table>

Table 3.5: Thresholds used for the DSA test.

primarily because the model will have been calibrated using erroneous data. The erroneous data will make it difficult for the model to determine the true relationships between various parameters. Consequently it would be expected that a model trained using good data would perform better than a model trained with poor data. Thus a pragmatic approach to quantifying the performance of the different ILD data cleaning methods is to examine their effect on the performance of the ULTT model. Since ULTT models are discussed in part B of this thesis, this experiment is reported in section 8.5. These experiments clearly showed the DSA treatment to be significantly better than other treatments for the majority of ULTT models.

3.7 Proposed Quality Control Methodology for ILD data

The proposed data quality methodology for the filtering of ILD data will be based on a temporal resolution of one day. That is each ILD will be given a daily status. The determination of its daily status will depend on two tests:

- **Missing Data** - If there is any missing data then the ILD will be deemed to have failed the test. The only exception to this is if the missing data refers to a known reboot of the UTC system. SCOOT machines are rebooted once a week, resulting in an outage of around 4 hours each time.

- **DSA test** - The DSA treatment outlined in section 3.4.2 will be used to test all the data which passed the first missing data test. The thresholds used by the DSA were set by looking at graphs of the four various test statistics from several detectors. All graphs were strongly bi-modal with one mode corresponding to valid data and one mode corresponding to invalid data. It was then possible to set the threshold in the middle of these two clear modes. No records were seen to lie in this middle region. Table 3.5 lists the thresholds used for each of the four tests that comprise the DSA treatment.

These quality indicators will be stored in a database table. There will be one record for each day and for each ILD. The schema is shown in Table 3.6. This approach to ILD
3.8. Conclusion

3.8. Conclusion

This chapter has discussed the issue of ILD data quality. It firstly identified various sources of error that may result in erroneous calculations of flow and occupancy being made. Such errors can be reduced if the ILDs are maintained well. It then reviewed existing techniques that can be used to identify and filter out erroneous ILD data. There is much scope to create novel ILD data quality filters based on multi-variate analysis. One such filter, the *Adjacent Detectors Test* was suggested. In addition a methodology was proposed to quantify the performance of each data cleaning treatment. This methodology will be undertaken in section 8.5. This will show that the Daily Statistics Algorithm is the best ILD filter to use with the ILD data used in this research. The chapter ended by proposing a framework for ensuring the quality of data derived from raw ILD data. This framework will be used to clean the ILD data used in this research. The next chapter will comprehensively describe the real world datasets that have been used in this thesis.

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Key Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detector ID</td>
<td>Yes</td>
<td>Unique ID of the ILD</td>
</tr>
<tr>
<td>Date</td>
<td>Yes</td>
<td>Date of the raw data</td>
</tr>
<tr>
<td>Availability</td>
<td>No</td>
<td>Indicates the percentage of records available in the day - i.e. shows how much data was missing.</td>
</tr>
<tr>
<td>DSA flag</td>
<td>No</td>
<td>Indicator whether the ILD passed the DSA test for the day.</td>
</tr>
<tr>
<td>Version Number</td>
<td>No</td>
<td>Indicate the version number of the software used to undertake the check. If a bug is found in the software, or the definition of a valid record changed (i.e. the DSA thresholds used), it may be necessary to re-evaluate the data quality checks undertaken using the old version of the software.</td>
</tr>
</tbody>
</table>

Table 3.6: Schema of database table used to store daily ILD quality control information

data quality will be applied to the ILD data used in this thesis (see chapter 4).
Chapter 4

ILD data used in the research

The previous chapter has given an introduction to ILDs and highlighted issues concerning the quality of data from these detectors. This chapter builds on the previous chapter by outlining the specific ILD data used in this research. More generally it provides a comprehensive description of the datasets that are used in parts B and C of this thesis.

The first section of the chapter introduces the London SCOOT system that controls the traffic lights in London. All the ILD data used in this research came from ILDs connected to the London SCOOT system. One key initial aim of this research was to archive all the raw data from the ILDs. Thus the second part of the chapter reviews existing traffic archive systems. Section 4.3 introduces the London SCOOT Archive Database (LSAD). LSAD is a system that captures and stores the raw data from single loop ILDs (i.e. the 0-bit and 1-bit data measured at a 250ms temporal resolution), and then processes them into basic traffic parameters. The fourth section introduces the other SCOOT based traffic archives, Comet and ASTRID, and highlights how LSAD complements these systems.

Originally it was envisaged that LSAD would provide the raw ILD data for the ULTT models developed in part B, and the research into travel time variability of part C. However, due to reasons beyond the control of either Imperial College London or Transport for London (TfL) it was not possible to make LSAD operational until the end of the research. Thus ILD data was obtained from another TfL system. The last sections of this chapter describes the datasets that are used in the research of parts B and C.

4.1 London’s SCOOT system

4.1.1 Introduction to SCOOT

SCOOT, (Split, Cycle, and Offset Optimisation Technique), is an urban traffic control (UTC) software package that optimises traffic signal timings on a road network. Its objective is to reduce the total delay (calculated in vehicle-hours) on the part of the network that is under its control. The SCOOT system relies on a network of inductive loop detectors placed in the road for its data. These detectors are polled at a frequency of 4Hz, i.e. every 250ms, to see whether there is a vehicle above it. The calibrated
SCOOT system is then able to calculate the length of queue at each link entering the junction. Using this information it can then determine a set of signal timings to reduce the overall delay at the junction. Indeed, it is possible to include several junctions in what is known as a ‘SCOOT region’, so that the traffic signals at different junctions become coordinated, further reducing the total delay. Studies have shown SCOOT to be successful in reducing ULTT. Hunt et al. (1981) report significant reduction of travel time in the order of 6% in Glasgow and 5% in Coventry. Similarly Robertson & Bretherton (1991) report a reduction in delay of 39% during the morning rush-hour, and 48% in the evening rush-hour in Southampton. Work by Fox et al. (1998) on a road in Kingston in south west London has shown that SCOOT particulary outperforms TRANSYT when there is a large variation of flow. More information on the SCOOT queuing model will be provided in section 5.6.1, and the interested reader is referred to Hunt et al. (1981), Robertson (1986), and Robertson & Bretherton (1991).

London uses the SCOOT system to reduce delay on its road network. SCOOT was first installed in Oxford Street in 1984. The system currently controls 1,355 traffic signals in London and is the largest such system in the world (O’Sullivan 2002). There are 5,270 ILDs connected to the London SCOOT network. As will be outlined in the rest of this chapter, these ILDs will provide the data for this thesis.

4.1.2 Characteristics of ILDs in London

This section identifies three distinct features of ILDs in London; binary nature, frequency, and multi-lane spanning. A brief discussion on the impact of these characteristics on a ULTT model is given.

**ILD analogue capability**

ILDs in London output solely a ‘0-bit’ or a ‘1-bit’. They do not output the analogue value of how much the inductance has been changed; only whether the amount of inductance change is above or below a certain threshold. This feature of ILDs in London means that profiles on individual vehicles passing over the ILD cannot be determined (see section 5.5.2).

**ILD frequency**

One notable feature of the ILDs used to capture data for the London SCOOT system is that they poll the road at a frequency of only 4Hz. This differs from the frequency of ILDs in North America with which most ILD research has been conducted. These ILDs run at a frequency of 60Hz. Operating at a reduced frequency has the following impacts:

- **Precision of occupancy** - the occupancy of each vehicle can only be recorded to the nearest 250 ms (as opposed to 16.7 ms for ILDs operating at 60Hz). Thus any variable calculated from occupancy will also have poorer precision.
• **Undercounting of vehicles** - if the time period between a leading vehicle not activating the ILD and the following vehicle re-activating the ILD is below 250 ms then it is possible that the ILD will record the two vehicles as only a single vehicle and hence undercount the actual flow. To this author’s knowledge there has been no explicit research of this in London. However it will be assumed that this will not be a problem since even in congested periods, vehicles are likely to leave a time gap greater than 250 ms between vehicles. Failure to do so would likely lead to an accident between vehicles.

**ILD lane spanning**

Another feature of ILDs in London is that they often span two lanes. Figure 4.1 shows an example of this on the small stretch of the Russell Square link that has two lanes. It shows the ILD n02/253e1 (refer to figure 4.6). It can be seen that vehicles from both lanes will pass over the detector. This is not usually a problem in off-peak hours since due to the presence of parked vehicles immediately downstream of the ILD, most vehicles tend to drive in the centre of the road and there is in effect only one lane. However, in peak periods when parking restrictions apply, both lanes are likely to be used, and undercounting of vehicles is likely as discussed in table 3.2.

The ULTT models developed in part B should ideally be robust to the above three characteristics of ILDs found in London.

### 4.2 Traffic archive systems: literature review

Before outlining the archive that will be used to store the London ILD data, it is informative to review several other traffic archive systems which have been implemented around the world. Brief descriptions of some of these systems is given below. This review is not meant to be comprehensive, but rather illustrates the range of applications and data marts available. A more comprehensive review of Traffic Archive Systems in the USA is given by Martin & Wu (2003).

#### 4.2.1 PeMS

PeMS (Performance Measurement System) collects and archives 30 second aggregated loop data from ILDs on California highways. It aggregates this 30-second data into 5-minute values and then calculates the flow, occupancy, and speed for this time period (Chen, Petty, Skabardonis, Varaiya & Jia 2001). Its primary purpose is as a research tool for the study of traffic on highways, and can be accessed over the internet. ¹ A Graphical User Interface (GUI) has been written to allow the user easy access to the data. PeMS is able to generate graphs desired by the user. There is also a map to allow users to locate the

¹http://pems.eecs.berkeley.edu/ It must be noted that each user must apply for a user id before they can gain access to the system.
Figure 4.1: ILDs in London often span two lanes - ILD n02/253el on the Russell Square link
section of road and detector ID that they are looking for. The data cleaning method used in PeMS has already been outlined in section 3.4.2.

4.2.2 ADUS project, INFORM and SMIS

The ADUS (Archived Data User Services) project, funded by the Federal Highway Administration, is studying how data collected by traffic management centres (such as the INFORM system in New York), can be used to better manage the road network (Hu et al. 2001). ADUS includes two systems; INFORM and SMIS. INFORM (Information FOR Motorists) covers 45 miles of highways in New York. It gathers 5-minute aggregated data from 2800 ILDs spaced at 0.5-mile intervals, and also collects data from 125 variable message signs and 80 video cameras. The second system was SMIS (Surveillance and Motorist Information System), based in Orlando in the USA. This system currently captures 30-second aggregated data from 845 ILDs.

4.2.3 Instrumented City

The instrumented city is a traffic data archive for use by the UK academic community. It archives data from SCOOT, including: delay, flow, vehicle stops, congestion, and signal timings. This data comes from SCOOT M02, M16, and M29 messages (Bell et al. 1996). In addition air pollution data from two fixed roadside pollution monitoring units is available. However, data necessary for this thesis, namely raw ILD data (i.e. SCOOT M19 messages) and link travel time measurements are not stored. The Instrumented Facility can be accessed over the internet.  

4.2.4 ROMANSE

ROMANSE (Road Management System for Europe) is a prototype system installed in Southampton that provides real time traffic and travel information with the intention of influencing travel behaviour (Wren & Jones 1996; Morris & Cherrett 2001). Although not a data archive per se, it merits mention since it utilises many interesting modes of both collecting raw data and then distributing processed traffic data. Its main sources of information are listed below:

- SCOOT - raw loop data is able to provide estimates of flow and occupancy as well as other basic traffic parameters outlined in section 3.2.

- ARTEMIS - a small network of closed circuit television (CCTV) cameras monitor traffic at critical points. Image processing of the image enables congestion to be spotted.

Distribution of traffic information includes the following:

\(^2\)http://www.its.leeds.ac.uk/facilities/icity/index.html
4.3. LSAD

- Parking Guidance System - located on key inbound arterials, this system displays up-to-date information of availability of spaces at car parks.
- Variable Message Signs - Provides information on the state of traffic on key arterials.
- STOPWATCH - Outputs expected arrival time of buses at bus-stops.

4.2.5 Highways Agency Regional Control Centres

The England Highways Agency, which controls the motorway and trunk road network in England, are in the process of implementing 7 regional control centres. These control centres will collect traffic data from a variety of sources. This data will include static information such as roadwork information, as well as dynamic information such as flow and occupancy measured from ILDs, and travel time information derived from ANPR systems. This data will be used to measure the performance of the Highways Agency road network, as well as to disseminate real-time information to drivers over the Traffic Information Highway (TIH). More information can be obtained from Highways Agency (2005) and Bozzo (2002).

4.3 LSAD

The previous section has described traffic archive systems implemented in other cities around the world. This section introduces the London SCOOT Archive Database (LSAD) which will archive raw loop detector data from London. LSAD will be a suite of programs running on different computers in different locations. The primary purpose of LSAD is to archive all the raw ILD data - that is record the state of each ILD (i.e. 0-bit or 1-bit) every 250ms. Another integral part of LSAD is to calculate the basic traffic parameters (flow, occupancy, etc) at any desired temporal resolution. Modules can then be added to LSAD which uses the data stored in LSAD to calculate other metrics such as ULTT. This section gives a brief review of LSAD.

4.3.1 Architecture of LSAD

LSAD is comprised of 4 stages. These are listed below:

- LSAD Stage 1 - Collection and temporary storage of raw ILD data
- LSAD Stage 2 - Transfer of data from Transport for London to Imperial College London
- LSAD Stage 3 - Processing and archiving of raw ILD data.
- LSAD Stage 4 - Retrieval and Analysis of semi-processed data

The architecture of LSAD is shown in Figure 4.2.
Stage #1

- SCOOT control software
- TFL software written specifically for LSAD
- SCOOT data retrieval

Stage #2

- DVD
- Read raw unprocessed data from DVD

Stage #3

1. Re-arrange the raw data
2. Calculate summary statistics.
3. Determine the quality of the data
4. Compress and archive the data

Stage #4

- Data retrieval services
- Data analysis services
- Web Server

Saturn Sun machine at Imperial College London

Transport for London, TfL

Imperial College London

Figure 4.2: Architecture of the London SCOOT Archive Database, LSAD
4.3.2 LSAD Stage 1 - Collection and temporary storage of raw ILD data

The first stage in the LSAD system collects the raw data from the inductive loop detectors located on the road network. Every second the urban traffic control (UTC) system, polls each of the OTUs (Outstation Transmission Unit) for the data of each detector in the previous second. The UTC system will forward this data to the SCOOT system and also to stage 1 of the LSAD software. This software has been written by TfL and runs on their DEC machines. This software simply stores the data it receives into files onto a large hard-disk. There is one file for every minute’s worth of data. By using a series of small files rather than one big file, the memory requirements of the program are reduced. Indeed, it is imperative that stage 1 of the LSAD process does not consume computer resources from the DEC machine that are needed for the successful operation of the SCOOT system. Thus stage 1 of the LSAD process has been designed to be as ‘light weight’ as possible.

4.3.3 LSAD Stage 2 - Transfer of data from Transport for London to Imperial College London

The uncompressed raw-data collected by LSAD stage 1 is written onto a DVD and then sent to Imperial College London. In future it may be decided to set up a network link between TfL and Imperial College London.

4.3.4 LSAD Stage 3 - Processing and archiving of raw ILD data.

LSAD stage 3 is the largest stage in the process. It is comprised of four major parts. These are explained below. Figure 4.3 shows a flow chart of the processes in this stage.

The first part of stage 3 concerns the rearranging of the raw ILD data. The raw ILD data obtained in stage 1 is stored by time and then by detector ID. There exists one file for each minute of raw data. In each file there are 60 lines of data; one line of
data for each second. Each line of raw data contains the status of each detector in the previous second; four bits for each detector. Data is stored by LSAD stage 1 in this format because no manipulation of the data is necessary and hence minimal strain is put on the UTC machine. However, it is desired to have one data file containing data for a single detector for a single day. This file will contain a string containing 345,600 characters; one character representing the status of the ILD every 250ms. The first task of LSAD stage 3 is to rearrange the data from LSAD stage 1 into this desired format.

The next process in LSAD stage 3 is to calculate the summary statistics for each detector every 15 minutes. The summary statistics are simply the flow and occupancy explained in section 3.2. They allow the user to identify data that they may wish to use (i.e. off peak and peak traffic flow). The summary statistics are also used in the data quality checking. The summary statistics are stored in a MySQL database.

The third process in LSAD stage 3 is to determine the quality of the data. The Daily Statistics Algorithm test outlined in sections 3.4.2 and 3.7 is used to quantify the quality of the raw data. The 15-minute summary statistics are used as input to this process.

The final process in LSAD stage 3 is to archive the files created in the first process. The data is first compressed using the commonly used Zip format. The compressed file is then stored on a Sun Saturn machine purchased by the Department of Computing with the aid of a grant from HEFEC’s joint research initiative in 2001. The directory structure follows the following structure:

Root Name > Detector Name > Current Year > file_detectorName_yyyymmdd

4.3.5 LSAD Stage 4 - Retrieval and Analysis of data from the Saturn machine.

Due to reasons outlined in section 4.5 it was not possible to build stage 4 of LSAD. However, it is envisaged that Stage 4 will contain a GUI (graphical user interface) front end allowing users easy access to the raw data. In addition, the GUI will also allow access to various analytical tools. These tools could include models to estimate link travel time over a selected link, or algorithms to impute missing data. Work is currently under progress by a new PhD student to complete this stage of LSAD.

4.3.6 Validation of data stored in LSAD

If data used in LSAD is to be used as a research tool for the study of urban networks, it is necessary to have confidence in the data stored in LSAD. Therefore it is important to validate data output at all stages in LSAD.

Validation of data in stage 1

It is difficult to explicitly validate that the ILDs are outputting 1-bits and 0-bits when they should (i.e. output a 1-bit when a metallic object is above it, and a zero otherwise).
It is possible to validate the number of 0-bit to 1-bit transitions in a period of time, by checking against a manual volume count over the same period. Validating that the total number of 1-bits (i.e., a measure of occupancy) is correct is far more challenging, and to the author’s knowledge no technique for doing so has been implemented. It may be possible in future to verify the LSAD derived occupancy against the occupancy estimated by an alternative source (e.g., a VIP system). However, this would not be a satisfactory solution considering the vast amount of detectors in operation.

Validation of data in stages 2 and 3.

Stages two and three revolve around sending the raw data from TfL’s machines to Imperial College London, and then calculating basic traffic parameters from this data. An independent data source that could be used to check the operation of these three stages are the SCOOT messages. The M29 SCOOT message calculates the flow and occupancy over a 15-minute period. This could be compared with the values of flow and occupancy output by LSAD stage 3.

4.4 LSAD, ASTRID, and COMET

It is envisaged that LSAD will run in addition to two existing TfL products called ASTRID and COMET. This section gives a brief introduction of these two systems and explain how LSAD will compliment the two systems.

ASTRID, Automatic SCOOT Traffic Information Database, is a database attached to the SCOOT software that stores data used in or derived from the SCOOT traffic model. A list of data items stored by it is given below. These are available at the level of link, node, region, or route (a series of links) (Bretherton 1996). The parameters are typically available at a temporal resolution of 15-minutes. (i.e. averaged over a 15-minute period).

- Flow - flow in vehicles per hour arriving at a stop line as modelled by SCOOT.
- Delay - total delay in vehicle hours per hour.
- Congestion - the percentage of 4-second intervals when a detector is fully occupied by traffic.
- Degree of saturation - arrival rate divided by the exit rate.
- Signal settings - cycle time, stage times, offsets.
- Vehicle Delay - mean delay per vehicle estimated from the SCOOT model.
- Journey time - the sum of the vehicle delay and the pre-measured cruise time of a vehicle travelling between the SCOOT detector and the stop line.
- Speed - derived from link length, cruise time, and vehicle delay.
• Congestion Index - derived from vehicle delay and cruise time. Indicates rising or falling levels in delay on a link.

COMET is a map based information and reporting tool that integrates data from a range of existing TfL systems. This includes travel time data from the congestion charging ANPR camera, and data from ASTRID (Barton & Brown 2003).

LSAD complements the data held in ASTRID and COMET, providing the following additional benefits:

• LSAD allows the analyst to select the temporal resolution at which they wish to work with the data.

• LSAD allows the analyst to choose the method by which data is aggregated. For example, Coifman recommends the use of filters when aggregating data. (Coifman 1996).

• LSAD allows the calculation of basic traffic parameters not stored in ASTRID. These include ALOTPV, and the average gap and average headway.

• LSAD allows for platoon effects to be investigated. Work presented in appendix B.2 has suggested that platooning has an important effect on the amount of mid-link delay experienced by vehicles.

• LSAD allows data items derived by the SCOOT model to be calculated independently. For example, the values of journey time given by the SCOOT model are not very accurate, as there has been little demand from the market until recently for the SCOOT model to produce accurate measures of delay and journey time. Work by Anderson & Bell (1997) has shown that other methods such as the Kimber & Hollis model (see section A.2.4) can give more accurate travel time estimates than that predicted by the SCOOT model.

• LSAD may allow the vehicle composition of the traffic stream to be studied - in particular it may allow the proportion of long vehicles on each link to be measured.

Since ASTRID is primarily a management tool, there has not been a need for it to store raw loop detector data. There is also the issue of storage space. Previous work by this author has calculated that 6000 ILDs running at 4Hz will generate 1.94Gb of data every day. Even when compressed, this data will require 209Mb of storage space (Robinson 2001). Only recently has it become economically viable to store such large amounts of data.

4.5 ASTRID ILD data used in this thesis

Due to circumstances beyond the control of either Imperial College London and Transport for London, brought about by the introduction of Congestion Charging in central London,
4.5. ASTRID ILD data used in this thesis

Figure 4.4: Map of the location of ILDs on the Tottenham Court Road Link

The ASTRID system was able to provide 15-minute aggregated flow and occupancy data. This data was obtained for all ILDs from the Tottenham Court Road and Russell Square Links outlined in table 2.9. These two links with their component ILDs are shown in figure 4.4 and figure 4.5. Data was collected over two separate time periods: 3rd March to 13th April 2003, and 1st April to 30th June 2004. Data was chosen for these periods since ANPR data was also available for these periods.

The major limitation of having to use ASTRID data rather than LSAD data was that the aggregation period of the flow and occupancy data was fixed at 15-minutes. Thus it has not been possible in part B of the thesis to test the effect of the aggregation period on the performance of a ULTT model. For example, it has not been possible to undertake research into whether a ULTT model performs better using 5-minute or 15-minute aggregated data.
4.6 ILD data quality

The ILD data quality framework outlined in section 3.7 was performed on data from both time-periods for both links. This framework is based on the Daily Statistics Algorithms identifying those days when data from ILDs appeared to be erroneous. Table 4.1 and table 4.2 show the performance of each of the detectors on the link in these two periods. It shows the number and percentage of days that the ILD passed the DSA test.

These results suggest that ILDs in London are still unreliable and need to be maintained efficiently to keep them working. Importantly tables 4.1 and 4.2 identified which ILDs should be used to provide data in the development of the ULTT models. The ILDs that were used are stated later in this chapter.

An interesting feature concerns the performance of ILDs in Tottenham Court Road. It can be seen that those ILDs that operated well in 2003 tended not to operate well in 2004 and vice-versa. There appears to be no rationale to this, although possible explanations could include roadworks breaking the ILDs and poor maintenance.

Figure 4.5: Map of the location of ILDs on the Russell Square Link
4.7 Primary data set used to train ULTT models: Russell Square 2003

In the research of part B, it was decided to use data primarily from the Russell Square link. This was because there tended to be a greater variation in travel times between peak and off-peak times (see figure 2.9). Real world data was provided by Transport for London (TfL), for a south-bound link around Russell Square in central London. This link is a single-lane link on a two-way road (although towards the southern end of this link there is a section of road with two lanes). Data from a total of 37 separate days in March and April 2003 were used. Travel time data for each individual vehicle was obtained from Automatic Number Plate Recognition (ANPR) cameras used to enforce London’s congestion charging scheme. The cameras were sited at either end of the link. Erroneous travel time records were filtered out using the Overtaking Rule (see section 2.7). The settings of the Overtaking Rule are given below:

### Table 4.1: DSA performance of the ILDs on the Tottenham Court Road link

<table>
<thead>
<tr>
<th>Detector ID</th>
<th>2003 - 42 days</th>
<th>2004 - 91 days</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3rd March - 13th April 2003</td>
<td>1st April - 30th June 2004</td>
</tr>
<tr>
<td></td>
<td>No days passed DSA test</td>
<td>% days passed DSA test</td>
</tr>
<tr>
<td>n02_057c1</td>
<td>27</td>
<td>64.3</td>
</tr>
<tr>
<td>n02_057c2</td>
<td>27</td>
<td>64.3</td>
</tr>
<tr>
<td>n02_056g1</td>
<td>27</td>
<td>64.3</td>
</tr>
<tr>
<td>n02_058a1</td>
<td>23</td>
<td>54.8</td>
</tr>
<tr>
<td>n02_058a2</td>
<td>18</td>
<td>42.9</td>
</tr>
<tr>
<td>n02_062c1</td>
<td>5</td>
<td>11.9</td>
</tr>
<tr>
<td>n02_062c2</td>
<td>5</td>
<td>11.9</td>
</tr>
<tr>
<td>n02_060g1</td>
<td>27</td>
<td>64.3</td>
</tr>
<tr>
<td>n02_060g2</td>
<td>27</td>
<td>64.3</td>
</tr>
<tr>
<td>n02_059c1</td>
<td>27</td>
<td>64.3</td>
</tr>
<tr>
<td>n02_059c2</td>
<td>27</td>
<td>64.3</td>
</tr>
<tr>
<td>n02_019v1</td>
<td>27</td>
<td>64.3</td>
</tr>
<tr>
<td>n02_019x1</td>
<td>23</td>
<td>54.8</td>
</tr>
</tbody>
</table>

### Table 4.2: DSA performance of the ILDs on the Russell Square link

<table>
<thead>
<tr>
<th>Detector ID</th>
<th>2003 - 42 days</th>
<th>2004 - 91 days</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3rd March - 13th April 2003</td>
<td>1st April - 30th June 2004</td>
</tr>
<tr>
<td></td>
<td>No days passed DSA test</td>
<td>% days passed DSA test</td>
</tr>
<tr>
<td>n02_063e1</td>
<td>26</td>
<td>61.9</td>
</tr>
<tr>
<td>n02_052a1</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>n02_170e1</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>n02_051j1</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>n02_169e1</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>n02_253e1</td>
<td>23</td>
<td>54.8</td>
</tr>
<tr>
<td>n02_055e1</td>
<td>27</td>
<td>64.3</td>
</tr>
</tbody>
</table>
4.8 Alternative data sets

- Number of following vehicles: 20
- Tolerance time: 0 seconds (NB. No differentiation between light and heavy traffic)

For each individual vehicle link travel time record the 15-minute flow and occupancy measured at three detectors on the Russell Square link were also recorded. The three detectors were:

- n02/063e1
- n02/253e1
- n02/055e1

Data from the other 3 ILDs were not used since during this period these detectors were broken (refer to table 4.2). Thus the feature vector consisted of 6 values. Flow and occupancy data were filtered using the Daily Statistic Algorithm using the thresholds given in table 3.5. Records whose ILD data failed the DSA test were removed from the data set. The resulting data set contained 24,718 records. A plot of the cumulative distribution function (cdf) of the travel times in this dataset is shown in figure D.1 found in the appendix.

One of the 37 days was then chosen at random to provide the validation data. From this single day, 200 travel time records were chosen at random to act as the validation data set. To ensure independence between the training and validation data sets, all records from the same day as the validation data were removed from the remaining data set. A subset of 15,000 records was chosen from this remaining data set to create the training data set. Due to the random nature of selecting both the validation and historical data sets, this whole process was repeated 50 times per experiment using different validation and training sets each time. Figure 4.6 shows a detailed map of the link.

Section D.1 in the appendix provides more information characterising this data set.

4.8 Alternative data sets

The development of the ULTT models used in part B was undertaken primarily using the 2003 Russell Square dataset described in the previous section. Since this link is a single-lane link it was decided to test the performance of the ULTT models on a multi-lane link. Thus an additional data set containing 2003 data from Tottenham Court Road was generated. Tottenham Court Road is a one-way road with 3-lanes. As part of the research of part C, additional data for the Russell Square link from 2004 was also generated. This section describes these two additional datasets.

4.8.1 Tottenham Court Road - 2003

An additional data set for the training and validation of the ULTT models was created using the 2003 data from Tottenham Court Road. This data set was used to test the
Figure 4.6: Data collected from the Russell Square link in central London in 2003.

performance of the ULTT models on a multi-lane road. This data covered the same time period as that outlined in section 4.7, i.e. from 3rd March to 13th April 2003. The same fields and methodology as outlined in section 4.7 were used. Flow and occupancy data were collected from the five ILDs listed below. Thus the feature vector consisted of 10 values. The ILDs were chosen since they performed well in the DSA filtering tests (see table 4.1).

- N02/057c1
- N02/056g1
- N02/060g1
- N02/059c1
- N02/019v1

The settings of the Overtaking Rule used for this link are given in Table 4.3. A detailed map of the link is shown in figure 4.7. After the Overtaking Rule and DSA treatments methods were applied to the ANPR and ILD data, a total of 104,810 records remained. A plot of the cdf of the travel times in this dataset is shown in figure D.9.

Section D.3 in the appendix provides more information characterising this data set.

4.8.2 Russell Square - 2004

A final dataset was created using data from the Russell Square site between 1st April and 30th June 2004. From table 4.2 it can be seen that there was a marked variation in the performance of the ILDs in the space of a year. Notably, detector n02/063e1 was broken throughout the 2004 period. It was thus necessary to replace data from this detector with
4.8. Alternative data sets

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of following vehicles</td>
<td>20 vehicles</td>
</tr>
<tr>
<td>Tolerance Time - light traffic</td>
<td>70 seconds</td>
</tr>
<tr>
<td>Tolerance Time - heavy traffic</td>
<td>150 seconds</td>
</tr>
<tr>
<td>Threshold travel time differentiating light or heavy traffic</td>
<td>400 seconds</td>
</tr>
</tbody>
</table>

Table 4.3: Overtaking rule settings used on Tottenham Court Road Link

Figure 4.7: Data collected from the Tottenham Court Road in 2003.
4.9. Conclusions

Data from the nearest working detector, n02/052a1. The revised map of the link is shown in Figure 4.8.

After the ANPR and ILD data was cleaned using the Overtaking Rule and DSA respectively, a total of 98,973 individual vehicle travel time records remained. A plot of the cdf of the travel times in this dataset is shown in figure D.5. This dataset is used primarily in the research of part C of this thesis. However it is also used to test the robustness over time of the architecture of the ULTT models.

Section D.2 in the appendix provides more information characterising this data set.

4.8.3 Tottenham Court Road - 2004

From table 4.1 it can be seen that the quality of the ILD data on the Tottenham Court Road Link in 2004 was poorer than that in 2003. Although four ILDs output good data, this data was clustered around the central two ILD locations (i.e. N02/062c and N02/058a). This differed from the previous year where good ILD data was also available at either end of the link. Thus it was not possible to create a data set for this link in 2004.

4.9 Conclusions

This chapter has outlined the ILD data that is used in this research. Raw data from ILDs is stored in the London SCOOT Archive Database where data quality checks are undertaken. It is then possible to obtain the 15-minute flow and occupancy data for each link. However, due to the delayed implementation of LSAD it was necessary to obtain ILD data from ASTRID. Finally the datasets that were used in the ULTT model development, and in the study of travel time variability were described. The ULTT models will be developed in part B of this thesis.
Part B - Models

The first part of this thesis has looked at the data sources that will be used in any urban link travel time (ULTT) model. Part B of the thesis concentrates on the models themselves. Chapter 5 gives a comprehensive review of ULLT models, outlining how such a model can be characterised, describing existing models and finally identifying limitations with existing models. Chapter 6 introduces the k-Nearest Neighbors (k-NN) method, describing its theoretical background, limitations, and existing use in transport. Chapter 7 presents a methodology to optimise such a k-NN model for use as a ULLT model. Chapter 8 then compares the performance of various ULLT models on two links in central London. This chapter also considers the effect of the ILD data cleaning treatments on the performance of the ULLT Model.
Chapter 5

Review of Link Travel Time Literature

This chapter presents a broad introduction to urban link travel time (ULTT) models that use data from single loop ILDs as their primary input. It begins by defining what a ULTT model is, differentiating the various types of model from one another. Sections 5.2 and 5.3 introduce material which allows the relevant characteristics of a ULTT model to be identified and quantified. This begins in section 5.2 by a review of the components of urban link travel time. Section 5.3 then elicits some desired properties of ULTT models.

Sections 5.4 to 5.9 present a literature review on the different classes of link travel time (LTT) models. The term LTT is used since many of the models proposed in this section have been applied solely for use on highways rather than within urban environments. The various LTT models have been organised into the following 6 categories:

- Spot speed methods (section 5.4).
- Correlation methods (section 5.5)
- Queuing theory methods (section 5.6)
- Regression based methods (section 5.7)
- Artificial Neural Network methods (section 5.8)
- Emerging ULTT methods (section 5.9)

The LTT models in the literature are then summarised in section 5.10. It is shown in the literature review that there are particular difficulties of modelling turning movements and modelling mid-link delay. Approaches to doing this are studied in sections 5.11 and 5.12 respectively.

Section 5.13 and 5.14 look at what data should be input into a ULTT model. In particular, section 5.14 investigates whether the use of solely ILD data is sufficient for a ULTT model, and if not, what other data should be used. The chapter is concluded in
section 5.15, which suggests that there is much scope for the improvement of, and creating of, new ULTT models.

5.1 Definition of ULTT literature

Before continuing it is important to define what exactly is meant by a ULTT model. The ULTT model that will be developed in this thesis estimates the expected travel time over a link in the time period of interest $T_{P_{desired}}$. Inputs into the ULTT model include traffic characteristics (flow, occupancy, etc) or other variables measured in the same time period ($T_{P_{desired}}$), or static variables (e.g. road width).

There are other variations of ULTT models which will not be pursued beyond this section. One common variation of ULTT models are short term travel time forecasting models. These models tend to use as input, data from the previous $x$ time periods. This data may include estimates or measurements of the dependent variable from previous time periods. (i.e. an LTT forecasting model may include measurements of actual LTT from previous periods). Short-term travel time forecasting models are outside the scope of this thesis, although their usefulness and importance is noted. Vlahogianni et al. (2004) provides a comprehensive introduction to short-term travel time forecasting models. Another variation of ULTT model is a model whose goal is to estimate the ULTT for each individual vehicle travelling on a link. These models will have to take into account such factors as traffic light offsets, and position of vehicle in platoon, etc. To the author’s knowledge there is no specific literature on this kind of model. Such a model could be called a microstructure urban link travel time (MSULTT) model. One possible extension of the ULTT model developed as part of this thesis would be to develop such a MSULTT. However, both short-term travel time forecasting and MSULTT models are beyond the scope of this thesis and shall not be considered further.

5.2 Components of ULTT

It is helpful in the analysis of the existing ULTT models to identify the various components of ULTT. It will be seen that some ULTT models fail to model all components properly, and are therefore likely to be inaccurate when applied in practice. Much of the literature identifies two components of ULTT, namely the travel time due to delays caused by the final intersection, and the time taken to travel on the uncongested part of the link to the junction. The latter component is commonly termed the ‘cruise time.’ As will be shown in sections 5.4 to 5.9, much of the literature fails to identify another component of ULTT, namely mid-link delay. This delay is caused by buses stopping, pedestrian crossings, vehicles turning mid-link, etc. This component of ULTT is discussed in greater detail in section 5.12.

Finally it is helpful to further disaggregate the intersection delay time into each individual turning movement at the junction. It is common for the delay experienced by one
turning movement to be greater than another (e.g. vehicles turning against the oncoming traffic tend to be delayed longer than those vehicles travelling straight over the junction.) It is thus proposed to disaggregate link travel time into the following 3 components:

- **Constant-speed cruise time** - this is the time taken to drive the length of the link, assuming that the vehicle travels at a constant speed and is not affected by any incidents mid-link.

- **Mid-link delay time** - this is the delay caused by incidents occurring in the middle of the link. For example delay caused by upstream vehicles turning into a side road against the on-coming traffic, or parked vehicles blocking the carriageway, etc.

- **Turning movement intersection delay time** - this is the delay time for a specific turning movement, caused by the intersection that demarcated the end of the link.

ULTT can thus be disaggregated as shown in equation 5.1.

\[
ULTT = TT_{CS} + D_{MID} + D_{ITM}
\]  

(5.1)

\(ULTT\) = Urban Link Travel Time  
\(TT_{CS}\) = Constant speed cruise time  
\(D_{MID}\) = Mid-link delay  
\(D_{ITM}\) = Intersection turning movement delay

This desegregation of link travel time will be used in the rest of the thesis.

5.3 Characterising the performance of a ULTT model

Before presenting the ULTT literature review, it is important to determine how to characterise the performance of a ULTT model. This section looks at the desired properties of a ULTT model, and then identifies a way to quantify accuracy, which is arguably the most important criterion when choosing a ULTT model.

5.3.1 Desired properties of a ULTT model

One of the largest criticisms that can be made of the majority of the ULTT model literature is that they measure the performance of a ULTT model by considering solely the overall accuracy of the estimations produced by the model. Of course this metric is a very important, if not the most important attribute of any ULTT model. However, there are other important characteristics of a ULTT model that should be considered before selecting a ULTT model to use in a certain situation. The desired properties of a ULTT model are given in table 5.1. It should be noted that in addition to quantifying these properties, it is also desirable to know the confidence in these model estimates. For example in addition to knowing the central value of any LTT estimate, the confidence of this value is also
5.3. Characterising the performance of a ULTT model

desirable. Nevertheless, the most important measure of performance of a ULTT model is probably the overall accuracy. Ways of quantifying accuracy are presented below.

5.3.2 Quantifying Accuracy

Various metrics for quantifying accuracy have been presented already in section 2.9.2. These metrics have a common structure. The estimated value, \( \hat{x} \), is compared against the true value, \( x \). One problem in reporting accuracy of a ULTT model is how to define the travel time metric to use. This appears to have been overlooked within the literature. Should an aggregated measure of travel time be used when measuring the performance of the ULTT model or should individual vehicle travel times be used? These two possible ways of calculating the accuracy of the model are elaborated on below:

- **Individual vehicle travel times** - Calculate the accuracy of the model using the travel time of the individual vehicle which generated the travel time record. The accuracy of the model, expressed as the Mean Square Error (MSE) is given by equation 5.2.

\[
MSE_1 = \frac{1}{T_1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (\hat{x}_{j,i} - X_{j,i})^2
\]  

(5.2)

- denotes current time-period

\( J = \) total number of time-periods

\( i = \) denotes current individual travel time record in time-period \( j \).

\( n_j = \) total number of individual travel time records in time-period \( j \).

\( \hat{x}_{j,i} = \) Model estimate of the travel time of a vehicle in time-period \( j \).

\( X_{j,i} = \) travel time record of vehicle \( i \) in time-period \( j \).

and

\[
T_1 = \sum_{j=1}^{J} n_j
\]  

(5.3)

- **Aggregated travel times** - Calculate the accuracy of the model by comparing the estimated travel time against the average travel time of vehicles in the same 15 minute period. The accuracy of the model, expressed as the MSE, is given by equation 5.4.

\[
MSE_2 = \frac{1}{J} \sum_{j=1}^{J} (\hat{x}_j - \mu_j)^2
\]  

(5.4)
### 5.3. Characterising the performance of a ULTT model

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>The model should output an estimate of travel time that is close to the population mean travel time of valid vehicles ($\mu^*$) cf. equation 2.1.</td>
</tr>
<tr>
<td><strong>Continuity</strong></td>
<td>The model should continuously output estimates at the desired accuracy in all conditions (NB. some models may not work when certain conditions hold, such as in peak or off peak hours).</td>
</tr>
<tr>
<td><strong>Computational speed</strong></td>
<td>Ideally the model should be computationally quick, both when training the model and then in executing a query. In general there tends to be a trade-off between the two. Models which are computationally expensive to train (eager learning methods) are then quick when executing a query. On the other hand, lazy learning methods, which tend to have no training stage tend to be computationally demanding when executing a query.</td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
<td>The model should provide an estimate of travel time over the whole link. Some models, such as SCOOT, only output an estimate of travel time or delay on the section of roadway between the detector and downstream junction.</td>
</tr>
<tr>
<td><strong>Data Accessibility</strong></td>
<td>The model should only use input data that can be easily collected and accurately measured.</td>
</tr>
<tr>
<td><strong>Data Inclusiveness</strong></td>
<td>Ideally the model should be able to incorporate and use any relevant traffic data, be it from ILDs, GPS probe vehicles, or ANPR cameras. In the context of this work the problem is to merge high-quality low-quantity data from ANPR with high-quality low-quality ILD data.</td>
</tr>
<tr>
<td><strong>Insight</strong></td>
<td>The model should give the analyst information on how the system behaves. For example what is the sensitivity of the system to certain variables.</td>
</tr>
<tr>
<td><strong>Integrity</strong></td>
<td>The model should indicate when the estimate is likely to be erroneous or biased.</td>
</tr>
<tr>
<td><strong>Mid-link delay</strong></td>
<td>The model will incorporate mid-link incidents and delays.</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>The model should indicate the degree of refinement to which the measurement is made. For example, a distance can be stated to the nearest km, m, cm, or mm, etc.</td>
</tr>
<tr>
<td><strong>Robust</strong></td>
<td>Ideally the model should be able to adapt to any changes in the topography of the link or characteristics of road use, without need for extensive recalibration.</td>
</tr>
<tr>
<td><strong>Spatial resolution</strong></td>
<td>The model will be able to output travel time over any link of any length as desired by the practitioner.</td>
</tr>
<tr>
<td><strong>Temporal Resolution</strong></td>
<td>The model will be able to output travel time at the desired temporal resolution (e.g. an estimate every 15 minutes).</td>
</tr>
<tr>
<td><strong>Tractable</strong></td>
<td>The model should not be so complex that it is difficult to maintain. This is particularly important if the model is to be commercialised.</td>
</tr>
<tr>
<td><strong>Transferable</strong></td>
<td>The model should be able to work at different geographic sites, with minimal recalibration needed.</td>
</tr>
<tr>
<td><strong>Turning Movements</strong></td>
<td>The model will be able to give the link travel time for each turning movement at the downstream junction.</td>
</tr>
<tr>
<td><strong>Vehicle to Vehicle Distribution</strong></td>
<td>The model should provide information on the vehicle-to-vehicle distribution of the travel times on the link. (i.e. variability of travel times in a small time period of say 5 or 15 minutes when conditions are seen as stationary).</td>
</tr>
</tbody>
</table>

Table 5.1: Desired Properties of an urban link travel time model
\[ \mu_j = \text{average travel time of all vehicles in time-period } j. \] It is given by equation 5.5

\[ \mu_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{j,i} \quad (5.5) \]

Within this thesis the accuracy of ULTT models is calculated using the individual vehicle travel times. Models were evaluated in this way for the following reasons:

- This is the performance of the ULTT model as perceived by a driver of a vehicle which has generated a single travel time record.
- The raw data suggested its use - each travel time record contained the travel time of the individual vehicle and NOT the average travel time of all vehicles in the same time period.

The following proof suggests that optimising a ULTT model based on calculating accuracy using individual travel times will also optimise the ULTT model when calculating accuracy using aggregated travel times.

Expanding equation 5.2 and substituting 5.5 leads to equation 5.6:

\[
MSE_1 = \frac{1}{T_1} \sum_{j=1}^{J} \left[ \sum_{i=1}^{n_j} \left( \hat{x}_j^2 + X_{j,i}^2 - 2\hat{x}_j X_{j,i} \right) \right]
\]

\[ = \frac{1}{T_1} \sum_{j=1}^{J} \left[ n_j \hat{x}_j^2 + n_j \sum_{i=1}^{n_j} (X_{j,i}^2) - 2n_j \hat{x}_j \sum_{i=1}^{n_j} X_{j,i} \right] \]

\[ = \frac{1}{T_1} \sum_{j=1}^{J} \left[ n_j \hat{x}_j^2 + \sum_{i=1}^{n_j} (X_{j,i}^2) - 2n_j \hat{x}_j \mu_j \right] \quad (5.6) \]

Now the variance, \( \sigma_j^2 \) of the individual travel times in time period \( j \) is given by equation 5.7.

\[ \sigma_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} (X_{j,i} - \mu_j)^2 \quad (5.7) \]

Rearranging leads to equation 5.8.


\[ n_j \sigma_j^2 = \sum_{i=1}^{n_j} (X_{j,i}^2 + \mu_j^2 - 2\mu_j X_{j,i}) \]

\[ = \sum_{i=1}^{n_j} (X_{j,i}^2) + n_j \mu_j^2 - 2\mu_j \sum_{i=1}^{n_j} X_{j,i} \]

\[ = \sum_{i=1}^{n_j} (X_{j,i}^2) + n_j \mu_j^2 - 2n_j \mu_j^2 \]

\[ = \sum_{i=1}^{n_j} (X_{j,i}^2) - n_j \mu_j^2 \]  

(5.8)

Rearranging equation 5.8 leads to equation 5.9.

\[ \sum_{i=1}^{n_j} (X_{j,i}^2) = n_j \sigma_j^2 + n_j \mu_j^2 \]  

(5.9)

Substituting equation 5.9 into equation 5.6 gives 5.10

\[ MSE_1 = \frac{1}{T_1} \sum_{j=1}^{J} (n_j \hat{x}_j^2 + n_j \sigma_j^2 + n_j \mu_j^2 - 2n_j \hat{x}_j \mu_j) \]

\[ = \frac{1}{T_1} \sum_{j=1}^{J} n_j (\hat{x}_j^2 + \sigma_j^2 + \mu_j^2 - 2\hat{x}_j \mu_j) \]

\[ = \frac{1}{T_1} \sum_{j=1}^{J} n_j (\hat{x}_j^2 + \mu_j^2 - 2\hat{x}_j \mu_j) + \frac{1}{T_1} \sum_{j=1}^{J} n_j \sigma_j^2 \]  

(5.10)

In a similar fashion equation 5.4 can be expanded out giving 5.11

\[ MSE_2 = \frac{1}{J} \sum_{j=1}^{J} (\hat{x}_j^2 + \mu_j^2 - 2\hat{x}_j \mu_j)^2 \]  

(5.11)

If the number of records in each time period is a constant, \( N \), then \( n_j = N \) and \( T_1 = NJ \). Thus equation 5.10 can be rewritten as equation 5.12:

\[ MSE_1 = \frac{1}{J} \sum_{j=1}^{J} (\hat{x}_j^2 + \mu_j^2 - 2\hat{x}_j \mu_j)^2 + \frac{1}{J} \sum_{j=1}^{J} \sigma_j^2 \]  

(5.12)

Substituting equation 5.11 into equation 5.12 gives equation 5.13

\[ MSE_1 = MSE_2 + \frac{1}{J} \sum_{j=1}^{J} \sigma_j^2 \]  

(5.13)
Since the standard deviation of each period does not depend on the model attributes, then minimising $MSE_2$ will also result in minimising $MSE_1$. The case where the number of records in each time period is not constant is slightly more complicated. In these cases a model optimised to minimise $MSE_1$ will attempt to minimise the error in those time-periods with a large number of vehicles in them as opposed to time periods with few vehicles in them. Since network managers are most interested in effectively managing the network during peak periods (as opposed to at night) then a strong case can thus be made for ensuring that the metric that quantifies the accuracy of the model takes into account the fact that it is desirable to have an accurate model during these peak periods than off-peak periods. This is thus further justification for calculating the accuracy of the model using the individual vehicle travel times rather than the aggregated vehicle travel times.

Equation 5.13 suggests that if individual vehicle travel-time values are used, then the performance of the model is limited by the standard deviation of individual travel times in the same time-period. Such variability may be caused by factors such as traffic lights and other random factors. Robinson & Polak (2004) explored this further and showed that on a link with one traffic signal, having a cycle time, $C$, and a red time, $R$, the minimum values of MAE and RMSE are given by equations 5.14 and 5.15 respectively.

\[
RMSE_{TOTAL} \geq \left( \frac{R^3}{C} \left( \frac{1}{3} - \frac{R}{4C} \right) \right)^{\frac{1}{2}} \tag{5.14}
\]

\[
MAE_{TOTAL} \geq \frac{R^2}{C} \left( 1 - \frac{R}{C} + \frac{R^2}{4C^2} \right) \tag{5.15}
\]

This use of aggregated or individual vehicle travel times in the calculation of accuracy may explain why some literature reports very high accuracies for their models, whereas other literature reports poorer performance for the same model.

In this thesis the accuracy using individual vehicle travel times will be reported. This underlines the importance of comparing ULTT models using the same datasets. Such an approach will be taken in chapter 8.

5.4 Spot speed LTT models

Much of the work on developing LTT models has been focused on North American highways. These models primarily use measurements of spot speed (speed estimates at a single point in the road) to estimate LTT. This section begins by reviewing and clarifying the difference between time mean speed, space mean speed, and spot speed. It then reviews this literature and assesses its relevance for use within the urban context.
5.4. Spot speed LTT models

5.4.1 Spot-speed, time-mean-speed, and space-mean-speed

It is important to be clear on what speed the ILD is measuring. There are three main measurements of speed: spot-speed, time-mean-speed, and space-mean-speed. This section gives a brief explanation of these types of speed.

The spot speed of a vehicle is the instantaneous speed of an individual vehicle at a specific point in space.

Time-mean-speed (TMS) is defined as the average speed of all vehicles passing over a single point over a period of time $T$. It is given by equation 5.16 (Roess et al. 1998)

$$TMS = \frac{1}{n} \sum_{i} \frac{d}{t_i}$$

$n = $ number of travel times observed
$t_i = $ travel time for the $i$-th vehicle
$d = $ distance traversed

If the spot-speeds of individual vehicles at a point, $v_{\text{spot},i}$, are recorded then the time mean speed is given as the arithmetic average of these speeds.

$$TMS = \frac{1}{n} \sum_{i} v_{\text{spot},i}$$

Space-mean-speed (SMS) is defined as the average speed of all vehicles in a space of length $d$ at a single instant in time. This can be expressed in a verbose form as:

$$SMS = \frac{\text{Total distance travelled by all vehicles}}{\text{Total travel time of all vehicles}}$$

Or expressed mathematically by equation 5.19: (Roess et al. 1998)

$$SMS = \frac{nd}{\sum_{i} t_i}$$

If the spot-speeds of individual vehicles at a point, $v_{\text{spot},i}$, are recorded then the space mean speed is given as the harmonic mean of these speeds.

$$SMS = \frac{1}{\frac{1}{n} \sum_{i} \frac{1}{v_{\text{spot},i}}}$$

It can be shown that the space-mean-speed and time-mean-speed are related at the measurement point by the variance of the spot-speeds, $\sigma^2$. This is shown in equation 5.21. The equation assumes that $\sigma$ is constant over the link.

$$TMS = SMS + \frac{\sigma^2}{SMS}$$

Thus time-mean-speed will always be greater than the space-mean-speed. The only exception to this is when the spot-speeds of all vehicles are the same. In this situation the
5.4. Spot speed LTT models

5.4.1 Variance and mean-speed

The variance of the space-mean-speed will be zero and thus the time-mean-speed will be the same as the space-mean-speed.

From the perspective of link travel time estimation, it is preferred to measure the space-mean-speed, since SMS and travel time are related by equation 5.22. This assumes that SMS is constant over the whole link.

\[
\text{link travel time} = \frac{\text{link length}}{\text{SMS}} \quad (5.22)
\]

Substituting equation 5.19 in 5.22 gives equation 5.23.

\[
\text{average link travel time} = \frac{1}{n} \sum_{i} t_i \quad (5.23)
\]

Failure to identify whether time-mean-speed or space-mean-speed is being used can lead to erroneous analysis and conclusions (Coifman 2001). As a caveat to this section, it should be noted that the literature is not consistent in its use of the term ‘spot-speed.’ The definition used in this thesis follows that of Klein & Kelley (1996) who defined spot-speed as ‘the speed of an individual vehicle as it passes an observation point of the traffic stream.’ However, in other literature such as TRB (1975), spot-speed is used to refer to the time-mean speed.

5.4.2 Dual loop spot speed measures

It is important to make a distinction between single loop and dual loop ILD spot speed measurements. The ULTT model developed in this thesis will use data from single loop ILDs located in urban environments. On highways however, dual loop ILDs are commonly used, since in addition to measuring flow and occupancy they can also directly measure vehicle spot-speed and vehicle length. The method by which this is achieved is presented below.

Figure 5.1 shows a diagram of a typical dual-loop ILD. It is comprised of two ILDs; an upstream and downstream ILD. \( U_1 \) represents the point on the road at which the front
of a vehicle is located when it is first detected by the ILD. $U_2$ represents the point on the road at which the end of a vehicle is located when the ILD fails to detect its presence. These points may not be the same points as the physical start and end of the ILD. This is because an ILD may detect the presence of a vehicle before it has physically passed over the ILD. In other cases it may be necessary for a substantial proportion of the vehicle to be present over the ILD before its presence is detected. Indeed the points $U_1$ and $U_2$ may vary from vehicle to vehicle, depending upon the amount of metal content within the vehicle. Thus the effective length of the ILD is likely to be different to its physical length. $L_s$ is thus the distance between the effective downstream points of each of the upstream and downstream detectors (i.e. $D_2 - U_2$). If the time a vehicle passes points $U_1$, $U_2$ and $D_2$ are $t_{U1}$, $t_{U2}$, and $t_{D2}$ respectively then the vehicle spot speed, $v_{spot}$, is given by equation 5.24.

$$v_{spot} = \frac{L_s}{t_{D2} - t_{U2}}$$ (5.24)

Dual loop ILDs are commonly used on highways, and have been implemented on motorways in the UK as part of the MIDAS system (Mortimer 2002; White 2003). PATH (2001) give some guidelines on setting the separation distance between the two ILDs. If this distance is too large then measurements may be susceptible to lane changes, and it may also be difficult to determine which vehicle signature recorded at the downstream detectors correlates with which vehicle at the upstream detector. If the distances are too low, then the ILDs are susceptible to cross-talk (see table 3.1) and the estimate will be more sensitive to any measurement error. PATH (2001) suggest a separation distance of 9m on highways. In urban areas where speeds are far lower, then this distance could probably be reduced to reduce the difficulty of matching the signals at the two ILDs to the same vehicle. LTT can then be estimated using the vehicle spot speeds measured with the ILD. This is achieved using equations 5.20 and 5.22.

### 5.4.3 Basic single loop measures

Dual loops allow the direct measurement of vehicle spot speed. However, it is possible to estimate the spot speed of an individual vehicle passing over a single loop ILD by measuring the total time that the vehicle causes the ILD to output a 1-bit, $T_1$. This is achieved using equation 5.25.

$$v_{spot} = \frac{L_E}{T_1}$$ (5.25)

The variable $L_E$ is the effective vehicle length of a vehicle. Using the terminology presented in Figure 5.1, it is given by equation 5.26.

$$L_E = L_{LU} + L_V$$ (5.26)

Work by Cherrett et al. (2000) has shown that the effective length of a vehicle can
5.4. Spot speed LTT models

actually be less than the actual vehicle length, \( L_v \), i.e. in figure 5.1 it is possible for \( L_{LU} < 0 \).

In addition it is also possible to measure the space mean speed. This is achieved by
counting the total number of vehicles passing over the link, \( V \) and measuring the total
time the ILD outputs a 1-bit, \( T_{all} \). Space mean speed is then given by equation 5.27.

\[
v_{SMS} = \frac{V \cdot L_E}{T_{all}}
\]

(5.27)

Comparison with equation 5.18 shows that the numerator of equation 5.27 gives the
total distance travelled by the vehicles, and the denominator the total time. Equation
5.27 is commonly normalised by the total time elapsed and expressed in terms of flow (\( q \)),
and occupancy (\( o \)) as equation 5.28. Some literature refers to the reciprocal of \( L_E \) as the
g-factor (Turner et al. 1998).

\[
v_{SMS} = \frac{q \cdot L_E}{o}
\]

(5.28)

Equation 5.28 requires an estimate of the mean value for \( L_E \) to be used and this often
results in large errors when using this method. Work by Coifman (2001) has shown that
\( L_E \) can vary significantly over the course of the day from 19 to 51 feet (6 to 16 m), since
at night a larger proportion of the traffic stream consists of lorries. He suggested a way of
updating the estimate of \( L_E \) throughout the day by assuming that in free flow conditions
(i.e. occupancy less than 10\%) all vehicles travelled at the speed limit.

5.4.4 Biased estimators

Equation 5.28 assumes that all independent variables are deterministic. However, both the
measurement of occupancy and the effective length are random variables. The occupancy
measured is a random variable due to the discrete sampling nature of the ILD (for example,
in London where ILDs operate at 4Hz, a vehicle which takes 600ms to traverse an ILD, will
output either two or three sequential ‘1-bits’). Both Mikhalkin et al. (1972) and Dailey
(1999) have shown that failure to model the variables in equation 5.28 as random variables
leads to a biased estimate of spot speed.

Mikhalkin et al. (1972) took into account the discrete nature of measuring the number
of ‘1-bits’, \( D_1 \), and showed that for an individual vehicle, an approximation of an unbiased
estimate of speed is given by equation 5.29.

\[
v_{\text{spot}} = \frac{f \cdot L_E}{D_1} \left[ 1 + \frac{1}{D_1^2} + \frac{1}{D_1^4} \right]
\]

(5.29)

\( f \) = frequency at which the ILD measures presence.

Dailey (1999) showed that due to the variation in the vehicle length and individual
speeds of vehicles passing over the ILD, equation 5.28 gave a biased estimate. He showed
that given the variance in speeds of individual vehicles, \( \sigma_{\text{mean}}^2 \), an unbiased estimate of
the mean vehicle speed, \( v_{\text{mean}} \) was given by the solution of equation 5.30. Using simulated data he found that this reduced the bias in estimated speed from 3.1 miles per hour using equation 5.28 to 0.07 miles per hour.

\[
\left( \frac{\alpha}{q \cdot L_E} \right) \cdot v_{\text{mean}} + v_{\text{mean}}^2 + \sigma_{\text{mean}}^2 = 0
\] (5.30)

One criticism of this work is that Dailey does not explicitly state whether \( v_{\text{mean}} \) refers to the time mean or space mean speed. From his analysis it appears that he is referring to the time-mean-speed. It is thus not surprising that he finds the estimate of space mean speed 5.27 to be a biased estimator of the time-mean-speed.

5.4.5 Violation of assumption of homogenous SMS

The previous section has identified one possible problem with estimating link travel time using space mean speeds estimated from single or dual loop ILDs. Another greater problem with making an estimate of link travel time using spot-speed methods is in ensuring that the SMS used in the calculation is relevant to the whole link. However, the SMS presented in equations 5.24, 5.27, and 5.28 relate to the space mean speed over the very small section of road covered by the ILD. On highways speed will tend not to vary much across the whole link. However the geometry of urban road networks is far more complex, and the inclusion of traffic signals and other similar control devices, a heterogeneous vehicle fleet, and parked vehicles will often result in the speed of a vehicle varying considerably over the course of the link. For this reason it is felt that ULTT methods based on space mean speed estimates from single or dual loop ILDs are inappropriate within the urban context.

5.5 Correlation LTT models

5.5.1 Basic principles of cross correlation

The previous section has highlighted various difficulties in estimating LTT using estimates of space mean speed from one or more single loop ILDs. Another approach to using ILD data to estimate LTT has been suggested through the use of correlation techniques. These techniques work by matching vehicles detected at an upstream detector with vehicles detected at a downstream detector, and then measuring the time taken to get between the two points. The cross-correlation between two different functions, \( f(t) \) and \( h(t) \) is given by equation 5.31 (Poularikas & Seely 1991, p.104).

\[
R_{fh}(t) = f(t) \star h(t) = \int_{-\infty}^{\infty} f(\tau) h(\tau - t) d\tau = \int_{-\infty}^{\infty} f(\tau + t) h(\tau) d\tau
\] (5.31)

For example, consider figure 5.2. This shows two signals, \( f(t) \) and \( h(t) \). It can be seen that signal \( h(t) \) is the same as signal \( f(t) \), but with a time shift of two seconds. The third diagram shows the cross-correlation between the two functions. The cross-correlation function is given by a triangular function, and is seen to peak at two seconds. Thus
5.5. Correlation LTT models

Figure 5.2: Basic Cross Correlation.

The peak of the cross-correlation gives the time-gap between the two signals. Correlation techniques have thus been proposed to estimate the travel time between adjacent ILDs. Two specific types of correlation technique have been proposed, these are outlined below.

5.5.2 Signature matching

The first technique is called signature matching. Although the majority of ILDs output a series of binary 0 and 1 bits denoting whether there is a metallic object above them or not, other ILDs record the actual analogue change in inductance. Since the change in inductance is a function of the mass of metal in proximity of the detector, the inductance will vary over the length of the vehicle. Thus each vehicle will, in principle, have its own inductance signature as the amount of metal over each vehicle varies along its length. If the sampling frequency is high, say 60Hz, then signatures from various vehicles will be distinct and consistent enough to allow them to be distinguished from other vehicles. Work by Kühne suggests that 10% of the traffic volume can be matched successfully using this technique (Turner et al. 1998). One point which Turner et al. (1998) do not note is that the 10% of traffic volume that can be matched successfully may not have travel times representative of the rest of the vehicle composition. For example, it may be that those vehicles that can be matched successfully will tend to be lorries and thus more likely to travel slower than other vehicles. Signature matching may therefore be prone to bias in providing an estimate of the population mean travel time.

5.5.3 Platoon Matching

The second correlation-based technique is called platoon matching. Since most ILDs do not record the actual inductance and return only a binary 0 or 1 bit, an alternative approach is to correlate platoons of vehicles passing over the upstream and downstream detectors. Such platoon matching has been studied by Dailey (1993). Dailey models the time series of the flow (the author terms this concentration) of vehicles at a downstream detector, $q_D$, as being related to the flow of vehicles at an upstream detector, $q_U$, by the
following equation.

\[ q_D(t) = b \cdot q_U(t - \tau_0) + \int_{x_1}^{x_2} S(x, t) dx + e(t) \]  

(5.32)

\( \tau_0 = \text{mean link travel time.} \)

\( q_D(t) = \text{flow at downstream detector located at point } x_2 \text{ at time } t \)

\( q_U(t) = \text{flow at upstream detector located at point } x_1 \text{ at time } t \)

\( b = \text{dispersion factor.} \)

\( S(x, t) = \text{Changes in flow from addition or removal of vehicles along the link.} \)

\( e(t) = \text{error (Dailey terms this the noise).} \)

The first term represents the flow downstream due to the flow recorded upstream \( \tau_0 \) seconds previously. This term is modified by the dispersion factor. The second term represents the vehicles, which have entered or exited the link between the two points. The third term represents the error term. By forming a cross correlation function using equation 5.32 it is possible to find the most likely mean link travel time, \( \tau_0 \) in the fashion outlined in section 5.5.1. Using simulated data, Dailey found that accurate estimates of travel time could only be obtained with this method if there was sufficient correlation (a correlation coefficient greater than 0.4) between the two flows over the two detectors. At high occupancies, greater than 15%, this condition tends not to be achieved.

Petty et al. (1998) also proposed a ULTT model based on platoon matching. He relates the upstream and downstream flows by the following equation:

\[ q_D(t_D) = \int_{T_B}^{t_D} f(t_D - t_U)q_U(t_U)dt_U + e \]  

(5.33)

\( q_D(t_D) = \text{flow at downstream detector at time } t_D \)

\( q_u(t_U) = \text{flow at upstream detector at time } t_U \). Nb. \( t_U \leq t_D \)

\( f(t_D - t_U) = \text{probability that the vehicle has a travel time of } t_D - t_U \text{ seconds.} \)

\( T_B = \text{Start time at which analysis begins - } T_B - t_D \text{ denotes the slowest time for a vehicle to travel between the upstream and downstream detector.} \)

\( e = \text{error term} \)

The above equation states that the flow at the downstream detector at time \( t_D \), is the sum over all times \( t_U \), of the flow at the upstream detector at time \( t_U \), multiplied by the probability that the time taken to travel between the two detectors is \( t_D - t_U \). The estimated link travel-time is found by finding the most likely value of \( f(t_D - t_U) \). Equation 5.33 has been used in platoon dispersion analysis, and is the basis of the diffusion theory used by Pacey to model platoon dispersion (Seddon 1972). The method of Petty et al. (1998) was found to give good estimates of link travel time, although the estimated travel time distribution tended to be multi-modal and hence the method was prone to finding more than one possible travel time. These other modes tended to be located at times
more than double the actual travel time, so could be eliminated using intuition. Petty
also recommended the use of a low-pass filter to eliminate such spurious peaks (modes).
Once again this model was proposed for the highway environment. In an urban context
there is likely to be more noise, and it may not be so obvious to determine which mode
refers to the correct travel time. Petty et al. (1998) also showed that their model gave
essentially the same results as that of Dailey (1993).

The important aspect for all platoon matching techniques is that the microstructure
of the traffic stream remains intact. Assuming that all vehicles in the traffic stream travel
at the same speed and stay on the link, the microstructure will be broken only if there are
bottlenecks along the route. It can be shown that for a microstructure to remain intact,
the following condition between the minimum flow at the bottleneck and the upstream
occupancy must hold.

\[ o \leq \frac{q_a}{vg} \]  

\( o \) = occupancy measured upstream of the bottleneck
\( v \) = velocity of the vehicles in the platoon.
\( g \) = constant relating occupancy to density.
\( q_a \) = Flow through the bottleneck

The full derivation of this is shown in appendix B.1. This equation shows that as the
effects of the incident worsen (i.e. the flow through the bottleneck, \( q_a \), is reduced), so
the occupancy at which the microstructure breaks-down decreases. More importantly it
shows that in the urban context, where traffic lights effectively reduce the flow through the
bottleneck to zero, the microstructure will be lost. Thus the platoon matching technique
is unlikely to work in the urban context.

5.6 Queuing theory based LTT models

The first two categories of LTT models have been shown to be inadequate for use in urban
environments. As stated in section 5.2, link travel time can be disaggregated into various
components, including intersection delay. A common approach taken by the literature in
modelling intersection delay, has been through the application of queuing theory. Indeed
many of the UTC systems that have been developed to reduce total delay across the
network rely on an underlying model that determines the queues at a junction - for example
SCOOT (Hunt et al. 1981). These models can be modified to provide an estimate of ULTT.
For the benefit of certain readers, appendix A.2 presents an introduction to basic queuing
theory.
5.6. Queuing theory based LTT models

5.6.1 SCOOT and TRANSYT models

Many UTC systems use queuing models which receive input from ILDs to estimate delays on urban links. One of the most well established UTC systems is TRANSYT (TRAffic Network StudY Tool) which has been used to optimise traffic signals since 1966 when it was installed in Glasgow and on Cromwell Road just south of Imperial College London. (Robertson 1969). TRANSYT was designed to create fixed time plans for traffic signals. SCOOT (Split, Cycle and Offset Optimisation Technique) (Hunt et al. 1981) uses a similar approach to TRANSYT but was designed as a real-time traffic signal optimiser, able to adapt traffic signals dependent on current traffic patterns on the road network. Both TRANSYT and SCOOT aim to minimise a certain performance index, usually the total delay in an area. In order to measure demand on a link, SCOOT builds up what is called a cyclic flow profile. (Robertson & Bretherton 1991). The cyclic flow profile is a time series of the LPU counts (link profile units) which are calculated every four seconds. The number of LPUs in any given 4 second period is a function of the flow and occupancy measured at the ILD on the SCOOT link. On average one car travelling at normal speed is equivalent to 17 LPUs (Hounsell & McLeod 1990). Using the preset ‘link cruise time,’ which indicates the average travel time between the detector and the stop line, SCOOT is then able to calculate the number of vehicles which arrive at the junction stop line. With knowledge of the traffic signal timings and saturation flow, SCOOT can then calculate the number of vehicles in any 4-second period exiting the junction. Hence SCOOT can calculate the delay at the link. SCOOT is thus able to make an estimate of travel time on the ‘SCOOT Link’ (the section of road between the ILD and the junction stop line) by summing up the delay time with the link cruise time. In effect the SCOOT link is a D-E link, using the terminology of section 2.1.2. Section 4.1.1 has already presented results of SCOOT’s performance. Although SCOOT is effective in reducing overall delay on a network, work by Anderson & Bell (1997) suggests its estimates of link travel time to be poor. A comprehensive study on the performance of the SCOOT model carried out in Southampton by Carden et al. (1989) suggests that although the SCOOT delay estimates are within 5% of the measured delay on average, there is a standard deviation of 80% of the measured delay.

5.6.2 Other Queuing models

One criticism of the SCOOT model outlined above is that it estimates the travel time only for the SCOOT link, which is a D-E link. Assuming that a network is disaggregated into a series of D-D, or S-S links, it can be seen that the SCOOT model fails to provide spatial coverage for a proportion of the network. Other queuing models which specifically model LTT for S-S links have been developed. Takaba et al. (1991) proposed two models based on queuing theory. The first of these is called the sandglass model. Takaba disaggregated a journey into links, each of which could be modelled as a sandglass (egg timer). The
sandglass model is given below.

\[ T = \frac{N}{q^*} + \frac{L_L - L_Q}{v_{FF}} \]  \hspace{1cm} (5.35)

\( T = \) Link Travel Time  
\( N = \) Number of vehicles in the queue  
\( q^* = \) Capacity flow on the link  
\( L_L = \) Length of link  
\( L_Q = \) Length of queue  
\( v_{FF} = \) Free-flow speed

The first component on the RHS represents the time spent queuing at the intersection. The second component represents the free-flow travel time on the uncongested section of road. This model is therefore a deterministic queuing model. Using the following relationship for the density, \( k \),

\[ k = \frac{N}{L + Q} \]  \hspace{1cm} (5.36)

and then approximating the non-linear relationship between density and flow (see figure A.2) by the following linear model proposed by Usami et al. (1986),

\[ k = k_m - \alpha q^* \]  \hspace{1cm} (5.37)

\( k = \) density  
\( k_m = \) jam density  
\( \alpha = \) parameter to be estimated

travel time was given by the following formula:

\[ T = \frac{k_m L_Q}{q^*} - \alpha L_Q + \frac{L_L - L_Q}{v_{FF}} \]  \hspace{1cm} (5.38)

Takaba et al. (1991) termed this the delay-time model. The first and second terms in equation 5.38 represent the time spent moving in the congested roadway section, and the third term represents the time spent travelling at free-flow speed in the uncongested part of the roadway.

Takaba et al. (1991) tested these two models on a major arterial in Tokyo, and found the sandglass model to be slightly more accurate than the delay-time model. It was noted that this may have been due to poor estimates of the jam density, saturation flow, and free-flow speed in the delay-time model. Nevertheless the delay-time model may be easier to implement than the sandglass model, since the latter requires the queue length to be measured directly.

Anderson & Bell (1997) undertook work comparing 3 queuing theory based models.
The three models were:

- Kimber and Hollis method - see section A.2.4 in the appendix.

- Simple queuing method - this method uses the smoothed count and occupancy to determine vehicles entering the queue, and uses the split and cycle times to determine the vehicles exiting the queue.

- SCOOT model

The above work found that none of the methods were satisfactory over both the two links investigated. This work suggests that queueing models may be a poor choice for ULTT models.

A notable approach to estimate ULTT using queuing theory was taken by Shimizu et al. (2000). The total delay for each turning movement at an intersection was added to an estimate of the time taken to travel over the non-queuing part to provide an overall link travel time. Unlike other research in this area Shimizu modelled the delay for each turning movement separately. This was done by adding an offset to the delay time of vehicles turning left and vehicles turning right. The estimates of travel time over route 1 (consisting of 5 individual links) were found to be incorrect by an average of 7.0%, whilst the estimates of travel time over route 2 (consisting of 4 individual links) were found to be incorrect by only 5.4%. These errors were explained by the inability to measure the free-flow speed correctly - i.e. before the vehicle reached the queue. However the approach of Shimizu et al. (2000) required the acquisition of many input variables; 10 for each link, 9 for each intersection, as well as an assumption on the average vehicle length and average vehicle speed. It would be impractical to collect this information outside of the test environment. It has also been shown by Coifman (2001) that large errors in the estimation of travel time or speed can occur when an assumption is made about the average vehicle length. For the above reasons, the approach by Shimizu is not considered further. However, the model proposed by Shimizu should be noted since it is the closest example in the literature of a MSULTT model, which estimates the travel time for individual vehicles (see section 5.1).

### 5.6.3 Sensitivity to saturation flow

Anderson & Bell (1997) proposed that the poor performance of the queuing models was due to their sensitivity to saturation flow. This tends to be fixed in queuing models, whereas the reality suggests it tends to vary. The variation of saturation flow was borne out in a study by McDonald & Hounsell (1986). Their research showed that saturation flow is affected by environmental conditions, a decrease of 6% being observed between dry and wet conditions. Their research even suggested that saturation flow decreased with increased green time. A larger variation of saturation flow was observed by Usami et al. (1986), on a road in Tokyo. The saturation rate was observed to vary from 0.33 veh / sec
5.7. Regression based LTT models

to 0.49 veh / sec. Choosing a constant saturation flow would thus lead to an error of as much as ± 20%. The critical nature of correctly determining the saturation flow has also been recognised in the PICADY software written by TRL to determine capacities, queue length, and delay (Binning 2001).

5.6.4 Difficulty of collecting data inputs

The biggest criticism of the queuing based models is that they rely either on a large number of data inputs, or on data inputs which can be difficult to measure accurately. As alluded to above, saturation flow and capacity flow can be difficult to measure directly. Even the traffic signal settings such as cycle time, stage times, and offsets may be difficult to collect. It is certainly possible to obtain this information from SCOOT. However to obtain this data for each 5-minute cycle an M16 message is required. In London 1355 signals are controlled by SCOOT. Thus over 16,000 M16 messages would have to be generated and processed each hour. Many TMCs will not wish to impose this extra load on their UTC computers. However, if it is not possible to collect traffic signal data, it may be possible to infer these from analysing the raw ILD data. For example Fourier Analysis on a time series of 4-second aggregated flow data could be used to estimate the cycle time. Nevertheless obtaining all the necessary data required for a queuing ULTT model is likely to be expensive both computationally, and in communications costs. They are thus not considered in the comparison of ULTT models in chapter 8.

5.7 Regression based LTT models

Due to the complications of obtaining input variables required for queuing based models, simpler models have been developed trying to relate easily available traffic variables to link travel time. These relationships are formulated using regression based models. These models tend to be based on simplifications of queuing theory, so do have theoretical justification in parts. For a comprehensive introduction to regression models the reader is referred to Gujarati (1995).

One of the first papers relating characteristics of the road and traffic stream to link travel time was by Wardrop (1968). He related the average space-mean-speed to the traffic flow, width of the road, number of controlled intersections, and the split time using equation 5.39.

\[ \frac{1}{v} = \left( 31 - \frac{140}{w} - 0.02444 \frac{q}{w} \right)^{-1} + f \cdot \left( 1000 - 6.8 \frac{q}{\lambda w} \right)^{-1} + \epsilon \]  

\( 5.39 \)

\( v = \) average journey speed (miles / hour) - i.e. space mean speed
\( w = \) average carriageway width (ft)
\( q = \) average traffic flow on a road (pcu / hour)
\( f = \) Number of controlled intersections per mile.
\( \lambda = \) Average proportion of effective green time (split)
5.7. Regression based LTT models

<table>
<thead>
<tr>
<th>Factor</th>
<th>Effect on average journey speed of an increase in the factor.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow on the road</td>
<td>D</td>
</tr>
<tr>
<td>Proportion of dual carriageway</td>
<td>I</td>
</tr>
<tr>
<td>Density of major intersections: A major intersection is defined as an intersection such as a roundabout or traffic signals where the road under consideration does not have priority.</td>
<td>D</td>
</tr>
<tr>
<td>Accessibility: The sum of all minor intersections and density of private drives on the road. A minor road is a road that must give way to traffic on the road under consideration.</td>
<td>D</td>
</tr>
<tr>
<td>Adverse Weather (NB. the extent on the effect depends on the type of adverse weather, with ice and fog having the severest effect on the decreasing journey speed)</td>
<td>D</td>
</tr>
</tbody>
</table>

Table 5.2: Effect of various traffic and road topology characteristics on the average journey speed in a study undertaken by Freeman Fox and Associates (1972).

$\epsilon = \text{random error assumed to be independently and identically normally distributed.}$

The first term in the brackets represents the average speed of the vehicle over the link, whilst the second bracket term represents the delay per intersection. The width of the road, $w$, appears in the denominator. Width acts as a proxy to the capacity. Replacing the width, $w$, with the capacity, $c$, reveals the familiar term of traffic intensity, $\frac{q}{c}$, in the formula. Wardrop does not give any indication of the accuracy of the above model. Neither are any limits of its use given. For example as flow, $q$, tends towards zero, the delay at each intersection tends towards 0.001 second. However, basic queuing theory suggests that as the flow tends towards zero the average vehicular delay at each intersection is given by $\frac{1}{2}(1 - \lambda)^2C$ (see equation A.14). Likewise at certain flows it is possible to have a division by zero in both terms.

Freeman Fox and Associates (1972) undertook similar work to Wardrop (1968). They derived a series of equations using regression analysis to relate the journey speed on a suburban arterial with various characteristics of the road. To this author’s best knowledge there is no other additional literature which explicitly incorporates topographical factors in a link travel time model. Table 5.2 summarises a list of factors incorporated into their models and its effect on the average journey speed.

Young (1988) undertook research on establishing a relationship between loop ALOTPV and delay at signal controlled junctions. It was found that at high occupancies there was a linear relationship between occupancy and delay, so long as a detector was placed on the road every 40m and queuing did not extend permanently beyond the most upstream detector. No theoretical rationale was given for this behaviour. However from equation
5.7. Regression based LTT models

It is seen that occupancy is inversely proportional to speed and hence proportional to travel time. Thus as the occupancy increases it would be expected for the travel time to also increase.

Gault & Taylor (1981) derived two models for determining link travel time. Their first model, termed the ‘Arrival Time model,’ related UTLT with the time the vehicle arrived at the detector and the average occupancy over the previous 12.5 seconds. It is formulated below.

\[ t = (1 - \sigma) a t^* + g \phi^{1.6} + K + \epsilon \]  

(5.40)

where

\[ a = a_0 + a_1 f + a_2 T_0 + a_3 x \]  

(5.41)

\[ g = g_0 + g_1 f + g_2 T_0 + g_3 x \]  

(5.42)

\[ K = k_0 + k_1 f + k_2 T_0 + k_3 x \]  

(5.43)

\[ t \] = travel time over link

\[ t^* \] = Time relating arrival time at the ILD to the beginning of the red phase of the traffic signal.

\[ \sigma \] = Heaviside step function (i.e. dummy variable). When the traffic signals are red, \( \sigma = 0 \), when the traffic signals are green \( \sigma = 1 \).

\[ T_0 \] = free flow travel time

\[ f \] = offset between the signals at each end of the test link

\[ x \] = degree of saturation (arrival flow / capacity flow)

\[ \phi \] = occupancy measured over the previous 12.5 seconds.

No standard indication of the accuracy of the regression such as the t-statistic, p-value, or \( R^2 \) value was given. This model was shown to perform poorly, often underestimating the travel time by half. It also could not be applied to situations where the average occupancy was above 50%. The 90% confidence interval for the standard error was 12.42 to 21.72 seconds.

Their second model proposed a linear-regression model, which was non-linear in its explanatory variables, relating link travel time to occupancy, degree of saturation, undelayed travel time, and downstream green ratio to upstream green ratio. This model is known in the literature as the ‘Occupancy Model’, ‘Newcastle model’ or Gault and Taylor model, and is given by the formula below:

\[ t = aO + b + \epsilon \]  

(5.44)
\[ a = a_0 + a_1 T_0 + a_2 x + a_3 g \]  \hspace{1cm} (5.45)

\[ b = b_0 + b_1 T_0 + b_2 x + b_3 g \]  \hspace{1cm} (5.46)

\( O = \) average detector occupancy
\( g = \) ratio of downstream split to upstream green split.
\( a_i, b_i = \) parameters to be determined.

No standard indication of the accuracy of the regression was given although an undefined value known as ‘chi-squared per point’ was stated. In simulation tests this method performed slightly better than the arrival time model, with the 90% confidence interval for the standard error being 3.11 to 9.66 seconds.

Sisiopiku et al. (1994) undertook empirical research exploring the effect that occupancy and flow rate had on travel time. They discovered that in low traffic, travel time is not sensitive to small changes in either flow or occupancy. However between occupancies of 17 and 60% travel time was linearly related to occupancy. Occupancy was found to be a better predictor of travel time than flow.

As part of the ADVANCE project, Sisiopiku & Rouphail (1994b) proposed the following model.

\[ delay = \beta_0 + \beta_1 r + \beta_2 O + \beta_3 s + \epsilon \]  \hspace{1cm} (5.47)

\( r = \) ratio of the length of the detector to the junction with the link length
\( s = \) split, i.e. green time divided by the cycle time.
\( O = \) average detector occupancy
\( T_0 = \) free flow travel time
\( \beta_i = \) parameters to be estimated

The travel time is easily found by summing the free flow travel time and the delay as shown in equation 5.48. This model is called the Illinois model.

\[ t = T_0 + delay \]  \hspace{1cm} (5.48)

One criticism of the above model is that the delay of equation 5.47 refers solely to the intersection delay, and does not include the mid-link delay term, nor does it take into account differing delays for differing turning movements.

Zhang & He (1998) proposed two models - one based on simple spot speed measurements and the other on the degree of saturation. These are written below.

\[ v_{ss} = \frac{q}{og} \]  \hspace{1cm} (5.49)
\[ v_{ss} = \text{spot speed} \]
\[ q = \text{flow} \]
\[ o = \text{occupancy} \]
\[ g = \text{constant to convert occupancy to density}. \]  
\[ \text{Zhang & He} \ (1998) \text{ used } 0.379. \]

\[ v_x = c - ae^{bx} \]  \hspace{1cm} (5.50)

\[ v_x = \text{speed based on the degree of saturation} \]
\[ x = \text{degree of saturation} \]
\[ a, b, c = \text{parameters} \]

They found that the first model performed better in heavy traffic, and the second model performed well in light traffic. The two models were thus merged into one by convex combination. This model is known as the Iowa model.

\[ v_{ls} = \lambda v_x + (1 - \lambda) \cdot v_{ss} \]
\[ 0 \leq \lambda \leq 1 \]  \hspace{1cm} (5.51)

\[ v_{ls} = \text{average link SMS}. \]
\[ \lambda = \text{Constant to be decided}. \]  
\[ \text{Zhang & He} \ (1998) \text{ suggested the following choices for lambda: } \lambda = 0 \text{ for heavy traffic}, \lambda = 1 \text{ for light traffic}, \text{ and } \lambda = \frac{1}{2} \text{ for moderate traffic}. \]

\text{Zhang & He} \ (1998) \text{ do not state how the parameters are estimated, stating only that the parameters are determined by minimising the difference between the estimated mean journey speed and observed mean journey speed. Neither did they provide any values indicating the confidence in these parameters. It is assumed however that the Classical Linear Regression Model was used. They also did not provide a way to objectively determine the traffic state, } \lambda. \]

\text{Zhang & He} \ (1998) \text{ compared the results from this model with the Newcastle and Illinois model. The models were compared with one another using the residual errors as the basis of the performance index. They found that all 3 models performed well at low speeds. At high speeds the Iowa model performed better than the other two, although it was found that the Newcastle model performed better than the Illinois model.}

Other variables that have been used in similar regression analyses include the 30-second aggregate Average Loop Occupancy Time Per Vehicle (ALOTPV) \text{(Cherrett et al. 2001). A regression model that included only this variable was shown to outperform a model that included only the 30-second occupancy data.}

Other regression models have been formulated for use on highways. For example, \text{Kwon et al.} \ (2000) \text{ used flow, occupancy, time of departure, and day of week in their regression model. The use of temporal metrics within ULTT models will be discussed in further detail in section 5.14.4.}
It is noted in conclusion that much of the regression modelling reported above is hampered by poor practices such as failure to report goodness of fit, checking for multicollinearity, indication of the significance of parameters, and analysis of the residuals. Gujarati (1995) provides a comprehensive account of undertaking regression modelling in a rigorous fashion. Nevertheless, regression models do offer a simple and quick way to model ULTT and will be included in the comparison of all ULTT models in chapter 8.

5.8 Artificial Neural Network Models

Since the early 1990’s Artificial Neural Networks, ANNs, have been widely used in transportation engineering. ANNs allow complex non-linear relationships between variables to be determined. A comprehensive review of the uses of ANNs can be found in Dougherty (1995), Fausett (1994), and Skapura (1996). Unfortunately ANNs are in many ways like a black box. Although it is possible to dissect the final model to view the relationships between variables (Dougherty 1995), this is very challenging and very rarely undertaken. Thus care should be taken when using an ANN to ensure that the output appears reasonable and that the model is operated well within the limits with which it was trained. Nevertheless, the ability of ANNs to identify patterns in data is very strong and results from these techniques have been promising. The first part of this section presents, in chronological order, the ANN ULTT literature. ANN research in other transportation fields is considered at the end of the section.

Cherrett et al. (2001) devised two feed-forward ANN models to calculate the spot-speed of a vehicle using data from a single loop ILD operating at 4Hz. The training method was not stipulated so it is assumed that a back-propagation method was used. The first of these models used individual vehicle data as input to the ANN. This included the number of 250ms time-periods the vehicle was over the ILD, as well as the time gap between the leading and following vehicle. The second ANN model used as its input 30-second aggregated average occupancy and ALOTPV data. Real world data was obtained in Southampton from a survey carried out in 1995 (Cherrett et al. 1995). The first ANN model was shown to be more accurate than the second model. The first model also outperformed two regression based models that were proposed. The second model only outperformed one of the regression models. They suggest that less commonly used ILD derived metrics such as the gap between the leading and following vehicle should be included in any ULTT model. This is further evidence to suggest that an ULTT should consider the microstructure of the traffic stream such as whether vehicles travel in platoons or not.

Cherrett et al. (2002) undertook further work comparing an ANN based ULTT model with an approach whereby the travel time between two adjacent ILDs was estimated by using the value of spot-speeds measured at the two ILDs. As before the input to the ANN model was 30-second aggregated ALOTPV data from ILDs on the link. Using a paired t-test, this research identified the ANN method as providing more accurate estimates of travel time. This was the case even when the number of ILDs supplying data to the ULTT
models was reduced to two.

Jiang & Zhang (2001) constructed an ANN using only flow information as input, whilst ANPR data was used as the source of actual travel time. The authors claim their ANN could predict the travel time accurately. However there is little theoretical background to suggest that flow on its own can determine travel time - the Lighthill Whitham Richards (LWR) model (refer to figure A.2) suggests that there are two applicable speeds for each value of flow. It was probable that the road under study was only operating in the uncongested part of the flow-speed diagram. If this is the case, this implies that the model has not been trained properly - the ANN should be trained over the full range of inputs (Skapura 1996). Empirical studies by Sisiopiku et al. (1994) also suggest there is little relationship between flow and travel time in the urban environment.

There appears to be limited research into the optimal configuration of ANNs for use as ULTT models. For example, the proposed ANN ULTT models in the literature seem to use the standard multi-layer perceptron neural network using a back-propagation algorithm. However in other areas greater research into the actual architecture of the ANN has been undertaken. In the field of traffic flow forecasting Chen, Grant-Muller, Musson & Montgomery (2001) found that a self-organised map multi-layer perceptron ANN outperformed a standard multi-layer perceptron ANN. A self-organised map multi-layer perceptron ANN is part of the family of Modular Neural Networks (MNN). A MNN is comprised of \( n \) different ANN models. When presented with input data, the MNN firstly classifies this data into one of \( n \) categories. The data is then processed by the ANN associated with this category. This approach has been used in the related field of short term travel time forecasting. Park & Rilett (1998) found that the MNN was 17% more accurate than standard ANNs. However Park & Rilett (1998) note the criticisms of the MNN models being its labour intensive nature of partitioning the input data, requiring the modeller to select a clustering algorithm, identifying the classification parameters, and deciding appropriate weight vectors. Further research by Park et al. (1999) uses a spectral basis neural network (SNN) in the field of short term travel time forecasting. A SNN model applies a non-linear transformation of the input variable before the data is entered into a standard ANN. The non-linear functions chosen for the mapping are based on sinusoidal functions. Park et al. (1999) found their model to outperform the standard ANN model as well as a model based on exponential filtering and Kalman filtering. It is suggested that further research on the use of the MNN and SNN models to model ULTT should be undertaken.

Another approach of using ANN’s for traffic flow characterisation was by Palacharla & Nelson (1995). They used flow and occupancy data as input into a fuzzy neural network (FNN). This was achieved by performing fuzzification on the input and defuzzification of the output data. A total of 16 flow ranges were defined and the extent to which any given flow value fell into each of these ranges calculated using a fuzzy membership function. Likewise each occupancy value was divided amongst 9 ranges, and each output delay time value was divided amongst 12 values. This fuzzy data was then used as input and output
of a standard ANN. Although the system found that the FNN performed better than a fuzzy expert system, no comparison was given against an ANN model where data had not undergone the fuzzification process. Thus it is not possible to indicate the benefit of using the FNN.

This chapter has suggested that ULTT models based on the use of ANNs could have a good performance. An ANN ULTT model is thus included in the ULTT model evaluation in chapter 8.

5.9 Emerging ULTT models

The models suggested so far have several limitations. Such problems with the existing literature have been noted by Vlahogianni et al. (2004) in their review on the similar problem of short-term travel time forecasting:

*A step forward to the achievement of better accuracy seems to be the use of hybrid models which are newly initiated and usually refer to combinations of parametric and non-parametric techniques.*

This section introduces some hybrid and non-parametric methods which have been suggested for use as ULTT models.

5.9.1 Fusion Techniques

The chapter has already identified many classes of ULTT model. An obvious extension is to derive a ULTT model which combines estimates from various other classes of ULTT model; for example combining the estimates of travel time obtained from a regression and spot speed model. Taking this idea one step forward, it may be possible to combine the estimate of travel time obtained from a ULTT model that uses ILD data as its input with measurements of actual link travel time derived from probe vehicles.

This approach has been taken by Xie et al. (2004). They propose the use of an ANN to combine estimates of travel time obtained from a regression-based model, and an estimate of travel time obtained from probe vehicles. A third input into the ANN is the number of probe vehicle readings. Equation 2.5 predicts that as the number of probe vehicle readings increases, so the confidence in the estimate of travel times from the probe vehicles will also increase. Thus it is desired to have a model that increases the weight of the estimate of travel times from the probe vehicles, as the number of probe vehicles increases. The architecture of the model of Xie et al. (2004) is illustrated in figure 5.3.

Using simulated data, Xie et al. (2004) find that this fusion approach provides estimates of travel time far more accurate than those obtained from using solely the probe vehicle estimator or regression based estimator. When comparing the estimated against actual value, the hybrid method had a 2-RMSE of 2.52 km/h whilst the probe vehicle and regression methods had a 2-RMSE of 4.62 km/h and 5.12 km/h respectively. However, this model has yet to be validated within the real world. Xie et al. (2004) notably omit
Figure 5.3: The hybrid ULTT model proposed by Xie et al. (2004) uses an ANN to fuse estimates of ULTT from a regression based model and travel times from probe vehicles.

Any mention of how they would filter out erroneous readings from GPS probe vehicles.

Another novel approach to fusing current probe vehicle estimates with historical data has been proposed by Chen & Chien (2001). They proposed the use of a Kalman filter to fuse these two estimates of travel-time together in their short-term forecasting model. They tested their method using real data obtained from the I-80 highway in New Jersey, USA. However, like much of the other literature they fail to compare their results of accuracy with other standard methods such as the regression methods. It is thus difficult to tell how good this model actually is. Nevertheless, the work of Chen & Chien (2001) suggests that the Kalman filter could be an ideal way to fuse data. A description of the Kalman filter is beyond the scope of this thesis, although its potential use to fuse estimates from various ULTT models together is noted. The interested reader is referred to Brown & Hwang (1992) for more information on the Kalman filter.

5.9.2 k-Nearest Neighbors techniques

The k-Nearest Neighbor approach relies on using similar historical records to estimate the appropriate output. Three instances of its use for the estimation of link travel time have been found in the literature. This is work by Handley et al. (1998), You & Kim (2000), You (2000), and Bajwa et al. (2003b,a). This method, and its use within the transport literature, including the references cited above, will be outlined in greater detail in chapter
5.10 Summary of Existing ULTT models

Sections 5.4 to 5.9 have presented a comprehensive review of the existing literature of ULTT models. Table 5.3 summarises this literature.
<table>
<thead>
<tr>
<th>Model</th>
<th>Type</th>
<th>Link Type</th>
<th>Non ILD data used</th>
<th>Turning Movement</th>
<th>Mid Link Delay</th>
<th>urban / simulated / real data</th>
<th>data cleaning, ILD</th>
<th>data cleaning, travel times</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mikhalkin et al. (1972)</td>
<td>spot-speed</td>
<td>n/a</td>
<td>none</td>
<td>Explicitly Modelled</td>
<td>Explicitly Modelled</td>
<td>highway</td>
<td>simulated</td>
<td>none</td>
</tr>
<tr>
<td>Dailey (1999)</td>
<td>spot-speed</td>
<td>n/a</td>
<td>Average length of vehicle, variability of speed</td>
<td>n</td>
<td>n</td>
<td>highway</td>
<td>real</td>
<td>none</td>
</tr>
<tr>
<td>Dailey (1993)</td>
<td>correlation</td>
<td>D-D</td>
<td>none</td>
<td>n</td>
<td>n</td>
<td>highway</td>
<td>real</td>
<td>none</td>
</tr>
<tr>
<td>Petty et al. (1998)</td>
<td>correlation</td>
<td>D-D</td>
<td>none</td>
<td>n</td>
<td>n</td>
<td>highway</td>
<td>real</td>
<td>none</td>
</tr>
<tr>
<td>Kimber &amp; Hollis (1979)</td>
<td>queuing</td>
<td>D-E</td>
<td>junction type, capacity, initial length of queue, saturation flow</td>
<td>n</td>
<td>n</td>
<td>urban</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hunt et al. (1981)</td>
<td>queuing</td>
<td>D-E</td>
<td>saturation flow, cruise speed</td>
<td>n</td>
<td>n</td>
<td>urban</td>
<td>real</td>
<td>yes - cleaned by SCOOT</td>
</tr>
<tr>
<td>Young (1988)</td>
<td>queuing</td>
<td>D-E</td>
<td>saturation flow, cruise speed</td>
<td>n</td>
<td>n</td>
<td>urban</td>
<td>real</td>
<td>none</td>
</tr>
<tr>
<td>Takaba et al. (1991)</td>
<td>queuing</td>
<td>E-E</td>
<td>capacity flow, length of queue, free flow speed, jam density</td>
<td>n</td>
<td>n</td>
<td>urban</td>
<td>real</td>
<td>none</td>
</tr>
<tr>
<td>Takaba et al. (1991)</td>
<td>queuing</td>
<td>E-E</td>
<td>capacity flow, length of queue, free flow speed, saturation flow</td>
<td>n</td>
<td>n</td>
<td>urban</td>
<td>real</td>
<td>none</td>
</tr>
<tr>
<td>Anderson &amp; Bell (1997)</td>
<td>queuing</td>
<td>D-E</td>
<td>signal settings, saturation flow</td>
<td>n</td>
<td>n</td>
<td>urban</td>
<td>real</td>
<td>yes - cleaned by SCOOT</td>
</tr>
<tr>
<td>Wardrop (1968)</td>
<td>regression</td>
<td>E-E</td>
<td>carriageway width, no of intersections per mile, average journey speed, signal settings</td>
<td>n</td>
<td>n</td>
<td>urban</td>
<td>real</td>
<td>none</td>
</tr>
</tbody>
</table>

Continued on next page...
### Table 5.3: Summary of ULTT models

<table>
<thead>
<tr>
<th>Model</th>
<th>Non ILD data used</th>
<th>Turning</th>
<th>Mid Link</th>
<th>Movement</th>
<th>Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freeman &amp; Associates (1972)</td>
<td></td>
<td>E-E</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Gault &amp; Taylor (1981)</td>
<td></td>
<td>E-E</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Gault &amp; Taylor (1981)</td>
<td></td>
<td>E-E</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Sisopiku &amp; Rouphail (1994)</td>
<td></td>
<td>E-E</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Zhang &amp; He (1998)</td>
<td></td>
<td>E-E</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Kwon et al. (2000)</td>
<td></td>
<td>E-E</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Cherrett et al. (2001)</td>
<td></td>
<td>E-E</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Jiang &amp; Zhang (2001)</td>
<td></td>
<td>E-E</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Xie et al. (2004)</td>
<td></td>
<td>E-E</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Bayas et al. (2005a)</td>
<td></td>
<td>E-E</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Bajwa et al. (2005a)</td>
<td></td>
<td>E-E</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
</tbody>
</table>

n = no, y = yes, n/a = not applicable.
5.11 Modelling turning movements

One of the major failings of many of the ULTT models described so far, is that they fail to model different turning movements. This fact has been noted by many of the authors of the models (Sisiopiku & Rouphail 1994; Anderson & Bell 1997). At many signalised intersections, vehicles turning against the oncoming flow will have to wait either for gaps in the oncoming flow, or for a special stage - usually at the end of the green stage for straight through traffic (Diep 1986). Of the models presented above, only Shimizu et al. (2000) modelled the delay for each turning movement separately. This was achieved by adding an additional term to that of the delay of vehicles going straight forward.

As shown in figure 2.10 it is possible to design the links to include the relevant turning movement. This is especially true if the network is disaggregated into D-D links. Because D-D links are used within this research the potential difficulties in modelling turning movements is not an issue. Thus turning movements are not modelled explicitly by the k-NN ULTT model proposed in chapter 7.

5.12 Modelling Mid-Link Delay

Another important limitation of many of the existing ULTT models is their failure to explicitly model mid-link delay. This component of travel time can be significant in urban areas. Examples of the causes of this type of delay are given below:

- Leading vehicles making a turn into a side road may cause following vehicles to queue behind it.
- Parked vehicles reducing the effective width of the road, forcing vehicles to move into the oncoming lane to pass them.
- Buses stopping
- Cyclists
- Road works
- Accidents

This section reviews some of the literature on modelling mid-link delay and incidents. Mid-link delay can be further divided into expected delays (such as those caused by buses stopping and vehicles turning mid-link), and incidents (accidents etc).

5.12.1 Stopping and Turning vehicles

Wong et al. (1998) looked at the delay at a signal controlled intersection with a bus-stop upstream. A simulation model was written using the MODSIM language, and validated using data from two arterial roads in Hong Kong. Total delay was modelled by adding two
5.12. Modelling Mid-Link Delay

The first one represented the queuing delay caused by the bus stopping. This was based on an M/M/1 queuing system and given by the equation 5.52 (refer also to A.2.5).

\[ D_{BSQ} = \frac{x_b}{q(1-x_b)} \]  

\( D_{BSQ} \) = Delay caused by the bus stop
\( q \) = vehicle flow
\( x_b \) = degree of saturation at the bus-stop

The second additional delay term was a correction term, CT, that modelled the interaction between the bus stop and the signal.

\[ CT = 106.5 \frac{q^{0.07} \Omega^{1.27} c^{0.57}}{L^{0.09} q^{0.53}} \]  

\( q \) = flow
\( \Omega \) = proportion of time the bus stop is blocked, i.e. dwell time multiplied by bus frequency
\( c \) = Cycle time
\( L \) = Distance of the bus stop from the intersection
\( g \) = Effective green time

It was found that the amount of delay increased as both the bus frequency and the dwell time of the bus at the bus stop increased. Delay also increased with increasing flow. As will be shown later, it is important to note the traffic arrival pattern. Wong et al. (1998) assumed that the headway distribution could be approximated by the following shifted negative exponential distribution in light traffic:

\[ h_L(t) = \frac{1}{T-\tau} exp \left( -\frac{t-\tau}{T-\tau} \right) \]  

\( h_L(t) \) = probability that headway is t seconds for light traffic.
\( T \) = Mean headway
\( t \) = Minimum headway

At heavier flows a normal distribution was assumed:

\[ h_H(t) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} exp \left( -\frac{(t-T)^2}{2\pi} \right) \]  

\( h_H(t) \) = probability that headway is t seconds for heavy traffic.
\( s \) = standard deviation of the headway distribution.

The distribution of headways at intermediate flows was given by a weighted average
of the above two distributions based on the ratio of the flow to the saturation flow.

\[ h(t) = s - \frac{q}{s} h_L(t) + \frac{q}{s} h_H(t) \]  

(5.56)

\( h(t) \) = probability that headway is \( t \) seconds for intermediate traffic.

\( s \) = saturation flow

\( q \) = flow

Bonneson (1998) undertook research on delay to major street traffic due to right-turn activity (i.e. vehicles that do not have to cross the opposing lane). His model uses Newtonian mechanics to determine the speed that a following vehicle has to decelerate to in order to avoid a collision. Using this speed, the time wasted accelerating and decelerating to this speed can be determined. This is given by the equation below.

\[ d = \frac{(v_{ff} - f_m)^2}{2v_{ff}} \left( \frac{1}{a_b} + \frac{1}{a_a} \right) \]  

(5.57)

\( d \) = delay

\( v_{ff} \) = free-flow speed

\( v_m \) = Speed that following vehicle must slow down to so as to avoid a collision

\( a_a \) = acceleration

\( a_b \) = braking deceleration

It is assumed that no vehicles have to stop since the turning vehicle will itself not stop. Bonneson assumed a shifted negative exponential distribution of headway as outlined in equation 5.54.

Both the two models of mid-link delay have been based on using a Poisson arrival flow. In urban environments, traffic lights tend to bunch vehicles together in platoons. Ideally the analysis work carried out by both Wong et al. (1998) and Bonneson (1998) would have been repeated assuming platooned arrival. Work in appendix B.2 has shown that the expected mid-link delay will increase as the time gap between vehicles decreases. A small time gap between vehicles is a feature of platoons, and hence platooning will increase the expected mid-link delay. This work also showed that the constant speed cruise time of equation 5.1 also increases when vehicles are traveling in platoons. This occurs because the desired speed of individual vehicles is not homogenous. Some vehicles want to travel faster than other vehicles. However, when vehicles are traveling in a platoon on a single-lane road where overtaking is not allowed, the speed of each vehicle is limited to the speed of the leading vehicle.

Of the ULTT models, only those proposed by Wardrop (1968) and Freeman Fox and Associates (1972) incorporate characteristics of the road topography, that could affect the amount of mid-link delay, in their model. However, since it has been shown that mid-link delay is a function of flow, then other models, notably the ANN and regression based
models, may implicitly incorporate the effects of mid-link delay when calibrating their parameters (regression) or internal weights (ANN).

5.12.2 Incident Delay

Recent research has attempted to model incidents on highways. By assuming that the occurrence of incidents on a highway are governed by a Poisson process and that the duration of each incident is Gamma distributed, Cohen & Southworth (1999) were able to formulate the mean and variance of a motorist’s delay per vehicle-mile for each class of incident.

\[
\mu_{dk} = \frac{\lambda_k C \left( m^2_k + s^2_k \right) \left( \frac{V}{C} - r_k \right) \left( g - r_k \right)}{2 \left( g - \frac{V}{C} \right)}
\]

\[
\sigma^2_{dk} = \frac{\frac{4}{3} \mu_{dk} m_k \left( 1 - \frac{r_k}{r_k} \right) \left( s^2_k - m^2_k \right)}{\left( s^2_k - m^2_k \right) - \mu^2_{dk}}
\]

\( \mu_{dk} \) = mean delay due to incident-type \( k \) (in hours per vehicle-mile)
\( \sigma^2_{dk} \) = variance of delay due to incident-type \( k \)
\( \lambda_k \) = Rate of incidents of type \( k \) per volume-length.
\( m_k \) = The mean of the Gamma distribution
\( s_k \) = The standard deviation of the Gamma distribution
\( k \) = Incident type indicator
\( r_k \) = capacity reduction factor due to incident \( k \). The effective capacity due to the incident is given by \( r \cdot C \)
\( g \) = 'getaway volume' expressed as a fraction of the capacity, indicating how quickly vehicles can exit the incident after it is cleared.
\( V \) = Average volume on the freeway
\( C \) = Capacity of the freeway

This methodology would allow all incidents, be they delay due to buses, vehicles turning right etc, to be parameterised using the 3 parameters, \( \lambda_k \), \( m_k \), and \( s_k \). It is seen that the effects of the incident are worse in congested conditions, when the degree of saturation \( \frac{V}{C} \) approaches 1. It should be noted that similarly to the equation for uniform delay at intersection, this method will result in unrealistically high results as the degree of saturation approaches the value of \( g \). Thus there is sensitivity in the equation to the value of \( g \). The value of \( g \) will be difficult to measure, and indeed Cohen & Southworth (1999) had to use a microsimulation model FRESIM to estimate it. In addition, the assumption of the Poisson and Gamma distribution of incident occurrence and duration may not be appropriate for all incident types, and this is something that should be explored further. Nevertheless the work of Cohen & Southworth (1999) offers a possible framework by which incidents could be modelled. It may be possible to extend this approach so that all mid-
link delays (be they due to stopping buses, vehicles turning mid-link) could be modelled as an incident. However, this work is beyond the scope of this thesis and will not be considered further.

5.12.3 INCIBEN and INCA

Recent work has been undertaken by the Department for Transport, UK, to enable the calculation of travel time variability (TTV) due to incidents on highways. The work developed an EXCEL spreadsheet called INCIBEN, which was then extended and renamed INCA (Mott MacDonald 2000). INCA (INcident Cost benefit Analysis), takes as its input a description of the network (links, length, lanes, capacities, and flows), models the effect of an incident, and then outputs the delay, TTV, and the costs of such delay. There are 5 types of incident. Each incident is described by the following 5 characteristics (van Vuren & Ahuja 2001):

- Rate of occurrence
- Mean incident duration
- Duration and variance weighting
- Severity (number of lanes blocked, capacity reduction factor)
- Maximum diversion proportion - i.e. proportion of vehicles which will reroute from the link to avoid the incident.

Disaggregating the incident delay into 3 periods, queue build up time, transition period, and final period, and using simple queuing theory, expressions are obtained for the total delay due to the incident. This is found to be proportional to the square of the build up time. In addition an expression for the variance of travel time is also formulated (HETA 1999).

\[
\sigma_{TOT}^2 = \sigma_{OTH}^2 + p(1-p)\mu_{INC}^2 + p^2\sigma_{INC}^2 + p(1-p)\sigma_{INC}^2
\]  

\[
\sigma_{TOT}^2 = \text{Total variance of travel time}
\]

\[
\sigma_{OTH}^2 = \text{variance of other non-incidents}
\]

\[
\sigma_{INC}^2 = \text{variance of incident durations}
\]

\[
\mu_{INC}^2 = \text{mean duration of incident}
\]

\[
p = \text{probability of an incident occurring}
\]

Each incident type is characterised as having a separate value of build up time, and a factor that takes into account the effect of the number of lanes closed. A method for deriving the build up time for each incident type is given. The model also takes into account the effect of diversions by reducing the arrival flow by a certain factor.
5.12.4 Conclusion

This section has presented several frameworks to model mid-link delay. All the models require additional data beyond that provided by ILDs. This will be investigated further in section 5.14. Nevertheless it was noted that existing regression and ANN models may already partially incorporate the effects of mid-link delay during the calibration process. The k-NN ULTT model developed in chapter 7 does not explicitly model mid-link delay, although as will be noted in section 5.14, ILD data may allow effect of incidents to be incorporated into a model implicitly.

5.13 Required data input into ULTT models

This chapter has so far highlighted the diverse range of data that could be used as input into any ULTT model. This section summarises the possible data inputs, whilst the following section investigates the effect on a ULTT model if only ILD data is made available.

Reviewing the above literature, it is possible to classify ULTT model inputs under one of the following categories:

- **TChar** - Traffic characteristic (flow, occupancy)
- **TCont** - Traffic control (traffic light settings, speed limit)
- **RNT** - Road network topography (width of road, number of traffic lights)
- **VP** - Characteristics of the vehicle population (number of bus routes, etc)

Table 5.4 presents a list of all input data variables used within this chapter. The number of possible model inputs is further increased since variables such as flow and occupancy can be calculated using many possible temporal values of aggregation. Indeed variables such as flow and occupancy from previous time-periods could be input into a ULTT model.

To the authors’ knowledge there is no explicit research which has attempted to identify those variables which help to account for the most variability in ULTT, although research into sub sectors of this have been carried out. For example Sisiopiku et al. (1994) showed that occupancy was a better predictor of ULTT than flow. The probable reason for the lack of rigorous work in this area is the great expense in creating such a data set. As section 5.3.1 stated, one of the desirable properties of a ULTT model is that all its input data should be easy to collect. For these reasons, the ULTT models used in chapter 8 will use solely data from ILDs as input. Without imposing such a restriction on the input to the ULTT models, the scope of this thesis would grow too large. However, it is acknowledged that there is a need for further research in identifying the parameters that should be input into a ULTT model.
5.14. Limitations of using data solely from ILDs

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Cat.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow</td>
<td>TChar</td>
<td>Average flow measured at an ILD.</td>
</tr>
<tr>
<td>Occupancy</td>
<td>TChar</td>
<td>Average occupancy measured at an ILD.</td>
</tr>
<tr>
<td>ALOTPV</td>
<td>TChar</td>
<td>See section 3.2.4</td>
</tr>
<tr>
<td>Average time gap</td>
<td>TChar</td>
<td>See section 3.2.5 (similar to mean headway)</td>
</tr>
<tr>
<td>Spot Speed</td>
<td>TChar</td>
<td>Spot speed of vehicles measured at an ILD</td>
</tr>
<tr>
<td>Platoon dispersion factor</td>
<td>TChar</td>
<td>See section 5.6.2</td>
</tr>
<tr>
<td>vehicles in queue</td>
<td>TChar</td>
<td>See section 5.6.2</td>
</tr>
<tr>
<td>Length of queue</td>
<td>TChar</td>
<td>See section 5.6.2</td>
</tr>
<tr>
<td>Vehicle length</td>
<td>TChar</td>
<td>Average length of a vehicle travelling over the ILD</td>
</tr>
<tr>
<td>Congestion Index</td>
<td>TChar</td>
<td>See Zhang &amp; He (1998)</td>
</tr>
<tr>
<td>Saturation flow</td>
<td>TChar</td>
<td>Saturation flow of a junction</td>
</tr>
<tr>
<td>cycle time</td>
<td>Tcont</td>
<td>Cycle time at a signal controlled junction.</td>
</tr>
<tr>
<td>Green time</td>
<td>Tcont</td>
<td>Green time for a link at a signal controlled junction.</td>
</tr>
<tr>
<td>Signal offset</td>
<td>Tcont</td>
<td>Offsets between green stages for each link at a signal controlled junction.</td>
</tr>
<tr>
<td>Speed limit</td>
<td>Tcont</td>
<td>Speed limit for vehicles on a link</td>
</tr>
<tr>
<td>Link length</td>
<td>RNT</td>
<td>Length of the link</td>
</tr>
<tr>
<td>No lanes</td>
<td>RNT</td>
<td>No of lanes on the link or the number of bus lanes</td>
</tr>
<tr>
<td>major intersections</td>
<td>RNT</td>
<td>No of major intersections on the link</td>
</tr>
<tr>
<td>minor intersections</td>
<td>RNT</td>
<td>No of minor intersections on the link</td>
</tr>
<tr>
<td>Distance bus stop</td>
<td>RNT</td>
<td>Distance of bus stop to junction</td>
</tr>
<tr>
<td>Parking bays</td>
<td>RNT</td>
<td>Indicator of whether on-road parking is allowed</td>
</tr>
<tr>
<td>Proportion buses</td>
<td>VP</td>
<td>Proportion of buses on a link</td>
</tr>
<tr>
<td>Bus stop delay</td>
<td>VP</td>
<td>Delay caused by a bus stopping</td>
</tr>
<tr>
<td>Acceleration</td>
<td>VP</td>
<td>Vehicle acceleration performance</td>
</tr>
<tr>
<td>Braking</td>
<td>VP</td>
<td>Vehicle braking performance</td>
</tr>
<tr>
<td>Adverse weather</td>
<td>Other</td>
<td>Indicator of adverse weather such as snow, or fog (see for example Montgomery &amp; May (1987)).</td>
</tr>
<tr>
<td>Accident likelihood</td>
<td>Other</td>
<td>Likelihood of an accident occurring</td>
</tr>
<tr>
<td>Accident delay</td>
<td>Other</td>
<td>Delay caused by an accident</td>
</tr>
</tbody>
</table>

Table 5.4: List of possible inputs to a ULLT model
5.14 Limitations of using data solely from ILDs

The ULTT models developed in this thesis use solely ILD data as their input. This limitation on the data input was necessary to limit the scope of the research. This section briefly discusses the likely effect on the performance of a ULTT model if only ILD data is available and provides two examples of cases where other data sources are desirable. This section finishes by suggesting the use of temporal metrics as proxy variables of factors which do affect ULTT.

There are two primary problems with using ILD data to measure link travel time:

- **Indirect measure of travel time**
  Firstly, unless the correlation method of section 5.5 is used, ILDs do not enable the ULTT to be measured directly. Instead, outputs available from the ILD (i.e. occupancy, flow, ALOTPV, etc) need to be related to travel time using a model, examples of which have been given in sections 5.4 to 5.9. These models may require a large number of assumptions and simplifications to be made thus reducing the performance of the ULTT model in estimating ULTT.

- **Point measurements**
  The second key problem with ILDs, is they only measure characteristics of the traffic flow at distinct fixed points and not over the whole link. If ILD data is to be used to estimate ULTT, then the ILD has to be in a location where the metrics measured by the ILD have a relationship with the ULTT.

To illustrate the above points, two examples of the limitations of using solely data from ILDs are presented below.

5.14.1 Example #1 - Data from ILDs other than flow and occupancy may be required.

Most ULTT models use solely flow, occupancy, or some ratio of the two as input. However, there may be situations where other data derived from ILDs will be needed. As an example, consider the road network shown in figure 5.4.

The link is 500m long. There is a pedestrian crossing 50m from the upstream junction. The link has a single-loop ILD located 100m downstream from the pedestrian crossing. Assume that the traffic flow is constant throughout the whole day. There is a school located on the north side of the road and a housing estate on the south side. The pedestrian crossing is only used at two periods of the day when children are crossing from school to the housing estate. Obviously when the pedestrian crossing is in use, the travel time over the link will be significantly longer than for times when the crossing is not in use.

However, flow counts and even average occupancy values aggregated over a period of 5 minutes will not be able to distinguish between periods when the pedestrian crossing is in and not in use. If only 5 minute aggregated flow or occupancy were available, then the only
5.14. Limitations of using data solely from ILDs

5.14.1 Example #1 - Importance of the location of an ILD on the link in detecting mid-link delay

The ability of ILD data to indicate mid-link delay will often depend on the location of the ILD. Consider figure 5.5.

An ILD, $d_1$, is placed in the middle of a section of road, P, where parking is allowed. (but placed so that it does not detect parked cars) Another detector $d_2$, is placed 100m downstream of where parking was allowed. Consider that vehicles only ever park on this section of road on a Friday. If there are no parked cars, then the speed of the vehicles measured at $d_1$ is the same as the speed at $d_2$. Hence the ALOTpv (see equation 3.4) at way a model could accurately give the correct travel time were if it were to incorporate a variable indicating the time of day and whether it were a school day or not.

Nevertheless it may be possible to infer that the pedestrian crossing is in use by looking at the time series profile of flow aggregated at small time resolutions (for example the flow and occupancy aggregated over a time of only 4 seconds). Using such data it may be possible to infer that the pedestrian crossing was in use - this would be characterised by a time period where no vehicles flowed over the ILD followed by a small period with high flow. In other words the presence of a platoon of vehicles when usually such platoons did not occur would indicate that the pedestrian crossing had been used. Further research needs to be undertaken on quantifying the extent to which vehicles in a traffic stream are travelling in platoons. Such a metric could then be an input to a ULTT model.

5.14.2 Example #2 - Importance of the location of an ILD on the link in detecting mid-link delay

The ability of ILD data to indicate mid-link delay will often depend on the location of the ILD. Consider figure 5.5.

An ILD, $d_1$, is placed in the middle of a section of road, P, where parking is allowed. (but placed so that it does not detect parked cars) Another detector $d_2$, is placed 100m downstream of where parking was allowed. Consider that vehicles only ever park on this section of road on a Friday. If there are no parked cars, then the speed of the vehicles measured at $d_1$ is the same as the speed at $d_2$. Hence the ALOTpv (see equation 3.4) at

![Figure 5.4: Data from ILDs other than flow and occupancy may be required.](image)

![Figure 5.5: Location of the ILD may determine its ability to indicate mid-link delay](image)
5.14. Limitations of using data solely from ILDs

both positions will be the same. If however there are parked vehicles, then vehicles will be forced to slow down when driving over that section. This will result in the ALOTPV being greater at $d_1$. However, by the time the vehicles reaches $d_2$ they will have accelerated up to their desired speed and the ALOTPV will be the same as in the case where no vehicles were parked over section, P. Thus the detector $d_2$ will not able to detect any change in traffic characteristics but detector $d_1$ will. This example clearly shows that the location of the ILD is crucial in being able to determine mid-link delay.

5.14.3 Eliciting non-traffic variables that influence travel time

The above discussion shows that data solely from ILDs may be insufficient to model the link travel time correctly. Other factors may often be needed. In the above examples, the time of day, indicator of whether the day was a school day, and day of week, would be required in any ULTT model.

It is important to determine the exogenous factors such as time of day, weather, type of day, etc, that affect travel time. Of particular importance is to elicit those variables, which affect travel time, but do not affect the data output from the ILDs. Example 2 above has already illustrated one case of this. A list of exogenous factors that may affect the travel time on a link, yet not affect the output signal from an ILD are listed in table 5.5. These factors may be candidates for inclusion in any ULTT model.

5.14.4 The use of temporal metrics as a proxy of underlying factors

One additional factor that may be incorporated into a ULTT model is a temporal metric, such as an indicator of whether the traffic is in peak or off-peak conditions, day-of-week, or hour-of-day. It must be stressed that the temporal metrics do not themselves influence ULTT. However, these temporal metrics are often highly correlated with underlying factors such as demand and capacity on a link that do strongly affect travel time. Ideally the underlying factors should be explicitly incorporated into the model rather than using a temporal metric as a proxy to this.

For example a model may include time of day within it. At certain times of day, for example at 8:30am, the link travel time will be longer. However, it is not the fact that it is 8:30am that causes the travel time to be longer, but rather the fact that at 8:30am there is a high demand on the road network. Rather than incorporating the time-of-day into the model, the model can be made more generic by incorporating a factor indicating the demand on the road network. Nevertheless, in the absence of such data, then these temporal metrics act as good proxy variables. ULTT models that use solely temporal metrics are included in the comparison of ULTT models in chapter 8.
5.15 Conclusions

This chapter has presented a comprehensive review of the ULTT literature. It first made the differentiation between a ULTT model and a short-term travel time forecasting model. It then identified desirable properties of a ULTT model. Although overall accuracy is arguably the most important performance criterion, there are several other desirable properties that a ULTT model should have.

Six categories of ULTT model have been identified. It was shown that those models based on spot-speed estimates and the correlation methods are inappropriate for use in urban environments. Queuing ULTT models were also seen to have limitations since they require input data which in many cases is infeasible to collect on a large scale. The ULTT models with the most potential were shown to be the regression based models and ANN models (although it is noted that ANNs are related to regression models).

The limitations of the existing models have been noted. In particular many current ULTT models fail to properly model the individual turning movements at junctions, and

<table>
<thead>
<tr>
<th>Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Darkness</td>
<td>Burrow (1986) explored the effect of darkness on the capacity of a junction. He found that the capacity of different junction types were reduced at night by the following amounts; Signal Controlled Intersection = 2.4%, Roundabouts = 4.9%, Major/Minor Priority junctions = 5.6%. This reduction to capacity is caused by the reduction in visibility causing drivers to drive more cautiously.</td>
</tr>
<tr>
<td>Weather</td>
<td>It is known that saturation flow is reduced in poor weather (McDonald &amp; Hounsell 1986; Usami et al. 1986). A reduction in saturation will lead to an increase in intersection delay. This may not necessarily be measured by an ILD, particularly if it is sufficiently upstream of the junction so that queuing does not extend all the way back from it. Indeed the SCOOT model allows for different saturation values to be used at different times of day to overcome this problem. Ibrahim &amp; Hall (1994) analysed data aggregated at 5-minute intervals from a freeway in Ontario. It was found that heavy rain caused the free-flow speed to fall by 10km/h whereas snow reduced the free-flow speed by 35-50km/h.</td>
</tr>
<tr>
<td>Special Event</td>
<td>An indicator of a special event such as a football match or market in town that may affect traffic flow patterns that cannot be detected from analysing ILD data. An example of this was given in section 5.14.2.</td>
</tr>
<tr>
<td>School Day</td>
<td>An indicator of whether the day is a school day or not. The example given in section 5.14.1 illustrates how a school day can affect traffic patterns which may not be detected by data from ILDs.</td>
</tr>
<tr>
<td>Traffic Signal Settings</td>
<td>The cycle time and effective red time are key variables affecting any intersection delay (see appendix A.2). These variables should be incorporated into any ULTT model if they are available. It may be possible to estimate the cycle time using Fourier Analysis of the raw ILD data.</td>
</tr>
</tbody>
</table>

Table 5.5: Additional factors that may affect link travel time yet not affect the output from ILDs
also fail to model mid-link delay properly. A review was undertaken on how these effects can be modelled. The chapter then concluded by investigating whether the performance of a ULTT model would be limited if only ILD data was used. It was suggested that a simple proxy variable for other immeasurable variables, would be temporal metrics such as hour-of-day or day-of-week.

In conclusion to this chapter it was suggested that there is much potential for the development of new ULTT models, particularly using non-parametric methods. Such a non-parametric method will be introduced in the next chapter.
Chapter 6

The k-Nearest Neighbors Method

It was suggested in the previous chapter that existing urban link travel time (ULTT) models may have poor accuracy, transferability, and distribution attributes. This section considers in detail an alternative ULTT model based on the k-Nearest Neighbors (k-NN) method. The first half of the chapter gives an introduction to the k-NN method, outlining all assumptions, and key-design parameters. Problems with the computational requirements of the k-NN method are introduced and methods to overcome these presented. Finally use of the k-NN within the transport field is reviewed and critiqued.

6.1 Basic Concept

The basic idea behind the k-Nearest Neighbor (k-NN) approach is to match the current input variables with historical observations that have similar input variables. The set of input variables is collectively called the feature vector. Each feature vector describes a point in multivariate space, known as the feature space. If the feature vector contains \( p \) attributes, then the feature space is \( p \)-dimensional. The output from the k-NN model is defined to be some function of the known outputs of those records with similar feature vectors. This is explained with the aid of the example below.

Figure 6.1 shows a scatterplot of the independent variable, \( x \), against a dependent variable \( y \). In this example the feature vector is comprised of only one attribute; i.e. \( x \). Suppose it is desired to use the k-NN method to estimate the value of \( y \) when \( x = x_1 \). If \( k \) is chosen to be 5, then the 5 historical observations with a value of \( x \) closest to \( x_1 \) are identified. These observations are shown by the shaded points in the diagram. The \( y \) values of these 5 closest observations are then used to estimate the value of \( y \) when \( x = x_1 \). It is common to use the arithmetic mean of these five \( y \) values.

\(^1\)The American spelling for neighbours has been used since the term ‘Nearest Neighbors’ is a technical term and consistency should be kept with existing literature.
6.2 Theoretical basis of the k-NN method

The previous section has outlined the basic concept of the k-NN method. The k-NN method has been applied in the related fields of classification, density estimation, smoothing, and non-parametric regression. There is therefore a considerable amount of literature on the theoretical basis and characteristics of the k-NN method. It is not possible to review all this material within this thesis. Instead only the fundamentals and those aspects that are relevant in the use of the k-NN method as a ULTT model are discussed. The next sub-section provides a brief historical background of the development of the k-NN method. Section 6.3 then provides a brief outline of the assumptions on which use of the k-NN method relies. Section 6.4 then discusses the key parameters of the k-NN method if it is to be used an a ULTT model. For a comprehensive introduction to the k-NN method the reader is referred to Fukunaga (1990) and Dasarathy (1991).

6.2.1 Historical background of the k-NN method

The first reference to the k-NN method was by Fix & Hodges (1951, 1952) for use as a classifier. However due to limitations of computation power it was not until the late 1960’s and early 1970’s that interest in the method grew. The potential of the k-NN method increased further, when in 1967, Cover & Hart (1967), made an important discovery about the use of the k-NN as a classifier. They discovered that given an infinite sample set, if a 1-nearest neighbor (1-NN) model was used then the probability of error of the k-NN classification rule would not exceed twice the error of the best performance possible (commonly referred to as the Bayes error). This was important since it suggested that the k-NN method offered a simple, yet accurate way of undertaking classification. This result encouraged far wider research into the k-NN method not only as a classifier but also for use in density estimation and non-parametric regression.

Figure 6.1: Example of the use of the k-NN method when \( k = 5 \).
6.3 Assumptions of the k-NN method

The k-NN method relies on several assumptions. It is important for any potential practitioner of the k-NN method to be aware of these assumptions, and where appropriate to test that these assumptions are met. Three basic assumptions of the k-NN method are outlined below.

6.3.1 Similar a posteriori distributions

The k-NN method relies on the assumption ‘that observations which are close together in feature-space are likely to belong to the same class or to have the same a posteriori distributions of their respective classes’ (Devijver & Kittler 1982). It is thus important that the feature vector contains all attributes to allow observations to be classified into their correct class of distribution.

As an example consider the equation \( y = mx + \epsilon \). \( x \) is the independent variable, and \( m \) depends on what class the specimen being measured comes from. \( \epsilon \) is an error term (noise). If there are two classes of the specimen A (\( m = 4 \)), and B (\( m = 1 \)) then the resulting scatter plot of measurements of \( x \) against \( y \) may look like that of figure 6.2.

![Figure 6.2: The k-NN method assumes that the historical observation come from the same class of distribution.](image)

The k-NN method \( (k = 6) \) will be used to estimate the expected value of \( y \) when \( x = x_3 \) and the data comes from class A \( (m = 4) \). If the feature vector contains only the value of \( x \), then the 6 nearest Neighbors will come from both classes of distribution A and B. This is shown by the left diagram of figure 6.2. The resulting estimate will thus be biased in the direction of the other class (class B). However, if both \( x \) and \( m \) are included in the feature vector, then only historical records from the correct class are chosen, resulting in a more accurate estimate of \( y \) for class A when \( x = x_3 \). This is illustrated on the right hand diagram. This simple example thus illustrates the importance of identifying the important attributes that should constitute the feature vector.
6.3.2 Large historical database

It can be shown that the distance from the input feature-vector to the k-nearest neighbors will tend to zero as the number of historical observations, \( N \), tends to infinity (Devijver & Kittler 1982; Cover & Hart 1967). In addition as long as \( k/N \) tends to zero as both \( k \) and \( N \) tend to infinity, the risk (a measure of the performance of the classifier) of the k-NN classification rule approaches the Bayes risk (which is the least risk possible, occurring when complete knowledge of the underlying system is available) (Cover 1968). It is thus desirable to have a large historical database. However increasing the size of the database increases the computation time of the method, although techniques do exist to reduce this problem (Dasarathy 1991).

6.3.3 Random distribution of the k-nearest neighbors about the input feature vector.

A further assumption made in the k-NN method is that the \( k \) nearest neighbor historical observations are distributed randomly in all \( n \) dimensions around the input feature vector. However, if the input feature-vector is located at the boundary of the feature-space of existing observations, then this assumption will not hold since most selected neighbors will tend to be on one side of the input feature vector. Thus input feature vectors located at the boundary will be subjected to larger bias error. Techniques do exist for overcoming these boundary problems. For example, instead of calculating the arithmetic mean of travel time for the \( k \)-nearest neighbors, these \( k \)-nearest neighbors can be used to define a suitable regression line through the local area (Simonoff 1996). This is considered in further detail in section 6.4.3.

6.4 Key parameters of the k-NN method

An Artificial Neural Network model has several key parameters that must be chosen by the practitioner of the model. These parameters include the number of hidden layers, the network architecture, the transfer function, and the training method used. Collectively these parameters describe the architecture of the ANN model. In a similar way, the k-NN method has an associated architecture. There are 4 key design parameters of the k-NN method that define its architecture. These are discussed in detail below.

6.4.1 Value of \( k \)

The first key parameter of the k-NN method is the number of nearest neighbors, \( k \), that should be selected. There is a comprehensive literature on this subject for all streams of k-NN usage. Consider the use of the k-NN in non-parametric regression. The k-NN kernel
estimate, \( \hat{m}_k(x) \), is defined by equation 6.1 (Härdle 1990, p. 42).

\[
\hat{m}_k(x) = n^{-1} \sum_{i=1}^{n} W_{ki}(x) Y_i \tag{6.1}
\]

\( \hat{m}_k(x) \) = Weighted average of the response variables around the point \( x \).

\( n \) = no of records in the feature space

\( Y_i \) = response variable of record \( i \).

\( W_{ki} \) = weight given to response variable \( i \).

If \( J_x \) is the set of the \( k \) nearest neighbors, then the weight, \( W_{ki} \), is given by equation 6.2

\[
W_{ki} = \frac{n}{k} \quad i \in J_x \\
W_{ki} = 0 \quad \text{otherwise} \tag{6.2}
\]

Lai (1977) showed that if the k-NN kernel estimate of equation 6.2 is used, then as \( k \) and \( n \) tend towards infinity, whilst \( k/n \to 0 \), then the bias and variance of the k-NN estimate is given by equations 6.3 and 6.4 respectively.

\[
E \hat{m}_k(x) - m(x) \approx \frac{1}{24f(x)^2} \left[ (m''f + 2m'f') (x) \right] \frac{k^2}{n} \tag{6.3}
\]

\[
\text{var}\{\hat{m}_k(x)\} \approx \frac{\sigma^2(x)}{k} \tag{6.4}
\]

Equations 6.3 and 6.4 show that as the bias is increased, so the variance is decreased, and vice-versa. Since the mean squared error is given by the sum of the bias\(^2\) and the variance, care must thus be taken in selecting the value of \( k \). Equations 6.3 and 6.4 refer to work where the k-NN method is used for non-parametric regression. Fukunaga & Hostetler (1973) performed similar work in the field of non-parametric density estimation. Fukunaga & Hostetler (1973, p. 273) found that the relationship between the Mean Squared Error of the density estimation (MSE), and the number of nearest neighbors was given by equation 6.5.

\[
MSE \propto \frac{1}{k} + \alpha^{-\frac{4}{n}} \left( \frac{k}{N} \right)^{\frac{4}{n}} \tag{6.5}
\]

\( k \) = number of nearest neighbors

\( \alpha \) = represents the rate of change of the underlying density distribution and is thus a measure of the non-linearity of the system. i.e. in linear systems, \( \alpha = 0 \).

\( n \) = number of dimensions in the feature space.

\( N \) = number of historical observations in the database.
The findings from Härdle (1990) and Fukunaga & Hostetler (1973) can be illustrated graphically. Figure 6.3 demonstrates how increasing $k$ reduces the variance error. It shows the use of the k-NN for estimating the linear function $y = x + e$, where $e$ is independently and identically distributed, with an expected value of 0, and a variance of $\sigma^2$. The raw data points are shown as small blue circles, and the true relationship between the two variables shown by the solid blue line. It can be seen from Figure 6.3 that when $k = 3$ (red line) the estimate is very rough and there is significant error to the true line (blue). This is because of the effect of the random component. However, as $k$ is increased to 25 (green line) the random error tends to become cancelled out, and the estimated value is far closer to the actual value.

Figure 6.4 demonstrates how increasing $k$ increases the bias error when using the k-NN to estimate an underlying non-linear system. The diagram shows a scatter plot of the noise-free function $y = x^2$, (blue circles), along with the k-NN estimates of the function for $k = 3$, and $k = 25$. It is seen that the error increases as $k$ increases. This occurs due to the non-linearity of the system.

![Scatterplot](image)

Figure 6.3: As $k$ increases the variance error decreases
6.4. Key parameters of the k-NN method

This work thus shows that care must be taken in selecting an appropriate value of \( k \). This is considered further in chapter 7 where the k-NN is optimised for use as a ULTT model.

6.4.2 Distance Metric

From equations 6.1 and 6.2, it can be seen that an important aspect of the k-NN method is in determining the set of \( k \) nearest neighbors, \( J_x \). The problem remains of how to define the ‘closeness’ between two points in multivariate space. ‘Closeness’ is usually quantified using a distance metric. The most common measurement of distance is the Euclidean distance. However, in multivariate space different variables may be measured on different scales. For example variable M may vary between 0 and 1000, whereas variable N may vary between 0 and 1. Thus a k-NN search using Euclidean distance will be more sensitive to a percentage change in M than to a similar change in N. It is thus likely to pick observations with similar M values regardless of its N value. In this case it is thus desirable to normalize each variable according to a measure of its range. Many possible measures...
of distance have thus been proposed. These are given below along with the appropriate equation for calculating the distance between two \( p \)-variate observations, \( x_a \) and \( x_b \).

**Euclidean**

Distance that the ‘crow flies’ between two points. Unscaled.

\[
d_{ab}^2 = (x_a - x_b) \cdot (x_a - x_b)'
\]  
\(d_{ab} = \text{distance between two points } a \text{ and } b\)
\(x_a = \text{vector defining location of point } a\)
\(x_b = \text{vector defining location of point } b\)

**Unit Map**

All values are mapped onto the unit hyper-cube - i.e. each value is normalised such that the minimum value of the attribute is mapped to 0 and the maximum value of the attribute is mapped to 1.

\[
d_{ab}^2 = (f(x_a) - f(x_b)) \cdot (f(x_a) - f(x_b))'
\]  
\(A = p \times p \text{ matrix, where } p \text{ is the number of features in the feature vector.}\)

**Quadratic form**

A widely used distance metric is based upon the quadratic form (Fukunaga & Hostetler 1973). This is given by equation 6.8.

\[
d_{ab}^2 = (x_a - x_b) \cdot A \cdot (x_a - x_b)'
\]  
Two common attribute-constants are:

- Standard deviation (SE-STD) i.e. \( V = \text{diag}(\sigma_1, \sigma_2, ..., \sigma_p) \)
- Variance (SE-VAR) i.e. \( V = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_p^2) \)

\[
d_{ab}^2 = (x_a - x_b) \cdot V^{-1} \cdot (x_a - x_b)'
\]  
A couple of points should be made about this distance metric. Firstly, equation 6.9 is a special case of equation 6.8, where all the non-leading diagonal entries have been set
6.4. Key parameters of the k-NN method

to zero. Secondly, it should be noted that the Standardised Euclidean normalised by the standard deviation is badly formed. This is because this metric is not dimensionless. Both the Mahalanobis and the SE-VAR metrics are dimensionless.

**Mahalanobis**

Another particular case of the quadratic form is the Mahalanobis metric. This metric weights each variable by its variance and covariance with other variables. It is given by equation 6.10.

\[ d_{ab}^2 = (x_a - x_b) \cdot \Sigma^{-1} \cdot (x_a - x_b)' \]  
\[ \Sigma = p \times p \text{ dimension Variance-Covariance matrix.} \]

**City Block**

The sum of the difference of each of the variables - i.e. the distance a taxi would take in a street-grid system getting from one point to another.

\[ d_{ab}^2 = \sum_{j=1}^{p} |x_{aj} - x_{bj}| \]  
\[ (6.11) \]

**Minkowski**

A generalised distance metric: \( c = 1 \) gives the city block metric \( c = 2 \) gives the Euclidean metric.

\[ d_{ab}^2 = \left[ \sum_{j=1}^{p} |x_{aj} - x_{bj}|^c \right]^{\frac{1}{c}} \]  
\[ (6.12) \]

Some attributes are non-metric - for example an indicator of a school day. In this case a ‘1’ may be used to represent a school day, and a ‘0’ used to represent a non-school day.

**Qi and Smith distance metric**

An alternative distance metric has been proposed by Qi & Smith (2004). Since this metric was used in applying the k-NN to the estimation of traffic flows it will be outlined in detail. Qi & Smith (2004) proposed a method of weighting the differences in each attribute between the current and historical observation, by the effect a difference in that attribute has on the output value.

Each attribute of the feature vector is first divided into \( m \) categories. The value of \( m \) is not necessarily the same for all attributes. The mean, \( \mu_m \), and standard deviation, \( \sigma_m \), of the output value of each category is calculated. The statistical significance of the
variation of the output value between categories is calculated, using the metric defined in equation 6.13. This metric is based on the t-statistic for independent pooled samples.

\[ D_{a,b} = \frac{\mu_a - \mu_b}{\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}\right)^{\frac{1}{2}}} \]  

(6.13)

\( D_{a,b} \) = the distance between the \( a \)-th and the \( b \)-th category.
\( \mu_i \) = the mean output value of category \( i \).
\( \sigma_i \) = the standard deviation of the output value of category \( i \).
\( n_i \) = the number of observations in category \( i \).

This distance metric is calculated between all combinations of categories, thus forming a distance matrix \( D_X \). This matrix has a size of \( m \times m \). The process is then repeated for all attributes, resulting in a separate distance matrix for all attributes.

These distance matrices can then be used to determine the distance between any two observations as follows:

- **Step 1** - Look at the value of the first attribute of each observation. Firstly determine which of the \( m \) categories, \( c_1 \) and \( c_2 \), the value for each of the two observations falls into. Using the distance matrix for the selected attribute, look-up the distance between categories \( c_1 \) and \( c_2 \).

- **Step 2** - Repeat step 1 for all attributes

- **Step 3** - The total distance between the two observations is the sum of all the attribute-distances calculated in step 1.

One of the advantages of this approach (although not mentioned by Qi & Smith (2004)) is that it allows for non-linearity in the effect of variation of each attribute on the output value. An attribute may have little effect on the output value below a certain value. However, above this value, the output value may be very sensitive to it. For example, delay at an intersection is very sensitive to the flow, when the flow is close to the capacity of the link. However, when the degree of saturation is below 0.7, then the delay tends not to be sensitive to flow. The distance metrics of 6.7 to 6.12 do not allow the influence of each attribute to differ according to the value of that attribute, and thus cannot take into account this non-linearity.

**Optimal Distance Metric**

The work above has identified various distance metrics that can be used in the k-NN method. Fukunaga & Hostetler (1973) undertook some research to determine the theoretical optimal quadratic metric (see equation 6.8). They showed that the optimum matrix \( A \) to use should take the same ellipsoidal shape as the underlying distribution. If the underlying distribution is Gaussian, then the optimum matrix, \( A \), of equation 6.8 is the
6.4. Key parameters of the k-NN method

The k-NN model presented in equation 6.1 shows that the estimate is dependent on the weights $W_{ki}$ used. These weights vary the importance of each of the nearest neighbors in contributing to the final estimate. This problem can be stated in a more general way.

Suppose that $J_y$ is the set of output values of the $k$ nearest neighbors to the input vector $x_c$. The estimate of the output value, $y_c$ to input vector $x_c$ is given by equation 6.14.

$$y_c = f(J_y) \quad (6.14)$$

The function $f()$ is termed the Local Estimation Method (LEM). A key design parameter for the k-NN method is to determine the appropriate LEM to use. Commonly the arithmetic mean of the $y$-values of each of the $k$-nearest neighbors is used, and indeed this is the case in equation 6.1. A more robust approach less susceptible to outliers would be to use the median value of the set $J_y$.

### Weighting by distance

If the Local Estimation Method used is the mean, then the value of $W_{ki}(x)$ in equation 6.1 will be a constant of 1. A small modification is to make the value of $W_{ki}(x)$ be dependent on the distance of the historical feature vector $(X_i)$ from the input feature vector $(x)$. Commonly the weights come from a kernel function and are given by equation 6.15 (Härdle 1990).

$$W_{ki}(x) = \frac{K(x - X_i)}{f(x)} \quad (6.15)$$
where

\[ \hat{f}(x) = n^{-1} \sum_{i=1}^{n} K(x - X_i) \]  \hspace{1cm} (6.16)

Common values for the Kernel function, \( K(u) \), are given in Table 6.1. \( u \) is given by equation 6.17.

\[ u = \frac{x - X_i}{\max(x - X_i)}, \hspace{1cm} X_i \in J_x \]  \hspace{1cm} (6.17)

\( J_x \) is the set of input feature vectors of the \( k \) nearest neighbors.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>( k(u) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epanechnikov</td>
<td>( \frac{3}{4} (-u^2 + 1) I )</td>
</tr>
<tr>
<td>Quartic</td>
<td>( \frac{15}{16} (1 - u^2)^2 I )</td>
</tr>
<tr>
<td>Triangular</td>
<td>( (1 -</td>
</tr>
<tr>
<td>Gauss</td>
<td>( (2\pi)^{-\frac{1}{2}} \exp \left( -\frac{u^2}{2} \right) I )</td>
</tr>
<tr>
<td>Uniform</td>
<td>( \left( \frac{1}{2} \right) I )</td>
</tr>
</tbody>
</table>

n.b. \( |u| \leq 1 \)

Table 6.1: Common Kernel Functions (Härdle 1990, p.138)

There is a similar technique to the k-NN method called Kernel Smoothing. Kernel Smoothing uses all records in the historical database to estimate the output value. However, each record’s influence is weighted by its distance from the input feature vector using the Kernel in equation 6.15.

**Regression**

One criticism of the use of the mean, median, and weighted mean, is that these estimators are susceptible to boundary bias. The \( k \) nearest points may tend to lie on one side of the input feature vector, thus giving a biased result. An alternative local estimation method is to use the \( k \) nearest neighbors to determine a linear regression model suitable for the local area. This model is then used to estimate the travel time for the input feature vector. Ordinary Least Squares (OLS) regression is used (Gujarati 1995). OLS assumes that the model is of the form:

\[ y = X\beta + e \]  \hspace{1cm} (6.18)

\( y \) = dependent variable - i.e. travel time. It is an \( n \times 1 \) vector
\( X \) = Matrix of independent variables. It is an \( n \times p \) matrix. i.e. Each row represents one observation (i.e. one nearest neighbor), with the columns representing each attribute.
\( n \) = Number of nearest neighbors
\( p \) = Number of attributes in the feature vector.
\[ \beta = p \times 1 \text{ vector of the unknown parameters.} \]
\[ e = \text{Is an } n \times 1 \text{ vector of error terms, mean } = 0, \text{ variance } = \sigma^2. \]

In OLS regression, \( \beta \) is given by equation 6.19.

\[ \beta = (X'X)^{-1} \cdot X \cdot y \tag{6.19} \]

Figure 6.5 gives an example of how the regression LEM would be used in the k-NN method. In this example, the size of the feature-vector is only one, i.e. there is one independent variable, \( x \). The underlying historical data comes from the function given by equation 6.20.

\[ y = 200 \cdot \sin\left(\frac{x}{20}\right) + e \tag{6.20} \]

Plot a of figure 6.5 shows the raw data. Suppose it is desired to know the value of the dependent variable \( y \), when \( x = 5 \). Assume that the Euclidean distance metric is used, and that the value of \( k \) has been set to 150. There are two primary stages:

- The first stage of the k-NN method is to find those feature-vectors in the historical data-set closest to the input feature vector (i.e. find those historical records which have a value of \( x \) closest to \( x_1 \)). These 150 nearest neighbors are shown by the red crosses in plot b.

- Having found these 150 nearest neighbors, the second stage involves estimating the value of \( y \) when \( x = 5 \) using some function of the \( y \) values of the 150 nearest neighbors. If the regression LEM is used, then an OLS regression curve is fitted to the data based on the regression model stated in equation 6.18. The expected value of \( y \) when \( x = 5 \) can then be determined using this regression model. This is shown in plot c of figure 6.5.

One advantage of using the regression LEM can be seen from the example of figure 6.5. It will be noticed that it is desired to know the value of \( y \) when \( x \) is very close to a boundary. Because of this, the 150 nearest neighbors tend to be to one side of \( x = 5 \), and the assumption of section 6.3.3 that the nearest neighbors are distributed about the desired value of \( x \) is not met. Because of this any estimation using a mean or median LEM will be biased. In the example of figure 6.5 the estimate of \( y \) using the mean is 122.1 and using the median is 135.3 when the expected value of \( y \) using equation 6.20 is 49.5. However, the regression LEM gives a far better estimate of 52.3. It can be seen that regression based LEMs will perform better at the boundary than mean and median LEMs.
6.4. Key parameters of the k-NN method

Figure 6.5: Example of the use of the regression LEM
Lowess

One deficiency of the Ordinary Least Squares method is that it is not robust. Outliers in the data can cause the parameters estimated by the model to be inaccurate. This results in inaccurate estimates of travel time. Modified versions of OLS regression have been proposed which are more robust to such outliers. One such modified regression technique is the LOWESS (LOcally WEighted Scatter plot Smoothing) method first proposed by Cleveland. The Pseudocode of this method is taken from Härdle (1990, p.192):

- **STEP #1** Fit a polynomial regression in a neighborhood of $x$, that is, find coefficients $\{ \beta_j \}_{j=0}^p$ which minimise equation 6.21:

$$n^{-1} \sum_{i=1}^{n} W_{ki}(x) \left( Y_i - \sum_{j=0}^{p} \beta_j x_j \right)^2$$

(6.21)

$n =$ number of nearest neighbors
$W_{ki}(x) =$ weight for each nearest neighbour $i$.
$\beta =$ coefficients of the regression equation
$Y_i =$ output value for record $i$.
$x_j =$ input value of record $j$.

- **STEP #2** Compute from the estimated residuals, $\epsilon_i$, the scale estimate, $s = \text{med}\{|\epsilon_i|\}$, and define the robustness weights $\delta_i = K(\epsilon_i/6)$, where $K$ denotes the quartic kernel, $K(u) = (15/16)(1 - u^2)^2 I(|u| \leq 1)$.

- **STEP #3** Fit a polynomial regression as in STEP #1 but with weights $\delta_i W_{ki}(x)$.

Essentially this method reduces the influence of each data point as its residual, $\epsilon_i$, increases. A minor modification to this method is to iterate stages 1 to 3 several times.

### 6.4.4 Attributes to include in the feature vector

One of the most crucial aspects to the success of the k-NN method is the correct selection of the attributes that should constitute the feature vector. As illustrated in figure 6.2, if too few attributes are used, then the model will not have enough features with which to differentiate the input observation from the historical observations, and inappropriate historical observations will be chosen. That is historical observations coming from a different underlying distribution of travel time will be chosen. However, if too many attributes are included in the feature vector, then the model will be prone to the curse of dimensionality. This describes the situation when irrelevant attributes in the feature vector dominate the distance metric, reducing the influence of the relevant attributes on the distance metric.
6.5 Computational Issues

Castelli (2001) illustrates the problem of the curse of dimensionality, where 20,000 random observations are selected of a feature vector comprised of 100 independent Gaussian random variables. If the distance of all observations to one another are calculated (i.e. \( \frac{n(n-1)}{2} = 199,990,000 \) distances), then all the distances lie in a range from 10.2 to 18.3. This differs from the intuitive expectation that there would be at least some points which had a distance close to zero from one-another.

Various methods of determining which attributes to include have been proposed. These feature selection methods can be divided into two categories, filter and wrapper methods (Liu & Motoda 1998). A filter method chooses the attributes independently of the final model, whereas a wrapper model uses the final model to evaluate the performance of the current feature vector. The latter method is prone to overfitting of the data. Due to reasons outlined in section 7.6, the feature vector used for the ULTT model will be fixed. Thus the study of feature selection is beyond the scope of this thesis. The interested reader is referred to Kittler (1978) and Kudo & Sklansky (2000) for a good summary of feature selection methods.

6.5 Computational Issues

One major disadvantage of using the k-NN method is that it is computationally expensive. The k-NN method is a lazy-learning method, deferring the decision on how to generalise beyond the training data only when each new feature-vector is encountered. Compare this with eager-learning models such as regression analysis and artificial neural networks. Once trained, these methods require very little computation time to estimate the output value for any new feature-vector encountered. The computation time for the k-NN method is dominated by the following two components:

- Calculating the distance between the feature-vector and all other \( n \) historical feature-vectors. This is proportional to \( n \).

- Sorting these \( n \) distances into order to determine the \( k \) nearest neighbors. This can be achieved using sorting methods such as Quicksort or Heapsort (Press et al. 1986).

   Nevertheless the sorting time is still proportional to \( n \cdot \log_2(n) \).

It is thus desirable to reduce the number of feature-vectors, \( n \), in the historical dataset. Two such adaptations of the k-NN approach have been suggested that reduce the size of the historical data set, the Condensed Nearest Neighbor rule (Hart 1968), and the Reduced Nearest Neighbor rule (Gates 1972). However both these methods work best when k-NN method is being used as a classification tool rather than as a density estimator. They also work best when only one nearest neighbor is being sought. There exist other heuristic approaches to reducing the number of historical feature-vectors. One such method is presented below.
6.5.1 Univariate Screening

The previous section has indicated that the computation time of the k-NN method is dependent on the number of distance measurements made to historical feature-vectors. If the size of the historical data-set is too large, then the k-NN method will become computationally infeasible. However, one of the key assumptions of the k-NN method is that there is a large historical database (see section 6.3.2). It is thus desirable to eliminate those feature-vectors which are obviously far from the input feature vector. The exact distance between the input and historical feature-vector, should only be calculated between those feature-vectors which are probably close to the input-feature vector. A heuristic approach, which is similar to one proposed by Yunck (1976), has been developed within this thesis to do this. The approach reduces the size of the data-set considered by undertaking univariate screening of the data, whereby each variable in the feature-space is considered independently of the other variables. Univariate screening is advantageous since the relevant fields in databases can be indexed to allow the valid records to be found very quickly.

As an example consider a feature vector comprised of flow and occupancy. Firstly the flows and occupancies are normalised by the standard deviation of each attribute. The flows of all the historical observations are then compared to the flow of the input observation, and the historical observation discarded from future consideration if its flow falls outside a certain range to that of the input observation. This is repeated with the occupancy variable. The number of valid observations is calculated and the process repeated using smaller thresholds, until the number of historical observations considered for further analysis is deemed to be a suitable size. This is illustrated in figure 6.6. Notice how the shaded area of valid observations reduces in size between iteration 1 and 2.

![Figure 6.6: Univariate screening employed to speed up search](image)

6.5.2 Problems with univariate screening

As illustrated in the simple two-dimensional case shown in figure 6.7, the univariate screening approach can exclude observations that are closer to the input observation than some
historical observations that are included. Suppose a square, centred on the input observation, contains exactly $k$ observations (the required number of matches). Observations such as B located at the corner of the square could be further from the input observation than observations such as A, located just outside the centre of the square. To overcome this, all observations within a circle having a radius of half the diagonal need to be considered. However, univariate screening techniques do not allow all observations within such a circle to be determined. Instead, it is necessary to increase the size of the square to enclose this circle, as shown in the diagram. For a feature-vector of two dimensions then the dimensions of the square containing all points that must be considered, should be $\sqrt{2}$ times the length of the smallest square that contains at least $k$ observations. This rule can be generalised so that for a feature-vector of $p$ dimensions the length of the square containing all points that must be considered should be $\sqrt{p}$ times the length of the smallest hyper-cube that contains $k$ observations.

Figure 6.7: Univariate screening, if applied rashly, may eliminate observations that are amongst the k-nearest neighbors.

This approach to quickly selecting the nearest neighbors has been used in this research. It is realised that the approach taken may result in the set of nearest neighbors being slightly wrong. However, without taking this heuristic approach to reducing the effective size of the historical data-set considered, it would not be computationally feasible to use the k-NN method.

6.6 k-NN Transport Applications: Literature Review

There is a small amount of literature on the use on the k-NN within the transportation field. Primarily this research has been in traffic flow estimation, although the method has been applied also in link travel time estimation. This section reviews this literature and identifies short-comings of current applications of the k-NN method.
6.6.1 Application To Traffic Flow Estimation

The first explicit use of the k-NN model in the transport literature was by Davis & Nihan (1991) to forecast flow and occupancy at a time interval $t$, given various flow and occupancy readings at the time interval $t-1$. They used real-world data from a highway in Washington state. However, this research was hindered by a lack of data, since only 89 historical records were available with a further 28 validation records. This lack of data violates one of the key criteria of the k-NN method outlined in section 6.3.2. Davis & Nihan (1991) considered three different settings for the value of $k$ (1, 5, and 10) as well as three different distance metrics (city block, Euclidean and the Euclidean standardised by the standard deviation). Only the mean was considered for use as the Local Estimation Method. When predicting flow, they found that the optimal values were $k = 1$ with use of the Euclidean distance metric. They also compared their model against a traditional time-series analysis using the Box-Jenkins approach. They found that the k-NN method did not outperform this model.

Smith & Demetsky (1997) used the k-NN method to forecast traffic flow. They compared the results of the k-NN model with those using historical averages, ARIMA models, and a back-propagation neural network. They found that the nearest neighbor approach gave the smallest error of all the methods and also gave an error distribution which tended not to be skewed. They also found that it was the most successful when transferred to a different site. One reason for the improved performance of the k-NN method may be because a large data set of 2400 historical records obtained from a highway in Northern Virginia was used. However, Smith & Demetsky (1997) failed to consider the effect of varying the key k-NN design parameters; only the Euclidean distance metric was used with the mean used for the LEM. The value of $k$ was not reported.

In later research by the same author (Smith et al. 2002), the author compared the performance of the k-NN model against a SARIMA model in estimating 15-minute aggregated flow on the M25 motorway. They used a dataset of over 4,000 records from Autumn 1996. The feature-vector contained the flow from the two previous 15 minute periods as well as the historical values of flow for the current and next 15-minute time period of the day. In this research Smith et al. (2002) did vary the value of $k$ used and found $k = 20$ to produce the best results. Various LEM functions were used ( Smith et al. (2002) used the term forecast generation method). They proposed the use of the mean, and then variations of a weighted mean (see equation 6.15). Only the Euclidean distance metric was proposed. However, although different combinations of $k$ and LEM were used, they found the k-NN unable to outperform a SARIMA model. These results suggest that the k-NN method may not be appropriate for use in short-term forecasting models.

Clark (2003) also proposed the use of the k-NN method to forecast the flow, speed, and occupancy recorded by an ILD. He also used MIDAS data from the M25. However, unlike Smith et al. (2002) he used 10-minute aggregated data, and used only 3,024 records from February and March 1998. He used an adaptation of the city-block method but failed to
consider other distance metrics. The distance metric is given by equation 6.22.

\[
d = \sum_{i=1}^{L} \left[ \frac{q_c - q_h}{w_q} \right]^2 + \sum_{i=1}^{L} \left[ \frac{v_c - v_h}{w_v} \right]^2 + \sum_{i=1}^{L} \left[ \frac{o_c - o_h}{w_o} \right]^2
\]

(6.22)

\(q\) = 10-minute average flow
\(o\) = 10-minute average occupancy
\(v\) = 10-minute average velocity
\(c, h\) = subscript denotes current and historic record
\(w_q, w_v, w_o\) = measure specific weights
\(L\) = number of lags used in the feature-vector

One weakness of this metric is that Clark (2003) fails to justify a method to determine the appropriate weights. Another weakness of the research is the failure to compare the performance of the k-NN method against time-series models. The work of (Smith et al. 2002) has suggested that a SARIMA model may be more appropriate.

Oswald et al. (2000) wrote a report outlining the methodology of using the k-NN method in traffic flow forecasting. However, this report gave little consideration about the key k-NN parameters, and how they can be optimised. Instead, the report concentrated on ways to structure and index multi-dimensional data. The focus of the paper appears to be to increase the speed (and not the accuracy) of the k-NN method.

### 6.6.2 Application To Travel Time Estimation

The previous section has given a brief overview of the use of the k-NN method in short-term traffic flow prediction. There are three instances in the literature where the k-NN method has been applied to estimate link travel time. Handley et al. (1998) used the k-NN method to estimate travel time on freeways in the San Diego area of the USA. Their research indicated that on freeways the most important input variables were flow and occupancy rather than other features such as time-of-day. However, it is not possible to comment in detail on this research, since much of this work was proprietary.

You & Kim (2000) used a k-NN method to estimate spot-speeds on a highway in Korea and travel time in an urban area in Seoul. When estimating travel time in the urban area, data derived from ILDs was not used as input. Rather travel times from probe vehicles obtained over the previous 15 to 60 minutes were used. You & Kim (2000) do not define the distance metric they used. However, they do report that they used a local linear trend line as their local estimation method. They report a MAPE in the region of 8 to 10%. However, they do not compare this accuracy with that achieved by other methods which makes their results difficult to interpret.

More recently Bajwa et al. (2003b,a) have used the k-NN method to estimate the travel time on a 12km expressway in Tokyo. As input to their feature vector they used the inverse of the time-mean speed aggregated over 5 minutes, obtained from 40 ultrasonic detectors
on this stretch of road. The feature vector also contained data from more than one-time period. In off-peak hours the authors used 2 time periods, and in peak hours (when they thought there would be heavy congestion) they used 6 time-periods of data. Thus their feature-vector contained between 80 and 240 variables. This very large feature vector led to computational limitations which forced them to only consider historical feature vectors which occurred in the same day-type and time-type. There were a total of 12 day-type time-type combinations. This assumption of recurrent travel-time patterns is one large limitation of their research and would result in their model not being able to react to rare incidents in off-peak hours for example. In the k-NN model that is developed in chapter 7 no such assumption of recurrent travel-time patterns is used. In order to find the nearest neighbors, Bajwa et al. (2003b,a), used the distance metric, \(d\), given by equation 6.23.

\[
d(p, t, h, t_s) = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \left[ \left( \frac{1}{v(i, t - j, p)} \cdot \frac{L(i)}{L} \right) - \left( \frac{1}{v(i, t_s - j, p)} \cdot \frac{L(i)}{L} \right) \right]^2
\]

\(L(i) = \) distance from detector \(i\) to detector \(i + 1\)
\(L = \) total distance of link
\(n_i = \) total number of detectors
\(n_j = \) total number of time-lags included in feature vector
\(v(i, t - j, p) = \) time-mean speed aggregated over 5-minutes, from detector \(i\), \(j\) time periods ago from current time \(t\), on prediction day \(p\).
\(v(i, t_s - j, p) = \) time-mean speed aggregated over 5-minutes, from detector \(i\), \(j\) time periods ago from current time \(t\), on historical day \(h\).

Essentially this is a city-block distance metric (see section 6.4.2) where each input has been weighted by the length of road that the detector represents. This distance metric is then used to select the k-nearest neighbors. The estimated travel time is then simply the arithmetic mean of the travel times of the nearest neighbors.

In order to determine the number of nearest neighbors to be used in each day-type time-type combination, and the number of time-lags, \(n_j\) to use, Bajwa et al. (2003b,a) used a genetic algorithm (see Mitchell 1997).

The speed data was not filtered for erroneous records. However, the training-set was filtered for erroneous travel-time records using the box plot technique. This method removes all records which are further than 1.5 times the interquartile range away from the upper and lower quartiles. To the best knowledge of this author, this is the only example in the ULTT literature where the travel-time data is explicitly filtered.

Bajwa et al. (2003b,a) presents novel techniques for determining the optimal number of nearest neighbors to select. It also considers the filtering of erroneous travel-time records. However, it fails to filter out erroneous speed measurements, and does not consider other distance metrics. Like other ULTT literature, no comparison of the k-NN method against
other models is undertaken.

6.6.3 Critique of existing usage of the k-NN method within transportation

The previous two sections have outlined the usage of the k-NN within the transportation literature. Various criticisms can be made of the existing literature. These are outlined below

- **Reference to underlying theory** - the literature tends to provide little reference to the underlying mathematical theory which underlines the k-NN method. No mention is made of notable contributions such as Cover & Hart (1967), Cover (1968), Fukunaga & Hostetler (1973) and Fukunaga (1990), and of other work including that by Härdele (1990). Such works have been introduced in this current chapter.

- **Comparison with other models** - the literature tends to fail to compare the performance of the k-NN method with other models. This makes it difficult to interpret whether the k-NN model is superior to other methods. Research is presented in chapter 9 which compares the k-NN model with other methods.

- **Data cleaning** - With the partial exception of Bajwa et al. (2003b, a) none of the literature explicitly considers cleaning the data used to train the k-NN method. Work in chapter 9 will show that the k-NN is sensitive in some circumstances to the presence of unfiltered data.

- **Experimental methodology** - much of the literature is weak in explaining the methodology. For example, none of the literature explicitly states how independence between the training and validation data was assured. If independence of the datasets is not assured then the same record may exist in both training and validation data set. This record would obviously be one of the selected nearest neighbors. In the research in this thesis independence between the training and historical data set is assured (see chapter 7).

- **Optimisation of the k-NN method** - other than Oswald et al. (2000) and Bajwa et al. (2003b, a) the literature does not offer guidelines on optimising the k-NN method. Oswald et al. (2000) also fails to state that the key design parameters may not necessarily be independent of one another and changing one parameter may affect the optimal parameter of another. A methodology to optimise the k-NN method for use as a ULTT model is presented in chapter 7.

The research presented in this thesis on the application of the k-NN for use as a ULTT model, will address the criticisms of the existing literature identified above.
6.7 Conclusions

This chapter has introduced the k-Nearest Neighbors (k-NN) method. It has given a theoretical foundation for the technique, including stating the assumptions upon which the k-NN method relies. The chapter then outlined the key design parameters. These were identified as: the attributes in the feature vector, the form of the distance metric, the value of $k$, and the local estimation method. The computational issues faced in using the k-NN method were then highlighted, and a pragmatic solution called *univariate screening* proposed. A review of the existing use of the k-NN method within the transport literature was given. Primarily this has been in the field of short-term flow forecasting. A critique of this literature was given. The next chapter will address these criticisms and develops a methodology by which the k-NN method can be optimised for use as a ULTT model.
Chapter 7

Characteristics of the k-NN method when used as a ULTT model

7.1 Introduction

Chapter 6 introduced the k nearest neighbors method, giving a brief review of the theory and assumptions underlying this methodology. This chapter considers the problems in applying the k-NN method for use as an Urban Link Travel Time (ULTT) model. The chapter begins by providing a simple example of how the k-NN method can be used as an ULTT model. The second part of this chapter then tests the assumption that records close together in feature space will tend to come from records with the same underlying travel time distribution. The third section presents research undertaken to determine the optimal settings for the k-NN ULTT model. Finally research is undertaken on the effect of the size of the historical database used on the performance of the k-NN ULTT model.

7.2 Using the k-NN as a ULTT model

This section provides a brief overview and illustration of how the k-NN method can be used to estimate ULTT using historical travel time and inductive loop detector (ILD) data. There are five stages to this process:

- Collection of historical data
- Specification of k-NN model parameters
- Collection of data for input feature vector
- Determination of the k closest neighbors
- Estimation of travel time from the k closest neighbors
Each of these processes is elaborated upon below.

### 7.2.1 Collection of historical data

In the first stage it is necessary to collect historical travel time records over the link, along with measurements of flow and occupancy from detectors over the link taken at the same time as the travel time record. For each vehicle travel time measurement, a record is added to the database. If the link contains three ILDs, then the record would contain the following attributes:

- Vehicle travel time record number (the key field).
- Travel time of the vehicle.
- 15-minute aggregated flow at ILD 1.
- 15-minute aggregated occupancy at ILD 1.
- 15-minute aggregated flow at ILD 2.
- 15-minute aggregated occupancy at ILD 2.
- 15-minute aggregated flow at ILD 3.
- 15-minute aggregated occupancy at ILD 3.

In a ULTT model the travel time of the vehicle is the dependent variable, \(Y\). The set of independent variables, \(X\), is the ILD data.

If there are 90,000 travel time records then the historical database may look something like table 7.1.

<table>
<thead>
<tr>
<th>Vehicle Record No</th>
<th>Y travel time</th>
<th>X 1</th>
<th>X 2</th>
<th>X 3</th>
<th>X 4</th>
<th>X 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>82.5</td>
<td>814</td>
<td>8.2</td>
<td>645</td>
<td>5.2</td>
<td>768</td>
</tr>
<tr>
<td>2</td>
<td>78.9</td>
<td>810</td>
<td>7.9</td>
<td>640</td>
<td>4.7</td>
<td>745</td>
</tr>
<tr>
<td>3</td>
<td>77.1</td>
<td>810</td>
<td>7.9</td>
<td>640</td>
<td>4.7</td>
<td>745</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>89,999</td>
<td>180.5</td>
<td>500</td>
<td>16.4</td>
<td>405</td>
<td>6.2</td>
<td>925</td>
</tr>
<tr>
<td>90,000</td>
<td>140.8</td>
<td>814</td>
<td>8.2</td>
<td>590</td>
<td>5.8</td>
<td>867</td>
</tr>
</tbody>
</table>

Table 7.1: Historical data required to use the k-NN method as a ULTT model

### 7.2.2 Specification of k-NN model parameters

In the second stage it is necessary to specify the k-NN design parameters that were outlined in section 6.4. In the case of this example, the four key design parameters are specified as follows:
• **Feature Vector:** the feature vector (i.e. set of independent variables) is defined as the 15 minute aggregated flows and occupancies from each of the three detectors (i.e. the variables $q_1$, $o_1$, $q_2$, $o_2$, $q_3$, and $o_3$.) For example, the feature vector for vehicle record 2 in table 7.1, $x_{h,2}$ is $[810, 7.9, 640, 4.7, 745, 5.8]$. The travel time corresponding to feature vector $x_{h,2}$ is 78.9 seconds.

• **Value of k:** In this example the 100 closest variables will be selected.

• **Distance Metric:** In this example the distance metric, Standardised Euclidean normalised by variance, will be chosen. This is given by equation 6.9.

• **Local Estimation Method:** In this example, the median value will be used.

These values have been selected solely for illustration purposes. A methodology for specifying the k-NN design parameters will be provided in this chapter from section 7.4.

### 7.2.3 Collection of data for input feature vector

After stage 2, the k-NN method is ready to be applied for use as an ULTT model. Suppose it is desired to estimate the travel time at time $t$. The first step is to collect the flow and occupancy data at time $t$ as specified in the feature-vector defined in stage 2 - i.e. collect the values of $q_1$, $o_1$, $q_2$, $o_2$, $q_3$, and $o_3$ at time $t$. This set of variables is known as the input feature vector, $x_c$.

### 7.2.4 Determination of the k closest neighbors

The fourth stage is to find the records in the historical database (illustrated in table 7.1.) that most resemble the traffic conditions on the road at time $t$. This is achieved by first measuring the distance between the input feature vector and each of the 90,000 historical records in the database.

The distance between the input feature vector and the $i$ - $th$ historical record of table 7.1 is given by equation 7.1.

$$d_{ab}^2 = (x_{h,i} - x_c) \cdot V^{-1} \cdot (x_{h,i} - x_c)'$$

where:

$$V = \begin{pmatrix}
\sigma_{q1}^2 & 0 & 0 & 0 & 0 \\
0 & \sigma_{o1}^2 & 0 & 0 & 0 \\
0 & 0 & \sigma_{q2}^2 & 0 & 0 \\
0 & 0 & 0 & \sigma_{o2}^2 & 0 \\
0 & 0 & 0 & 0 & \sigma_{q3}^2 \\
0 & 0 & 0 & 0 & 0 & \sigma_{o3}^2
\end{pmatrix}$$

$\sigma_{qi}^2$ = variance of the flows from detector $i$ calculated from all 90,000 historical records in table 7.1.
\[ \sigma_{oi}^2 = \text{variance of the occupancies from detector } i \text{ calculated from all 90,000 historical records in table 7.1.} \]

Once the distance between the input feature vector and each historical feature vector has been calculated, it is then possible to sort the feature vectors in order of distance. The 100 closest feature vectors can then be found. As an example, suppose the input feature vector \( x_c \) is \([812, 7.8, 610, 4.6, 720, 5.5]\), then table 7.2 shows the list of closest neighbors sorted by distance.

<table>
<thead>
<tr>
<th>rank</th>
<th>distance</th>
<th>Vehicle Record No</th>
<th>Y travel time</th>
<th>X</th>
<th>( q_1 )</th>
<th>( q_2 )</th>
<th>( q_3 )</th>
<th>( q_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.328</td>
<td>16,898</td>
<td>75.2</td>
<td></td>
<td>805</td>
<td>7.8</td>
<td>645</td>
<td>4.6</td>
</tr>
<tr>
<td>2</td>
<td>8.335</td>
<td>16,895</td>
<td>84.5</td>
<td></td>
<td>805</td>
<td>7.8</td>
<td>650</td>
<td>4.6</td>
</tr>
<tr>
<td>3</td>
<td>8.453</td>
<td>2</td>
<td>78.9</td>
<td></td>
<td>810</td>
<td>7.9</td>
<td>640</td>
<td>4.7</td>
</tr>
<tr>
<td>99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>10.245</td>
<td>2</td>
<td>90.3</td>
<td></td>
<td>840</td>
<td>7.6</td>
<td>635</td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td>10.257</td>
<td>2</td>
<td>79.2</td>
<td></td>
<td>800</td>
<td>7.9</td>
<td>590</td>
<td>4.4</td>
</tr>
</tbody>
</table>

Table 7.2: The 100 closest historical travel time records to the input feature vector, sorted by distance

### 7.2.5 Estimation of travel time from the k closest neighbors

In this example the value of \( k = 100 \) was chosen. Table 7.2 has identified the 100 records in the historical database closest to the input feature vector. The travel times associated with each of these travel time records (column Y in table 7.2) is then used to estimate the travel time for time \( t \). In this example the median of these 100 travel time records will be used.

### 7.2.6 Conclusions

This section has provided a small example of how the k-NN method can be applied to estimate travel time using data from inductive loop detectors. The next section proceeds to prove that the assumptions underlying the k-NN method are valid in the context of its use as a ULTT model. The work of section 7.4 onwards proposes a methodology to determine the optimal k-NN settings.

It should be noted that like most other ULTT models, the k-NN method requires historical data to be collected for the link for which ULTT is to be estimated. The difference between the k-NN method and other ULTT models, is that the k-NN only uses these historical records at the moment that the estimation is required. The k-NN method does not necessarily have to be trained, although as will be shown in later sections of this chapter, the performance of the k-NN method can be improved by optimising the values of the key design parameters.
7.3 Testing Basic Assumptions

As stated in section 6.3, the most crucial assumption upon which the k-NN method depends is that records which are close together in feature space come from the same underlying output distribution. In the context of using the k-NN method as a ULTT model, this means that records which have similar flows and occupancies should also tend to come from the same underlying travel time distribution. Due to the importance of this assumption, an experiment was undertaken to determine whether the degree to which this assumption held true when the k-NN method was used in a ULTT model.

7.3.1 Data

The data for this experiment was the Russell Square 2003 data set described in section 4.7. This contained a total of 24,718 individual vehicle travel time records which had been filtered using the DSA and Overtaking Rule. Each record contained a date-stamp, 15-minute aggregated occupancy, and flow data from three ILDs, as well as the individual vehicle link travel time. Data was collected over 37 days. In order to ensure that the validation data was independent from the training data, one day was chosen at random. This day is known as the validation day. Only data from this one day was used as validation data. A total of 15,000 records from the other 36 days was chosen at random to act as the training data.

Before the validation data records were selected from the validation day each travel-time record was checked to see whether it had at least 7 other travel-time records within 450 seconds (7 1/2 minutes) of it. Such a record was called a valid record. If the record did not satisfy these constraints then the travel-time record was not considered for use as a validation record. The reason for performing this filtering was that later on it is necessary to estimate the actual mean travel time and standard deviation of travel times in the same 15-minute period as the travel-time record. The number 7 was chosen since this would ensure that the estimated values of mean and standard deviation of travel time would be estimated from at least 8 travel-time records. Ideally more than 8 values would be selected. However, by increasing this value then the number of valid validation records would be reduced. It was felt that 7 other travel time records represented a balance between having sufficient records to calculate the mean and standard deviation of travel time, and having a representative sample of travel time records from all periods throughout the day, including in off-peak times. A total of 20 valid records from the validation day were thus randomly selected to act as the validation data.

Because of the random selection of the validation day, validation data, and training data, this whole process was repeated a total of 50 times. Thus within the experiment a total of 1000 validation records were considered.
7.3.2 Methodology

The objective of the experiment was to investigate whether records with feature vectors close to one another in feature space come from a similar travel time distribution. To achieve this an experiment, based on the k-NN method, was set up. The methodology used to run this experiment is presented in pseudo-code form in the list below:

1. A total of 20 validation records and 15,000 training records were selected as outlined in section 7.3.1.
2. A single validation record was selected from the 20 validation records.
3. The distance between the validation record was calculated to all 15,000 historical records. The Mahalanobis distance metric was used.
4. The 15,000 historical records were sorted by their distance to the validation record.
5. The 5000 closest records were then grouped together into bins of size 50 records (i.e. ranks 1-50, ranks 51-100, ..., ranks 4951-5000) to create a total of 100 bins.
6. A test was then performed on the travel time records in each of these 100 bins to see whether they came from the same underlying distribution as the validation record. This test is described in section 7.3.3.
7. Steps 2 to 6 were repeated with the other 19 validation records.
8. Steps 1 to 7 were repeated using another 49 sets of validation records as outlined in section 7.3.1.

7.3.3 Test of equivalent distributions

A test was performed on each bin of 50 historical records selected by the k-NN method to determine whether the distributional properties of travel time of the historical record and validation record were significantly different. To achieve this the travel time distribution of both the validation and historical record had to be characterised, using the mean and standard deviation.

The travel time distribution of the validation record was characterised by firstly finding those travel time records which lay within 450 seconds (i.e. a total window of 15 minutes) from the validation record. Previous screening described in section 7.3.1 ensured that there were at least 7 such records. These records were used to calculate the mean and standard deviation of the travel time distribution. Similarly the travel time distribution of the 50 historical records in each bin was characterised by calculating the mean and standard deviation of the 50 historical records.

Because there tended to be less than 20 records in each 15-minute window, it was not possible to use statistical tests such as the Chi-Square test or Kolmogorov-Smirnov test to determine whether the validation and historical records came from the same travel
time distribution. Instead an assumption was then made that the underlying travel time distribution was normal in shape, and hence the travel time distribution could be characterised using only the mean and standard deviation. Such an assumption is consistent with underlying empirical evidence (Smeed & Jeffcoate 1971; Dandy & McBean 1984).

However it is realised that an assumption of a normal distribution does allow the possibility, albeit very small, of negative travel times. The following test statistic, $z$, was then used to estimate whether the travel time of the validation and historical record came from the same underlying travel time distribution. (Kottegoda & Rosso 1998, p.271).

$$z = \frac{\mu_v - \mu_h}{\sigma_{\mu_{vh}}}$$  \hspace{1cm} (7.3)

where

$$\sigma_{\mu_{vh}} = \left( \frac{\sigma_v^2}{n_v} + \frac{\sigma_h^2}{n_h} \right)^{\frac{1}{2}}$$  \hspace{1cm} (7.4)

$z$ = test statistic

$n_v$ = no of travel time records used to calculate the characteristics of the validation record travel time distribution (i.e. all records within 450 seconds of the validation record).

$\mu_v$ = mean travel time of the validation record travel time distribution

$\sigma_v$ = standard deviation of the validation record travel time distribution

$\mu_h$ = mean travel time of the historical record travel time distribution

$\sigma_h$ = standard deviation of the historical record travel time distribution

$n_h$ = no of travel time records used to calculate the characteristics of the historical record travel time distribution.

Using the 95% confidence limit, the travel time records are said to come from different distributions if $z > 1.96$.

### 7.3.4 Results

In step 6 of section 7.3.2 a test, outlined in section 7.3.3, was undertaken to see whether the 50 travel time records in each of the 100 bins came from the same underlying distribution as the validation record. The results of this test were averaged over the 1000 validation records, to determine the percentage of records that belonged to a different travel time distribution (i.e. failed the test given by equation 7.3) for each of the 100 bins. The results are presented in figure 7.1. It should be noted that the x-axis contains 100 points (i.e. ranks 1-50, 51-100, ..., 4951 - 5000 are grouped together).

From figure 7.1 it can be seen that as the distance between the feature vectors of the validation and historical dataset increases (x-axis), so the likelihood that the two travel times of each record come from different underlying travel time distributions also increases (y-axis). This suggests that the primary assumption of the k-NN method is met. That is feature vectors which are close to one another in feature space are more likely to come...
7.3. Testing Basic Assumptions

Figure 7.1: Relationship between the probability that the validation and historical records have the same underlying travel time distribution with the rank of the distance between the two.

from records which have the same underlying travel time distribution.

7.3.5 Conclusion

The above experiment has clearly shown that feature vectors close to one another in feature space tend to have travel-times which come from the same underlying travel time distribution. Thus the most crucial assumption of the k-NN method is acceptable when the k-NN method is used to estimate travel time using flow and occupancy data from ILDs. In effect the aim of the k-NN model is to classify each feature vector as belonging to a certain underlying travel-time distribution. When the k-NN model locates the $k$ nearest neighbors it is effectively taking a sample of $k$ realisations of the underlying travel-time distribution. The next section considers how the k-NN method can be optimised for use as a ULTT model.
7.4 Determining the optimal k-NN settings

In section 6.4 the four key design parameters of the k-NN ULTT method were identified as follows:

- Attributes to include in the feature vector
- Distance Metric
- Local Estimation Method
- Number of nearest neighbors, \( k \), to use

This section describes experimental work that was carried out to identify what the optimal settings of these parameters were when the k-NN was used as a ULTT model.

7.4.1 Methodology

This section outlines the approach that was taken in identifying the optimal settings of the k-NN method. To this author’s knowledge there are no comprehensive guidelines in the literature for applying the k-NN as a non-parametric regression model to real world problems. Thus it was necessary to devise a 4-stage approach to identifying the optimal parameters for the k-NN ULTT model. This is outlined briefly below and illustrated in figure 7.2.

![Figure 7.2: Methodology for optimising the k-NN ULTT model](image)

Stage #1
Parameters are considered independently to eliminate levels which are obviously sub-optimal

Stage #2
Identify the optimal distance metric and local estimation method. Identify the approximate optimal value of \( k \).

Stage #3
Find the optimal value of \( k \) to a sufficiently high precision

Stage #3a
Model RMSE

Stage #3b
Line Search

Stage #4
Determine the sensitivity of the optimal k-NN ULTT model.

Stage #1
Pre-screening

Stage #2
Full Factorial Analysis

Stage #4
Sensitivity Analysis

Figure 7.2: Methodology for optimising the k-NN ULTT model
• **Stage #1: Pre-screening** (section 7.5): The task of identifying the optimal settings was made more difficult since the parameters are not independent of one another- i.e. changing one parameter may change the optimal value of another parameter. Thus it was desirable to perform a full factorial test where all combinations of the settings were tested. Since the k-NN method is computationally intensive such an approach would take too long if constraints were not imposed on the number of levels for each factor. Thus before the full factorial experiment was undertaken it was necessary to identify those levels for each factor which were obviously worse than the others. This was achieved by first considering the factors independently of one another. This pre-screening reduced the number of levels for each factor to a workable size.

• **Stage #2: Full factorial analysis** (section 7.6): Having pre-screened all levels which are obviously sub-optimal, it is possible to undertake a full-factorial analysis of the k-NN ULTT model. A full factorial analysis provides an approximate measure of the performance of the k-NN ULTT model for all combinations of the settings. This stage will identify the likely optimal combination of Local Estimation Method (LEM) and Distance Metric. However, it will only provide an approximate indicator of the optimal value of the number of nearest neighbors, $k$, since due to computational reasons it was only possible to use 14 values of $k$ using this optimisation approach.

• **Stage #3a: Modelling the error** (section 7.7): In order to identify the optimal value of $k$, research was undertaken to model the relationship between the RMSE of the model estimate and the value of $k$. A study was undertaken to determine whether the RMSE could be modelled using equation 6.5. In addition an attempt was made using other curve-fitting models to model the relationship between $k$ and the RMSE.

• **Stage #3b: Line Search** (section 7.8): An alternative approach to identifying the optimal value of the value of $k$ to a higher precision is to perform a line search. This section explores the use of a Fibonacci line search.

• **Stage #4: Sensitivity Analysis** (section 7.9): After stage #3 the optimal architecture of the k-NN ULTT model will have been identified. The final stage is to assess how sensitive this k-NN ULTT model is to variations in any of the settings.

The second stage, the full factorial grid search, involves assessing the performance of the k-NN ULTT model for every possible combination of levels for all factors. In order to make such an approach feasible it is desirable to limit the number of factors, and the number of levels for each factor. For this reason it was decided to reduce the number of design parameters from 4 to 3 by fixing the feature vector. The remaining three design parameters were the Local Estimation Method (LEM), Distance Metric (DM), and the number of nearest neighbors ($k$).
7.5. Stage #1: Factors considered independently of one-another

7.4.2 Data

For all experiments carried out in sections 7.5 and 7.9, the Russell Square 2003 dataset, described in section 4.7 was used. Each record contained a date-stamp, 15-minute aggregated occupancy, and flow data from three ILDs, as well as the individual vehicle link travel time. Validation records and training (historical) records were chosen in the manner described in section 7.3.1. This ensured independence between the two data sets. Due to the random selection of validation and training datasets, each experiment was repeated multiple times using different randomly selected validation and training data.

7.4.3 Performance Evaluation

The performance of the ULTT model was quantified by calculating the MAPE and RMSE between the estimated travel time from the k-NN ULTT model and the actual travel time (see equations 2.17 and 2.18). In these equations \(x_i\) refers to the actual travel time from an individual travel time record, and \( \hat{x}_i \) refers to the estimate of travel time for the same individual vehicle from the k-NN ULTT model. \(L\) refers to the total number of validation records used.

Unlike much of the literature, two separate ways of measuring the accuracy have been used. This is because the MAPE and RMSE metrics have slightly different qualities. By close inspection of equations 2.17 and 2.18 it can be seen that a ULTT model which provides accurate estimates at high actual travel times, but poor estimates at low actual travel times will have good RMSE properties but poor MAPE properties. Similarly a ULTT model which provides accurate estimates at low actual travel times, but poor estimates at high actual travel times will have poor RMSE properties but good MAPE properties. It is thus desirable for the k-NN ULTT model to have good RMSE and MAPE properties.

Throughout sections 7.5 and 7.9 it may be noted that the overall best possible performance of the k-NN ULTT model varies slightly from experiment to experiment. This is due to the different validation and historical datasets used in each experiment.

7.5 Stage #1: Factors considered independently of one-another

The objective of stage #1 of the optimisation methodology is to quickly identify those levels of each factor which were obviously much poorer than the rest. These levels will then be removed from further consideration. This is a pre-condition to performing a full factorial experiment. To identify those levels of each factor which were obviously much poorer than the rest, each of the 3 factors were studied independently of one another.
7.5.1 Methodology and Data

It was initially envisaged to undertake a total of three experiments. In each experiment two of the key design parameters would be held constant and the other one would be varied. However, initial research when seeking default parameter settings, suggested that there was a very strong correlation between the optimal LEM and the optimal value of \(k\). Thus only one experiment was undertaken in this section, that of varying the distance metric. The default value of \(k\) was set to 200, and the default LEM was the median. These default attributes were used since initial research suggested that these were close to the optimal levels.

In this experiment the data in section 7.4.2 was used. The experiment was repeated 50 times to take into account the random selection of validation and historical records. The performance of the ULTT models was evaluated as outlined in section 7.4.3.

<table>
<thead>
<tr>
<th>Distance Metric (see section 6.4.2)</th>
<th>Mean MAPE, %</th>
<th>Mean RMSE, secs</th>
<th>Mean computation time, secs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE-STD</td>
<td>17.95</td>
<td>104.3</td>
<td>12.6</td>
</tr>
<tr>
<td>SE-VAR</td>
<td>17.59</td>
<td>102.3</td>
<td>12.6</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>17.83</td>
<td>103.3</td>
<td>12.7</td>
</tr>
<tr>
<td>UnitMap</td>
<td>17.53</td>
<td>102.3</td>
<td>34.1</td>
</tr>
<tr>
<td>QiSmith - 10 bins</td>
<td>18.70</td>
<td>106.0</td>
<td>181.6</td>
</tr>
<tr>
<td>QiSmith - 50 bins</td>
<td>18.88</td>
<td>107.1</td>
<td>182.5</td>
</tr>
</tbody>
</table>

Table 7.3: Performance of each distance metric for use in estimating ULTT using the k-NN method.

7.5.2 Results

The effect of the distance metric (DM) on the performance of the k-NN was studied. From results given in table 7.3 it is seen that the Standardized Euclidean normalised by variance (SE-VAR) and UnitMap had the best performance when measured by both RMSE and MAPE. However, mapping the data to the unit hypercube (UnitMap) required around 3 times more computation time than the Standardized Euclidean based method. The Mahalanobis metric also performed well although not as well as the SE-VAR metric. This result seems to contradict the theoretical prediction of Fukunaga & Hostetler (1973), although it agrees with the conclusions drawn by Franco-Lopez et al. (2001) in their use of the k-NN method for the mapping of forest stand density. Noticeably the SE-VAR metric outperformed that Standardized Euclidean normalised by standard deviation (SE-STD). This is expected since as noted in section 6.4.2, the SE-STD metric is not dimensionless. All three variants of the Standardized Euclidean and Mahalanobis distance metric greatly outperformed the Qi & Smith (2004) method. This latter method was also very slow. This was due to two reasons. Firstly this method requires a large amount of time to initialise the distance matrix, \(D_{a,b}\) (see equation 6.13). Secondly it is necessary to constantly use \(D_{a,b}\) as a 'look-up table'. Because of its poor performance the Qi & Smith (2004) metric
was excluded from any further use.

7.5.3 Conclusions

The work carried out in this section suggests that it is only possible to omit the Qi & Smith (2004) distance metrics from further consideration. This research suggested that there was strong correlation between the optimal value of \( k \) and the local estimation method (LEM). It was therefore not possible to exclude any LEM at this stage, since it was necessary to consider each LEM-\( k \) combination.

7.6 Stage #2: Full factorial grid search

A full factorial grid search is a comprehensive yet computationally inefficient way of locating the optimal settings. It involves assessing the performance of the k-NN ULTT model for every possible combination of level for all factors. In order to make such an approach feasible it is desirable to limit the number of factors, and the number of levels for each factor. For this reason it was decided to reduce the number of factors from 4 to 3 by fixing the feature vector to include the flow and occupancy from each of the 3 detectors outlined in section 4.7. In addition, from the work of the pre-screening of distance metrics, presented in table 7.3, it was possible to eliminate the Qi and Smith distance metrics from further study.

7.6.1 Methodology and Data

The full factorial grid search involved assessing the performance of the ULTT model for all possible combinations of the key-parameters. The following levels were considered for each factor:

**Distance Metric** - The following four distance metrics were tested:

- **SE-STD** - Standardised Euclidean where all values are normalised by the standard deviation of the attribute - see equation 6.9.
- **SE-VAR** - Standardised Euclidean where all values are normalised by the variance of the attribute - see equation 6.9.
- Mahalanobis - see equation 6.10.
- **UnitMap** - all values are first mapped onto a unit hypersphere, such that the minimum value of each attribute is mapped to zero and the maximum value of each attribute is mapped to one. The Euclidean distance is then used - see equation 6.7.

**Local Estimation Method** - The following four local estimation methods were tested:

- mean
• median

• Regression - Ordinary Least Squares Regression. Refer to section 6.4.3.

• Lowess - refer to section 6.4.3.

**Number of nearest neighbors, k** - Unlike the first two factors, this factor has many possible levels - all integer values up to the size of the database. It was decided to select 14 values of $k$ to include in the full factorial. Ideally more would have been chosen, but this would have resulted in increased computation time making the full-factorial approach infeasible. Equations 6.3 and 6.5 suggest that the k-NN model is sensitive at low values of $k$. Thus a sampling strategy was devised to sample intensively at low values of $k$ yet still include samples at very high values of $k$. For the purposes of this experiment, the following 14 levels of $k$ were chosen: 20, 50, 100, 200, 300, 400, 500, 750, 1000, 1500, 2000, 3000, 4000, 5000.

In all there were a total of 224 different combinations of the settings. The k-NN ULTT model was assessed using each of these 224 different combinations. In this experiment the data described in section 7.4.2 was used. Due to the random nature of selecting the validation and historical data sets, the process of testing the performance of each of the 224 combinations of k-NN ULTT model was undertaken 25 times. Each iteration a new historical and validation dataset was chosen. The performance of the ULTT models was evaluated as outlined in section 7.4.3.

### 7.6.2 Results

This section presents the results of the full-factorial research when performance was calculated by MAPE, RMSE, and a hybrid measure.

**Results: MAPE**

Table 7.4 shows the combinations which gave the best average performance when MAPE was used as the performance index.

The results of table 7.4 suggest that the median LEM is clearly superior to all other LEMs when performance of the k-NN ULTT model is measured by MAPE. The performance of the model seems to be insensitive to the choice of distance metric although the SE-STD does not seem to perform as well as the other 3 DMs. This is not a surprising result, since it has already been noted that SE-STD is not dimensionless whereas the other three are.

**Results: RMSE**

Table 7.5 shows the combinations which gave the best average performance when RMSE was used as the performance index.
### Table 7.4: Best settings when performance is measured by MAPE

<table>
<thead>
<tr>
<th>Rank</th>
<th>Distance metric</th>
<th>Local estimation method</th>
<th>k</th>
<th>Rank MAPE</th>
<th>MAPE %</th>
<th>Rank RMSE</th>
<th>RMSE /secs</th>
<th>Rank Time /secs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UnitMap</td>
<td>median</td>
<td>400</td>
<td>1</td>
<td>18.51</td>
<td>96</td>
<td>103.4</td>
<td>178</td>
</tr>
<tr>
<td>2</td>
<td>UnitMap</td>
<td>median</td>
<td>300</td>
<td>2</td>
<td>18.51</td>
<td>92</td>
<td>103.0</td>
<td>162</td>
</tr>
<tr>
<td>3</td>
<td>UnitMap</td>
<td>median</td>
<td>200</td>
<td>3</td>
<td>18.53</td>
<td>90</td>
<td>102.9</td>
<td>146</td>
</tr>
<tr>
<td>4</td>
<td>SE-VAR</td>
<td>median</td>
<td>200</td>
<td>4</td>
<td>18.54</td>
<td>82</td>
<td>102.4</td>
<td>38</td>
</tr>
<tr>
<td>5</td>
<td>Mahalanobis</td>
<td>median</td>
<td>200</td>
<td>5</td>
<td>18.55</td>
<td>86</td>
<td>102.8</td>
<td>46</td>
</tr>
<tr>
<td>6</td>
<td>SE-VAR</td>
<td>median</td>
<td>300</td>
<td>6</td>
<td>18.55</td>
<td>89</td>
<td>102.8</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>UnitMap</td>
<td>median</td>
<td>500</td>
<td>7</td>
<td>18.55</td>
<td>106</td>
<td>104.1</td>
<td>194</td>
</tr>
<tr>
<td>8</td>
<td>UnitMap</td>
<td>median</td>
<td>100</td>
<td>8</td>
<td>18.56</td>
<td>77</td>
<td>102.2</td>
<td>138</td>
</tr>
<tr>
<td>9</td>
<td>SE-VAR</td>
<td>median</td>
<td>400</td>
<td>9</td>
<td>18.57</td>
<td>97</td>
<td>103.4</td>
<td>62</td>
</tr>
<tr>
<td>10</td>
<td>Mahalanobis</td>
<td>median</td>
<td>100</td>
<td>10</td>
<td>18.61</td>
<td>67</td>
<td>101.7</td>
<td>30</td>
</tr>
</tbody>
</table>

The results of table 7.5 suggest that the regression based LEMs (i.e. regression and Lowess) are superior to the mean and median when performance is measured by RMSE. The performance of the model seems to be insensitive to the choice of distance metric, even the SE-STD metric. Since the good performing combinations which included the SE-STD all use regression based LEMs, this suggests that regression based LEMs are less sensitive to distance metrics that are not dimensionless than the median and mean LEMs. This could be because the parameters of the linear model (see equation 6.18) used by the regression based LEMs are able to model and take into account the dimensional effect of the SE-STD metric.

### Table 7.5: Best settings when performance is measured by RMSE

<table>
<thead>
<tr>
<th>Rank</th>
<th>Distance metric</th>
<th>Local estimation method</th>
<th>k</th>
<th>Rank MAPE</th>
<th>MAPE %</th>
<th>Rank RMSE</th>
<th>RMSE /secs</th>
<th>Rank Time /secs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UnitMap</td>
<td>regression</td>
<td>1500</td>
<td>86</td>
<td>20.75</td>
<td>1</td>
<td>99.2</td>
<td>207</td>
</tr>
<tr>
<td>2</td>
<td>SE-STD</td>
<td>regression</td>
<td>3000</td>
<td>108</td>
<td>20.84</td>
<td>2</td>
<td>99.2</td>
<td>155</td>
</tr>
<tr>
<td>3</td>
<td>SE-VAR</td>
<td>regression</td>
<td>1500</td>
<td>84</td>
<td>20.74</td>
<td>3</td>
<td>99.3</td>
<td>115</td>
</tr>
<tr>
<td>4</td>
<td>SE-VAR</td>
<td>regression</td>
<td>2000</td>
<td>87</td>
<td>20.76</td>
<td>4</td>
<td>99.3</td>
<td>131</td>
</tr>
<tr>
<td>5</td>
<td>Mahalanobis</td>
<td>regression</td>
<td>1500</td>
<td>111</td>
<td>20.87</td>
<td>5</td>
<td>99.4</td>
<td>123</td>
</tr>
<tr>
<td>6</td>
<td>Mahalanobis</td>
<td>regression</td>
<td>3000</td>
<td>82</td>
<td>20.72</td>
<td>6</td>
<td>99.4</td>
<td>159</td>
</tr>
<tr>
<td>7</td>
<td>UnitMap</td>
<td>regression</td>
<td>3000</td>
<td>78</td>
<td>20.67</td>
<td>7</td>
<td>99.4</td>
<td>215</td>
</tr>
<tr>
<td>8</td>
<td>SE-STD</td>
<td>regression</td>
<td>4000</td>
<td>117</td>
<td>20.93</td>
<td>8</td>
<td>99.4</td>
<td>171</td>
</tr>
<tr>
<td>9</td>
<td>Mahalanobis</td>
<td>regression</td>
<td>2000</td>
<td>114</td>
<td>20.88</td>
<td>9</td>
<td>99.4</td>
<td>143</td>
</tr>
<tr>
<td>10</td>
<td>UnitMap</td>
<td>lowess</td>
<td>1500</td>
<td>25</td>
<td>18.99</td>
<td>10</td>
<td>99.5</td>
<td>208</td>
</tr>
</tbody>
</table>

### Results: Combined measure of performance

What is clearly noticeable from tables 7.4 and 7.5 is that those settings with a low rank in MAPE tend not to have such a good ranking when performance is measured by RMSE. Likewise those settings with a low RMSE rank tend not to have a low MAPE. Reasons for differences in the RMSE and MAPE results have been given in section 7.4.3. Ideally the optimal settings should perform well when measured by both RMSE and by MAPE. Ideally they should also be computationally quick. To determine the overall settings, the
following performance index, $PI$, was defined in equation 7.5.

$$PI = \text{RankMAPE} + \text{RankRMSE} + \frac{\text{rankTime}}{5}$$  \hspace{1cm} (7.5)

This performance index identified those settings with good MAPE and RMSE properties whilst at the same time preferring those settings which were computationally quick. The results of this are presented in table 7.6.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Distance metric</th>
<th>Local estimation method</th>
<th>k</th>
<th>Rank MAPE</th>
<th>MAPE %</th>
<th>Rank RMSE</th>
<th>RMSE /secs</th>
<th>Rank Time</th>
<th>Time /secs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SE-VAR</td>
<td>lowess</td>
<td>1500</td>
<td>26</td>
<td>19.00</td>
<td>13</td>
<td>99.5</td>
<td>116</td>
<td>34.9</td>
</tr>
<tr>
<td>2</td>
<td>SE-VAR</td>
<td>lowess</td>
<td>2000</td>
<td>33</td>
<td>19.08</td>
<td>15</td>
<td>99.6</td>
<td>132</td>
<td>38.6</td>
</tr>
<tr>
<td>3</td>
<td>Mahalanobis</td>
<td>lowess</td>
<td>1500</td>
<td>31</td>
<td>19.07</td>
<td>20</td>
<td>99.6</td>
<td>124</td>
<td>35.4</td>
</tr>
<tr>
<td>4</td>
<td>UnitMap</td>
<td>lowess</td>
<td>1500</td>
<td>25</td>
<td>18.99</td>
<td>10</td>
<td>99.5</td>
<td>208</td>
<td>139.8</td>
</tr>
<tr>
<td>5</td>
<td>Mahalanobis</td>
<td>median</td>
<td>100</td>
<td>10</td>
<td>18.61</td>
<td>67</td>
<td>101.7</td>
<td>30</td>
<td>15.6</td>
</tr>
<tr>
<td>6</td>
<td>Mahalanobis</td>
<td>lowess</td>
<td>1000</td>
<td>36</td>
<td>19.10</td>
<td>27</td>
<td>99.7</td>
<td>108</td>
<td>31.8</td>
</tr>
<tr>
<td>7</td>
<td>SE-VAR</td>
<td>median</td>
<td>200</td>
<td>4</td>
<td>18.54</td>
<td>82</td>
<td>102.4</td>
<td>38</td>
<td>18.0</td>
</tr>
<tr>
<td>8</td>
<td>SE-VAR</td>
<td>median</td>
<td>100</td>
<td>14</td>
<td>18.69</td>
<td>76</td>
<td>102.2</td>
<td>26</td>
<td>15.5</td>
</tr>
<tr>
<td>9</td>
<td>Mahalanobis</td>
<td>lowess</td>
<td>2000</td>
<td>40</td>
<td>19.14</td>
<td>28</td>
<td>99.7</td>
<td>144</td>
<td>39.1</td>
</tr>
<tr>
<td>10</td>
<td>SE-VAR</td>
<td>lowess</td>
<td>1000</td>
<td>35</td>
<td>19.10</td>
<td>44</td>
<td>100.0</td>
<td>100</td>
<td>31.5</td>
</tr>
</tbody>
</table>

Table 7.6: Best settings. $PI = \text{rank MAPE} + \text{rank RMSE} + (\text{rank time} / 5)$

The results of table 7.6 suggest that the best overall settings are those based upon a robust LEM. Indeed, the Lowess LEM, based upon the principle of robust regression appears to be the best choice. Interestingly those settings which use a Lowess LEM tend to have a far higher optimal value of $k$ than those settings that use a median LEM. This suggests that the optimal value of $k$ is strongly related to the LEM. Once again, the performance of the model seems to be insensitive to the choice of distance metric, although the SE-STD distance metric appears to be inferior to the other three.

7.6.3 Conclusions from the full factorial approach

The research in this section has indicated that a k-NN model which uses the Lowess LEM and the SE-VAR DM and which uses between 1000 and 2000 nearest neighbors provides a good solution. It also suggests a strong correlation between the optimal LEM and the optimal value of $k$. The selection of the distance metric does not appear to be important.

The full factorial technique is a rather blunt approach to optimising the k-NN ULTT model. One limitation of the technique is that due to the very high number of possible values of nearest neighbors, $k$, this approach cannot be used to precisely determine the optimal value of $k$. Sections 7.7 and 7.8 investigate some approaches that can be used to identify the optimal value of $k$ to a greater precision for a given LEM and distance metric.
7.7 Stage #3a: Modelling the error

Since the number of nearest neighbors, $k$, is a continuous (integer) variable, it is not computationally feasible to find the optimal value of $k$ to a good precision for a given combination of Local Estimation Method (LEM) and Distance Metric (DM) using the full factorial grid search. Stage #3 of the optimisation methodology attempts to overcome this. An approach to determining the optimal value of $k$ to a high precision is to attempt to model the error of the travel-time estimate as a function of $k$ for each fixed LEM and DM combination. Figures 7.3 and 7.4 show plots of this relationship using results from the full-factorial experiment. Values of error between the 14 selected values of $k$ (knots) are obtained by simple linear interpolation. Ideally it is desired to fit a smooth line through the data without distinct changes in gradient at the knots. This section discusses several techniques of achieving this. Section 7.7.1 attempts to fit an equation which has been suggested by the literature whilst section 7.7.2 attempts the use of some general curve fitting techniques and in particular attempts to fit a cubic spline onto the data.

Already, figure 7.3 and figure 7.4 clearly show that settings using the median or mean LEMs are relatively sensitive to the value of $k$ chosen, whereas the regression based LEMs tend not to be sensitive to the value of $k$ once the value of $k$ is greater than a certain amount.

7.7.1 Fitting the theoretical MSE curve

The first approach to fitting a line through the data, recognises that equation 6.5 suggests that the Mean Square Error (i.e. RMSE squared) can be modelled by equation 7.6. Figure 7.4 suggests that this will be a minima rather than a maxima.

$$MSE = \alpha_1 + \frac{\alpha_2}{k} + \alpha_3 \cdot k^\frac{4}{n}$$ (7.6)

The optimal value of $k$ can then be found by finding the point where the gradient of equation 7.6 is 0, i.e. $d(MSE)/d(k) = 0$. This optimal value of $k$, $k^*$, is then given by equation 7.7.

$$k^* = \left[ \frac{n \cdot \alpha_2}{4 \cdot \alpha_3} \right]^\frac{n}{4+n}$$ (7.7)

This approach was carried out on the data for all 16 different combinations of LEM and distance metric generated from the full-factorial grid search. The results are presented in table 7.7. It can be seen that this approach produced nonsensical results (i.e. negative error which have been limited to -999 for purpose of display) for the regression and Lowess models, as well as two of the median based models. The poor performance of the model for these LEMs can be explained since equation 7.6 relating MSE to the value of $k$ was formulated where the LEM was a weighted mean. For the mean-based models, the performance of the model was seen to be very good with high $r^2$ values. For this
7.7. Stage #3a: Modelling the error

Figure 7.3: Variation of MAPE with value of k for each LEM and distance metric combination.

case, all parameters ($\alpha_1, \alpha_2, \alpha_3$) were seen to be significant at the 95% level (refer to the corresponding t-statistics, $t_1, t_2, t_3$). As an example of this work figure 7.5 shows the fitted error curve for the settings where LEM = mean, DM = SE-VAR, and where $n$ was set to 6.

In the results of table 7.7, the number of dimensions, $n$, was set to 6 since there were 6 attributes in the feature vector. It was noticed that a better fit to equation 7.6 was often found to exist when $n$ was set to 8. The reasons for this are unknown.

Two noticeable problems were encountered when undertaking this MSE curve fitting. Firstly this method was not successful at modelling the MSE error for combinations which used the regression or Lowess LEM. Secondly, although the curve fitted to combinations using the mean or median LEM tended to have a high $r^2$ value, it was felt that in several cases with a slightly lower $r^2$ value the fitted MSE curve failed to correctly model the error at lower values of $k$. Thus the estimate of the optimal value of $k$ was obviously faulty. An example of this is in the modelling of the MSE where the LEM = median and the DM = Mahalanobis. This is shown in figure 7.6. Thus an alternative way of estimating the optimal value of $k$ for each combination of LEM and DM was required. This was achieved
7.7. Stage #3a: Modelling the error

![Graphs showing variation of RMSE with value of k for each LEM and distance metric combination.](image)

Figure 7.4: Variation of RMSE with value of k for each LEM and distance metric combination.

<table>
<thead>
<tr>
<th>Local Estimation Method</th>
<th>Distance metric</th>
<th>optimal k</th>
<th>RMSE</th>
<th>$r^2$</th>
<th>$\alpha_1$</th>
<th>$t_1$</th>
<th>$\alpha_2$</th>
<th>$t_2$</th>
<th>$\alpha_3$</th>
<th>$t_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>SE-STD</td>
<td>74.2</td>
<td>95.76</td>
<td>0.977</td>
<td>8867</td>
<td>108.80</td>
<td>9001</td>
<td>2.52</td>
<td>10.30</td>
<td>19.55</td>
</tr>
<tr>
<td>mean</td>
<td>SE-VAR</td>
<td>132.1</td>
<td>93.58</td>
<td>0.997</td>
<td>8429</td>
<td>422.37</td>
<td>17347</td>
<td>19.83</td>
<td>7.59</td>
<td>58.84</td>
</tr>
<tr>
<td>mean</td>
<td>Mahalanobis</td>
<td>103.0</td>
<td>93.75</td>
<td>0.986</td>
<td>8414</td>
<td>141.64</td>
<td>15394</td>
<td>5.91</td>
<td>10.21</td>
<td>26.58</td>
</tr>
<tr>
<td>mean</td>
<td>UnitMap</td>
<td>124.3</td>
<td>93.50</td>
<td>0.999</td>
<td>8447</td>
<td>750.40</td>
<td>14697</td>
<td>29.78</td>
<td>7.12</td>
<td>97.78</td>
</tr>
<tr>
<td>median</td>
<td>SE-STD</td>
<td>-999.0</td>
<td>-999.0</td>
<td>0.954</td>
<td>9149</td>
<td>72.07</td>
<td>-999</td>
<td>-1.16</td>
<td>10.10</td>
<td>12.30</td>
</tr>
<tr>
<td>median</td>
<td>SE-VAR</td>
<td>70.7</td>
<td>93.83</td>
<td>0.985</td>
<td>8557</td>
<td>155.41</td>
<td>7007</td>
<td>2.90</td>
<td>8.69</td>
<td>24.42</td>
</tr>
<tr>
<td>median</td>
<td>Mahalanobis</td>
<td>-999.0</td>
<td>-999.0</td>
<td>0.971</td>
<td>8700</td>
<td>83.13</td>
<td>-5</td>
<td>-0.00</td>
<td>11.13</td>
<td>16.45</td>
</tr>
<tr>
<td>median</td>
<td>UnitMap</td>
<td>71.5</td>
<td>93.81</td>
<td>0.992</td>
<td>8560</td>
<td>218.98</td>
<td>6843</td>
<td>3.99</td>
<td>8.33</td>
<td>32.95</td>
</tr>
<tr>
<td>regression</td>
<td>SE-STD</td>
<td>-999.0</td>
<td>-999.0</td>
<td>0.972</td>
<td>9385</td>
<td>23.23</td>
<td>99999</td>
<td>5.20</td>
<td>-1.99</td>
<td>-1.99</td>
</tr>
<tr>
<td>regression</td>
<td>SE-VAR</td>
<td>-999.0</td>
<td>-999.0</td>
<td>0.963</td>
<td>9343</td>
<td>19.64</td>
<td>99999</td>
<td>13.49</td>
<td>-5.13</td>
<td>-1.67</td>
</tr>
<tr>
<td>regression</td>
<td>Mahalanobis</td>
<td>-999.0</td>
<td>-999.0</td>
<td>0.971</td>
<td>9859</td>
<td>23.70</td>
<td>99999</td>
<td>14.71</td>
<td>-7.35</td>
<td>-2.73</td>
</tr>
<tr>
<td>regression</td>
<td>UnitMap</td>
<td>-999.0</td>
<td>-999.0</td>
<td>0.971</td>
<td>9274</td>
<td>22.48</td>
<td>99999</td>
<td>15.56</td>
<td>-4.78</td>
<td>-1.79</td>
</tr>
<tr>
<td>lowess</td>
<td>SE-STD</td>
<td>-999.0</td>
<td>-999.0</td>
<td>0.759</td>
<td>12402</td>
<td>7.74</td>
<td>99999</td>
<td>3.74</td>
<td>-20.37</td>
<td>-1.96</td>
</tr>
<tr>
<td>lowess</td>
<td>SE-VAR</td>
<td>-999.0</td>
<td>-999.0</td>
<td>0.715</td>
<td>13128</td>
<td>7.12</td>
<td>99999</td>
<td>3.11</td>
<td>-24.02</td>
<td>-2.01</td>
</tr>
<tr>
<td>lowess</td>
<td>Mahalanobis</td>
<td>-999.0</td>
<td>-999.0</td>
<td>0.724</td>
<td>13759</td>
<td>7.60</td>
<td>99999</td>
<td>2.97</td>
<td>-26.91</td>
<td>-2.30</td>
</tr>
<tr>
<td>lowess</td>
<td>UnitMap</td>
<td>-999.0</td>
<td>-999.0</td>
<td>0.742</td>
<td>12422</td>
<td>7.32</td>
<td>99999</td>
<td>3.55</td>
<td>-20.82</td>
<td>-1.90</td>
</tr>
</tbody>
</table>

Table 7.7: Modelling the RMSE - fitting equation 7.6 to the data
Figure 7.5: Fitted error curve for the setting LEM = mean, DM = SE-VAR.
7.7. Stage #3a: Modelling the error

Figure 7.6: Fitted error curve for the setting LEM = median, DM = Mahalanobis
by attempting to fit a cubic spline to the data.

### 7.7.2 Cubic spline fitting

A popular approach to creating a smooth line through data to approximate the true relationship, $f(x)$, is to fit a cubic polynomial, $g(x)$, of the form given in equation 7.8 between two adjacent data points (knots). This process is repeated across all the adjacent data points to create what is known as a cubic spline. This function is often known as a piecewise polynomial.

$$g(x) = \alpha_0 + \alpha_1 \cdot x + \alpha_2 \cdot x^2 + \alpha_3 \cdot x^3 \quad (7.8)$$

Equation 7.8 has four unknowns. To obtain a unique solution, four constraints are needed. Two of these constraints are that the value of $g(x)$ at both the data points must equal the actual value $f(x)$ at these points. The other two constraints ensure that the gradients of $g(x)$ at both knots of each piecewise polynomial are equal to the gradient of $g(x)$ of the adjacent polynomials at this point. The interested reader is referred to de Boor (1978) for more information.

One advantage of fitting the cubic polynomial is that it is trivial to differentiate, and hence easy to determine the minimum point. This approach was thus applied to the data from the full factorial experiment to determine the optimal value of $k$. Figure 7.7 shows the cubic spline fitted to the data of the optimal LEM-DM combination (i.e. LEM = Lowess, and DM = SE-VAR). Because of the very high RMSE at low $k$, the cubic spline curve has an overshoot and takes on a sinusoidal appearance. Due to this sinusoidal nature, the value of the minimum RMSE at $k = 600$ is erroneous, and the optimal value of $k$ is probably higher. An alternative to the cubic spline could be to fit a Hermite curve (de Boor 1978). However, since this is monotonic between knots then this will not be able to estimate the optimum value of $k$. It thus seems that curve fitting is not appropriate for this data. The next section investigates an alternative approach to identify the optimal value of $k$ to a greater precision.

### 7.8 Stage #3b: Line Search

Section 7.7 attempted to find the optimal value of $k$ to a high precision for a given DM and LEM by modelling the relationship between the RMSE and the value of $k$. With the exception of ULTT models which used the mean LEM these models performed poorly. An alternative empirical approach to determine the optimal value of $k$ is to undertake a Fibonacci line search (Cooper & Steinberg 1970). The Fibonacci line search iteratively tests the performance of the model in a fashion such that the value of $k$ approaches a local optimum value.

In the first instance the Lowess LEM and the SE-VAR DM were used since the work of the full factorial approach, presented in table 7.6, suggested that this was the optimal
7.8. Stage #3b: Line Search

Figure 7.7: Fitted error curve for the setting LEM = median, DM = Mahalanobis

<table>
<thead>
<tr>
<th>i</th>
<th>$k_{min}$</th>
<th>$k_{max}$</th>
<th>$k_{small}$</th>
<th>RMSE small</th>
<th>$k_{large}$</th>
<th>RMSE large</th>
<th>precision</th>
<th>optimal $k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>4000</td>
<td>1528</td>
<td>18.119</td>
<td>2473</td>
<td>18.159</td>
<td>4000.0</td>
<td>2000</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2473</td>
<td>945</td>
<td>18.201</td>
<td>1528</td>
<td>18.119</td>
<td>2473.0</td>
<td>1237</td>
</tr>
<tr>
<td>3</td>
<td>945</td>
<td>2473</td>
<td>1528</td>
<td>18.119</td>
<td>1889</td>
<td>18.150</td>
<td>1528.5</td>
<td>1709</td>
</tr>
<tr>
<td>4</td>
<td>945</td>
<td>1889</td>
<td>1305</td>
<td>18.150</td>
<td>1528</td>
<td>18.119</td>
<td>944.5</td>
<td>1417</td>
</tr>
<tr>
<td>5</td>
<td>1305</td>
<td>1889</td>
<td>1528</td>
<td>18.119</td>
<td>1666</td>
<td>18.115</td>
<td>583.9</td>
<td>1597</td>
</tr>
<tr>
<td>6</td>
<td>1528</td>
<td>1889</td>
<td>1666</td>
<td>18.115</td>
<td>1752</td>
<td>18.119</td>
<td>360.6</td>
<td>1709</td>
</tr>
<tr>
<td>7</td>
<td>1528</td>
<td>1752</td>
<td>1614</td>
<td>18.114</td>
<td>1666</td>
<td>18.115</td>
<td>223.3</td>
<td>1640</td>
</tr>
<tr>
<td>8</td>
<td>1528</td>
<td>1666</td>
<td>1580</td>
<td>18.115</td>
<td>1614</td>
<td>18.114</td>
<td>137.4</td>
<td>1597</td>
</tr>
<tr>
<td>9</td>
<td>1580</td>
<td>1666</td>
<td>1614</td>
<td>18.114</td>
<td>1631</td>
<td>18.115</td>
<td>85.9</td>
<td>1623</td>
</tr>
<tr>
<td>10</td>
<td>1580</td>
<td>1631</td>
<td>1597</td>
<td>18.117</td>
<td>1614</td>
<td>18.114</td>
<td>51.5</td>
<td>1606</td>
</tr>
<tr>
<td>11</td>
<td>1597</td>
<td>1631</td>
<td>1614</td>
<td>18.114</td>
<td>1614</td>
<td>18.114</td>
<td>34.3</td>
<td>1614</td>
</tr>
</tbody>
</table>

Table 7.8: Identification the optimal value of $k$ to reduce MAPE. LEM = lowess, DM = SE-VAR
Table 7.9: Identification the optimal value of \( k \) to reduce RMSE. LEM = lowess, DM = SE-VAR

<table>
<thead>
<tr>
<th>i</th>
<th>( k_{\text{min}} )</th>
<th>( k_{\text{max}} )</th>
<th>( k_{\text{small}} )</th>
<th>( \text{RMSE small} )</th>
<th>( k_{\text{large}} )</th>
<th>( \text{RMSE large} )</th>
<th>precision</th>
<th>optimal ( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>4000</td>
<td>1528</td>
<td>101.726</td>
<td>2473</td>
<td>101.621</td>
<td>4000.0</td>
<td>2000</td>
</tr>
<tr>
<td>2</td>
<td>1528</td>
<td>4000</td>
<td>2472</td>
<td>101.621</td>
<td>3056</td>
<td>101.707</td>
<td>2472.0</td>
<td>2764</td>
</tr>
<tr>
<td>3</td>
<td>1528</td>
<td>3056</td>
<td>2112</td>
<td>101.618</td>
<td>2472</td>
<td>101.621</td>
<td>1527.8</td>
<td>2292</td>
</tr>
<tr>
<td>4</td>
<td>1528</td>
<td>2472</td>
<td>1889</td>
<td>101.657</td>
<td>2112</td>
<td>101.618</td>
<td>944.2</td>
<td>2000</td>
</tr>
<tr>
<td>5</td>
<td>1889</td>
<td>2472</td>
<td>2112</td>
<td>101.618</td>
<td>2249</td>
<td>101.612</td>
<td>583.7</td>
<td>2180</td>
</tr>
<tr>
<td>6</td>
<td>2112</td>
<td>2472</td>
<td>2249</td>
<td>101.612</td>
<td>2335</td>
<td>101.617</td>
<td>360.5</td>
<td>2292</td>
</tr>
<tr>
<td>7</td>
<td>2112</td>
<td>2335</td>
<td>2198</td>
<td>101.609</td>
<td>2249</td>
<td>101.612</td>
<td>223.2</td>
<td>2223</td>
</tr>
<tr>
<td>8</td>
<td>2112</td>
<td>2249</td>
<td>2163</td>
<td>101.601</td>
<td>2198</td>
<td>101.609</td>
<td>137.3</td>
<td>2180</td>
</tr>
<tr>
<td>9</td>
<td>2112</td>
<td>2198</td>
<td>2146</td>
<td>101.602</td>
<td>2163</td>
<td>101.601</td>
<td>85.8</td>
<td>2155</td>
</tr>
<tr>
<td>10</td>
<td>2146</td>
<td>2198</td>
<td>2163</td>
<td>101.601</td>
<td>2180</td>
<td>101.604</td>
<td>51.5</td>
<td>2172</td>
</tr>
<tr>
<td>11</td>
<td>2146</td>
<td>2180</td>
<td>2163</td>
<td>101.601</td>
<td>2163</td>
<td>101.601</td>
<td>34.3</td>
<td>2163</td>
</tr>
</tbody>
</table>
combination for the LEM and DM. Figure 7.4 also suggested that there was only one minima and that the local minima found would in fact be the global optimum. A total of 11 iterations were required to determine the optimal point to within a precision of 25 points. The performance of each new $k$ value was evaluated using 250 validation records selected from a single day. The validation and historical data used have been described in section 7.4.2. The performance of the model was evaluated as per section 7.4.3. This process was repeated 50 times to allow for the random selection of the validation and historical records. Ideally more validation records would have been used. However, due to computational constraints this was not possible. Table 7.8 shows the iterations used to identify the optimal value of $k$ when MAPE was used as the performance index. The optimum value of $k$ to reduce the MAPE was found to be 1614.

Table 7.9 shows the iterations used to identify the optimal value of $k$ when RMSE was used as the performance index. The optimum value of $k$ to reduce the RMSE was found to be 2163.

One drawback of the Fibonacci method is that it is very computationally expensive. It took up to one day on a Pentium 4 2.8 GHz computer to perform the optimisation on a single LEM-DM combination. For this reason optimisation results for other LEM-DM combinations are not shown.

7.9 Stage #4: Sensitivity Analysis

Stages 1 to 3 allowed possible candidate optimal k-NN ULTT models to be identified. However, none of the approaches explicitly quantifies the sensitivity of the solution to changes in any of the parameters, nor whether this solution is significantly better than other solutions. Stage #4 thus carried out sensitivity analysis using regression analysis to see which factors and interaction terms significantly affected the performance of the ULTT model.

7.9.1 Methodology and Data

The data used for the sensitivity analysis came from the results of the full factorial grid search carried out in stage #2. The full factorial experiment generated 5600 records (25 runs of 224 different possible combinations of the k-NN ULTT model). A linear regression model of the form given in equation 6.18 was developed to see what variables affected the MAPE and the RMSE. The dependent variables $X$ were dummy variables that signified whether the parameter or interaction term were present in the data record. The independent variable $y$ was the measured error of the ULTT model expressed as either MAPE or RMSE. The work in section 7.6 suggested that the combination of LEM and $k$ was important. Thus dummy variables were defined to represent each of these 56 possible combinations (interaction terms) of LEM and $k$. In addition three variables representing the SE-STD, Mahalanobis, and UnitMap distance metrics were added. The SE-VAR
variable was omitted from the model since this was the reference variable. Thus there were 59 dummy variables in the model. Of the 5600 original records, 428 records had to be discarded since these records had very high errors which could distort the OLS model. These records came primarily from regression based models with very low $k$ values. Thus the regression model was performed with a total of 5172 input records. No significant correlations between the variables in the model were found.

### 7.9.2 Results

**Results: MAPE**

The results of the regression model where the MAPE was the dependent variable are presented in Table 7.10. The value of adjusted $r^2$ for this model was 0.272. For interaction terms of the LEM and value of $k$, the variable name is given by $LEM_k$. For example, $LOW_{400}$ is the interaction term of the Lowess LEM with a $k$ value of 400.

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>21.155</td>
<td>0.307</td>
<td>68.958</td>
<td>0</td>
</tr>
<tr>
<td>SE-STD</td>
<td>0.378</td>
<td>0.162</td>
<td>2.334</td>
<td>0.02</td>
</tr>
<tr>
<td>MAHALANO</td>
<td>0.152</td>
<td>0.162</td>
<td>0.939</td>
<td>0.348</td>
</tr>
<tr>
<td>UNIT-MAP</td>
<td>-0.201</td>
<td>0.162</td>
<td>-1.241</td>
<td>0.215</td>
</tr>
<tr>
<td>LOW-100</td>
<td>7.672</td>
<td>0.988</td>
<td>7.768</td>
<td>0</td>
</tr>
<tr>
<td>LOW-200</td>
<td>1.902</td>
<td>0.56</td>
<td>3.399</td>
<td>0.001</td>
</tr>
<tr>
<td>LOW-300</td>
<td>-0.326</td>
<td>0.524</td>
<td>-0.622</td>
<td>0.534</td>
</tr>
<tr>
<td>LOW-400</td>
<td>-1.265</td>
<td>0.516</td>
<td>-2.451</td>
<td>0.014</td>
</tr>
<tr>
<td>LOW-500</td>
<td>-1.456</td>
<td>0.505</td>
<td>-2.883</td>
<td>0.004</td>
</tr>
<tr>
<td>LOW-750</td>
<td>-1.98</td>
<td>0.504</td>
<td>-3.933</td>
<td>0</td>
</tr>
<tr>
<td>LOW-1000</td>
<td>-2.129</td>
<td>0.504</td>
<td>-4.227</td>
<td>0</td>
</tr>
<tr>
<td>LOW-1500</td>
<td>-2.169</td>
<td>0.504</td>
<td>-4.308</td>
<td>0</td>
</tr>
<tr>
<td>LOW-2000</td>
<td>-2.093</td>
<td>0.504</td>
<td>-4.156</td>
<td>0</td>
</tr>
<tr>
<td>LOW-3000</td>
<td>-2.095</td>
<td>0.504</td>
<td>-4.16</td>
<td>0</td>
</tr>
<tr>
<td>LOW-4000</td>
<td>-1.993</td>
<td>0.504</td>
<td>-3.959</td>
<td>0</td>
</tr>
<tr>
<td>LOW-5000</td>
<td>-1.876</td>
<td>0.504</td>
<td>-3.726</td>
<td>0</td>
</tr>
<tr>
<td>REG-50</td>
<td>12.878</td>
<td>0.509</td>
<td>25.314</td>
<td>0</td>
</tr>
<tr>
<td>REG-100</td>
<td>6.09</td>
<td>0.504</td>
<td>12.094</td>
<td>0</td>
</tr>
<tr>
<td>REG-200</td>
<td>3.617</td>
<td>0.504</td>
<td>7.182</td>
<td>0</td>
</tr>
<tr>
<td>REG-300</td>
<td>1.922</td>
<td>0.504</td>
<td>3.817</td>
<td>0</td>
</tr>
<tr>
<td>REG-400</td>
<td>1.016</td>
<td>0.504</td>
<td>2.019</td>
<td>0.044</td>
</tr>
<tr>
<td>REG-500</td>
<td>0.391</td>
<td>0.504</td>
<td>0.776</td>
<td>0.438</td>
</tr>
<tr>
<td>REG-750</td>
<td>-0.135</td>
<td>0.504</td>
<td>-0.269</td>
<td>0.788</td>
</tr>
<tr>
<td>REG-1000</td>
<td>-0.371</td>
<td>0.504</td>
<td>-0.737</td>
<td>0.461</td>
</tr>
<tr>
<td>REG-1500</td>
<td>-0.421</td>
<td>0.504</td>
<td>-0.837</td>
<td>0.403</td>
</tr>
<tr>
<td>REG-2000</td>
<td>-0.411</td>
<td>0.504</td>
<td>-0.817</td>
<td>0.414</td>
</tr>
<tr>
<td>REG-3000</td>
<td>-0.502</td>
<td>0.504</td>
<td>-0.997</td>
<td>0.319</td>
</tr>
<tr>
<td>REG-4000</td>
<td>-0.428</td>
<td>0.504</td>
<td>-0.849</td>
<td>0.396</td>
</tr>
<tr>
<td>REG-5000</td>
<td>-0.37</td>
<td>0.504</td>
<td>-0.735</td>
<td>0.463</td>
</tr>
</tbody>
</table>

Continued on next page...
### Table 7.10 – continued from previous page...

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD-50</td>
<td>-2.279</td>
<td>0.504</td>
<td>-4.526</td>
<td>0</td>
</tr>
<tr>
<td>MD-100</td>
<td>-2.55</td>
<td>0.504</td>
<td>-5.064</td>
<td>0</td>
</tr>
<tr>
<td>MD-200</td>
<td>-2.611</td>
<td>0.504</td>
<td>-5.186</td>
<td>0</td>
</tr>
<tr>
<td>MD-300</td>
<td>-2.549</td>
<td>0.504</td>
<td>-5.063</td>
<td>0</td>
</tr>
<tr>
<td>MD-400</td>
<td>-2.498</td>
<td>0.504</td>
<td>-4.96</td>
<td>0</td>
</tr>
<tr>
<td>MD-500</td>
<td>-2.418</td>
<td>0.504</td>
<td>-4.802</td>
<td>0</td>
</tr>
<tr>
<td>MD-750</td>
<td>-2.238</td>
<td>0.504</td>
<td>-4.445</td>
<td>0</td>
</tr>
<tr>
<td>MD-1000</td>
<td>-2.04</td>
<td>0.504</td>
<td>-4.051</td>
<td>0</td>
</tr>
<tr>
<td>MD-1500</td>
<td>-1.656</td>
<td>0.504</td>
<td>-3.288</td>
<td>0.001</td>
</tr>
<tr>
<td>MD-2000</td>
<td>-1.248</td>
<td>0.504</td>
<td>-2.479</td>
<td>0.013</td>
</tr>
<tr>
<td>MD-3000</td>
<td>-0.298</td>
<td>0.504</td>
<td>-0.592</td>
<td>0.554</td>
</tr>
<tr>
<td>MD-4000</td>
<td>0.677</td>
<td>0.504</td>
<td>1.344</td>
<td>0.179</td>
</tr>
<tr>
<td>MD-5000</td>
<td>1.494</td>
<td>0.504</td>
<td>2.967</td>
<td>0.003</td>
</tr>
<tr>
<td>MN-50</td>
<td>0.25</td>
<td>0.504</td>
<td>0.496</td>
<td>0.62</td>
</tr>
<tr>
<td>MN-100</td>
<td>-0.227</td>
<td>0.504</td>
<td>-0.45</td>
<td>0.652</td>
</tr>
<tr>
<td>MN-200</td>
<td>-0.399</td>
<td>0.504</td>
<td>-0.793</td>
<td>0.428</td>
</tr>
<tr>
<td>MN-300</td>
<td>-0.36</td>
<td>0.504</td>
<td>-0.715</td>
<td>0.475</td>
</tr>
<tr>
<td>MN-400</td>
<td>-0.344</td>
<td>0.504</td>
<td>-0.683</td>
<td>0.495</td>
</tr>
<tr>
<td>MN-500</td>
<td>-0.294</td>
<td>0.504</td>
<td>-0.585</td>
<td>0.559</td>
</tr>
<tr>
<td>MN-750</td>
<td>-0.238</td>
<td>0.504</td>
<td>-0.473</td>
<td>0.636</td>
</tr>
<tr>
<td>MN-1000</td>
<td>-0.192</td>
<td>0.504</td>
<td>-0.381</td>
<td>0.703</td>
</tr>
<tr>
<td>MN-1500</td>
<td>0.0418</td>
<td>0.504</td>
<td>0.083</td>
<td>0.934</td>
</tr>
<tr>
<td>MN-2000</td>
<td>0.523</td>
<td>0.504</td>
<td>1.039</td>
<td>0.299</td>
</tr>
<tr>
<td>MN-3000</td>
<td>1.553</td>
<td>0.504</td>
<td>3.084</td>
<td>0.002</td>
</tr>
<tr>
<td>MN-4000</td>
<td>2.312</td>
<td>0.504</td>
<td>4.592</td>
<td>0</td>
</tr>
<tr>
<td>MN-5000</td>
<td>2.901</td>
<td>0.504</td>
<td>5.762</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7.10: Regression model of the MAPE (MN = mean, MD = median, REG = regression, LOW = Lowess).

The results of table 7.10 suggest that there are two main regions where MAPE was significantly low (i.e. t-statistic < -1.96):

- LEM = median, number of nearest neighbors, $k$ between 50 and 2000, with optimal $k$ around 200.
- LEM = Lowess, number of nearest neighbors, $k$ between 400 and 5000, with optimal $k$ around 1500.

**Results: RMSE**

The results of the regression model where the RMSE was the dependent variable are presented in table 7.11. The value of adjusted $r^2$ for this model was 0.052. It is not known why this was much less than the $r^2$ for the MAPE model.
### Table 7.11

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>104.554</td>
<td>4.511</td>
<td>23.176</td>
<td>0</td>
</tr>
<tr>
<td>SE-STD</td>
<td>1.477</td>
<td>2.38</td>
<td>0.62</td>
<td>0.535</td>
</tr>
<tr>
<td>MAHALANO</td>
<td>2.126</td>
<td>2.382</td>
<td>0.892</td>
<td>0.372</td>
</tr>
<tr>
<td>UNIT-MAP</td>
<td>-0.866</td>
<td>2.381</td>
<td>-0.364</td>
<td>0.716</td>
</tr>
<tr>
<td>LOW-100</td>
<td>86.096</td>
<td>14.524</td>
<td>5.928</td>
<td>0</td>
</tr>
<tr>
<td>LOW-200</td>
<td>44.216</td>
<td>8.228</td>
<td>5.374</td>
<td>0</td>
</tr>
<tr>
<td>LOW-300</td>
<td>16.547</td>
<td>7.704</td>
<td>2.148</td>
<td>0.032</td>
</tr>
<tr>
<td>LOW-400</td>
<td>2.711</td>
<td>7.589</td>
<td>0.357</td>
<td>0.721</td>
</tr>
<tr>
<td>LOW-500</td>
<td>2.368</td>
<td>7.43</td>
<td>0.319</td>
<td>0.75</td>
</tr>
<tr>
<td>LOW-750</td>
<td>-4.478</td>
<td>7.405</td>
<td>-0.605</td>
<td>0.545</td>
</tr>
<tr>
<td>LOW-1000</td>
<td>-5.369</td>
<td>7.405</td>
<td>-0.725</td>
<td>0.468</td>
</tr>
<tr>
<td>LOW-1500</td>
<td>-5.567</td>
<td>7.405</td>
<td>-0.752</td>
<td>0.452</td>
</tr>
<tr>
<td>LOW-2000</td>
<td>-5.478</td>
<td>7.405</td>
<td>-0.74</td>
<td>0.459</td>
</tr>
<tr>
<td>LOW-3000</td>
<td>-5.516</td>
<td>7.405</td>
<td>-0.745</td>
<td>0.456</td>
</tr>
<tr>
<td>LOW-4000</td>
<td>-5.405</td>
<td>7.405</td>
<td>-0.73</td>
<td>0.465</td>
</tr>
<tr>
<td>LOW-5000</td>
<td>-5.363</td>
<td>7.405</td>
<td>-0.724</td>
<td>0.469</td>
</tr>
<tr>
<td>REG-50</td>
<td>81.653</td>
<td>7.481</td>
<td>10.915</td>
<td>0</td>
</tr>
<tr>
<td>REG-100</td>
<td>39.126</td>
<td>7.405</td>
<td>5.284</td>
<td>0</td>
</tr>
<tr>
<td>REG-200</td>
<td>25.529</td>
<td>7.405</td>
<td>3.448</td>
<td>0.001</td>
</tr>
<tr>
<td>REG-300</td>
<td>11.442</td>
<td>7.405</td>
<td>1.545</td>
<td>0.122</td>
</tr>
<tr>
<td>REG-400</td>
<td>3.725</td>
<td>7.405</td>
<td>0.503</td>
<td>0.615</td>
</tr>
<tr>
<td>REG-500</td>
<td>-1.387</td>
<td>7.405</td>
<td>-0.187</td>
<td>0.851</td>
</tr>
<tr>
<td>REG-750</td>
<td>-4.332</td>
<td>7.405</td>
<td>-0.585</td>
<td>0.559</td>
</tr>
<tr>
<td>REG-1000</td>
<td>-5.57</td>
<td>7.405</td>
<td>-0.752</td>
<td>0.452</td>
</tr>
<tr>
<td>REG-1500</td>
<td>-5.873</td>
<td>7.405</td>
<td>-0.793</td>
<td>0.428</td>
</tr>
<tr>
<td>REG-2000</td>
<td>-5.798</td>
<td>7.405</td>
<td>-0.783</td>
<td>0.434</td>
</tr>
<tr>
<td>REG-3000</td>
<td>-5.87</td>
<td>7.405</td>
<td>-0.793</td>
<td>0.428</td>
</tr>
<tr>
<td>REG-4000</td>
<td>-5.687</td>
<td>7.405</td>
<td>-0.768</td>
<td>0.442</td>
</tr>
<tr>
<td>REG-5000</td>
<td>-5.583</td>
<td>7.405</td>
<td>-0.754</td>
<td>0.451</td>
</tr>
<tr>
<td>MD-50</td>
<td>-2.872</td>
<td>7.405</td>
<td>-0.388</td>
<td>0.698</td>
</tr>
<tr>
<td>MD-100</td>
<td>-2.777</td>
<td>7.405</td>
<td>-0.375</td>
<td>0.708</td>
</tr>
<tr>
<td>MD-200</td>
<td>-1.945</td>
<td>7.405</td>
<td>-0.263</td>
<td>0.793</td>
</tr>
<tr>
<td>MD-300</td>
<td>-1.031</td>
<td>7.405</td>
<td>-0.139</td>
<td>0.889</td>
</tr>
<tr>
<td>MD-400</td>
<td>0.036</td>
<td>7.405</td>
<td>0.005</td>
<td>0.996</td>
</tr>
<tr>
<td>MD-500</td>
<td>1.003</td>
<td>7.405</td>
<td>0.135</td>
<td>0.892</td>
</tr>
<tr>
<td>MD-750</td>
<td>2.808</td>
<td>7.405</td>
<td>0.379</td>
<td>0.704</td>
</tr>
<tr>
<td>MD-1000</td>
<td>4.313</td>
<td>7.405</td>
<td>0.582</td>
<td>0.56</td>
</tr>
<tr>
<td>MD-1500</td>
<td>6.505</td>
<td>7.405</td>
<td>0.878</td>
<td>0.38</td>
</tr>
<tr>
<td>MD-2000</td>
<td>8.131</td>
<td>7.405</td>
<td>1.098</td>
<td>0.272</td>
</tr>
<tr>
<td>MD-3000</td>
<td>10.254</td>
<td>7.405</td>
<td>1.385</td>
<td>0.166</td>
</tr>
<tr>
<td>MD-4000</td>
<td>11.769</td>
<td>7.405</td>
<td>1.589</td>
<td>0.112</td>
</tr>
<tr>
<td>MD-5000</td>
<td>12.996</td>
<td>7.405</td>
<td>1.755</td>
<td>0.079</td>
</tr>
<tr>
<td>MN-50</td>
<td>-3.056</td>
<td>7.405</td>
<td>-0.413</td>
<td>0.68</td>
</tr>
<tr>
<td>MN-100</td>
<td>-3.683</td>
<td>7.405</td>
<td>-0.497</td>
<td>0.619</td>
</tr>
<tr>
<td>MN-200</td>
<td>-3.317</td>
<td>7.405</td>
<td>-0.448</td>
<td>0.654</td>
</tr>
<tr>
<td>MN-300</td>
<td>-2.589</td>
<td>7.405</td>
<td>-0.35</td>
<td>0.727</td>
</tr>
</tbody>
</table>

Continued on next page...
7.10. Recommended Approach to determining optimal parameters for a k-NN ULTT model

Table 7.11 – continued from previous page...

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN-400</td>
<td>-1.976</td>
<td>7.405</td>
<td>-0.267</td>
<td>0.79</td>
</tr>
<tr>
<td>MN-500</td>
<td>-1.369</td>
<td>7.405</td>
<td>-0.185</td>
<td>0.853</td>
</tr>
<tr>
<td>MN-750</td>
<td>-0.0678</td>
<td>7.405</td>
<td>-0.009</td>
<td>0.993</td>
</tr>
<tr>
<td>MN-1000</td>
<td>1.082</td>
<td>7.405</td>
<td>0.146</td>
<td>0.884</td>
</tr>
<tr>
<td>MN-1500</td>
<td>2.975</td>
<td>7.405</td>
<td>0.402</td>
<td>0.688</td>
</tr>
<tr>
<td>MN-2000</td>
<td>4.511</td>
<td>7.405</td>
<td>0.609</td>
<td>0.542</td>
</tr>
<tr>
<td>MN-3000</td>
<td>6.863</td>
<td>7.405</td>
<td>0.927</td>
<td>0.354</td>
</tr>
<tr>
<td>MN-4000</td>
<td>8.659</td>
<td>7.405</td>
<td>1.169</td>
<td>0.242</td>
</tr>
<tr>
<td>MN-5000</td>
<td>10.054</td>
<td>7.405</td>
<td>1.358</td>
<td>0.175</td>
</tr>
</tbody>
</table>

Table 7.11: Regression model of the RMSE (MN = mean, MD = median, REG = regression, LOW = Lowess).

The results of table 7.11 suggest that there were no areas where the RMSE was statistically significantly lower, although settings which used the Regression or Lowess LEM performed well when values of k between 750 and 5000 were used. However, since the adjusted $r^2$ value is so low, it is difficult to draw conclusions from this model.

Nevertheless the results from tables 7.10 and 7.11 enhance the conclusions drawn from the full factorial grid search in tables 7.4 and 7.5 by showing that the best ten k-NN ULTT models presented in these tables are consistently and significantly better than other k-NN ULTT models, particularly when error is measured by MAPE.

7.10 Recommended Approach to determining optimal parameters for a k-NN ULTT model

This chapter has investigated what methodology can be used to identify the optimal settings of a k-NN ULTT model. The work in this section suggests the following three steps to optimise a k-NN model:

- Pre-screening of parameters as outlined in section 7.5.
- Full Factorial grid search as outlined in section 7.6 to identify the optimal combination of local estimation method (LEM) and distance metric (DM).
- Fibonacci line search to identify the optimal value of k to use for the specific LEM and DM combination obtained in the previous stage.

This methodology has been applied to the Russell Square 2003 dataset to identify the optimal architecture of k-NN ULTT model. The optimal settings are outlined below. This methodology will be reused in chapter 8 to identify the key design parameters that should be used for a k-NN ULTT model based on the Russell Square 2004 dataset and Tottenham Court Road datasets. This chapter will also explore whether the optimal k-NN architecture varies from one site to another site.
7.11 Optimal Settings

This section presents the recommended architecture of the k-NN ULTT model for use with the Russell Square 2003 dataset. It has been shown that there are different optimal values of $k$ depending on whether the MAPE, RMSE, or some other performance index is to be optimised. Thus three k-NN models are proposed. One model will be used in chapter 8 when comparing the k-NN model against other ULTT models. The second k-NN model minimises the overall MAPE, whilst the third minimises the overall RMSE.

7.11.1 Optimal ULTT model settings

The first k-NN model is designed to have good MAPE and RMSE qualities. It is also designed to perform well when compared against other ULTT models. This was achieved by firstly selecting the optimal LEM and DM using the performance index given in equation 7.5. This work identified that SE-VAR DM, and the Lowess LEM as the optimal DM-LEM combination. This ensured that the optimal model would tend to have good MAPE and RMSE characteristics. However in section 7.8 it was shown that for this given LEM and DM combination, the optimal value of $k$ will depend on whether the model is being optimised using the MAPE or RMSE as the performance index. The other ILD based ULTT models that will be presented in section 8.1 are optimised to minimise the squared error. It was thus decided to select the value of $k$ such that the RMSE would be minimised. This work was carried out in section 7.8 where it was identified that the optimal value of $k$ was 2163. Thus the optimal k-NN settings for use as a ULTT model for the Russell Square 2003 dataset are as follows:

- Distance Metric = SE-VAR
- Local Estimation Method = Lowess
- Number of nearest Neighbors = 2163

These settings will be used when the k-NN method is compared against other ULTT models in chapter 8.

To obtain a better feel for the performance of this model a graph was produced of the estimated travel times generated by this model against the actual travel times. A total of 2500 records from Russell Square 2003 dataset described in section 7.4.2 were used to generate the graph displayed in figure 7.8. From this figure it can be seen that in the main part the k-NN model seems to produce unbiased estimates. However, at very low actual travel times the k-NN model overestimates the travel time. The k-NN model never made an estimate less than 134 (the minimum actual travel time was 78). Similarly at very high actual travel times the k-NN underestimates the travel time. The k-NN model never made an estimate of travel time greater than 736 secs (the maximum actual travel time was 2533 secs). The reduced performance of the model at very low or very high actual travel times could suggest that either:
Figure 7.8: Actual versus Estimated travel times for the overall best k-NN ULTT model

- The training data set does not contain enough vehicle records with very low or very high actual travel times.
- The feature vector has not been formed correctly and should include other variables to allow better differentiation of the times of very low or very high actual travel times.
- Inherent short term vehicle-to-vehicle variability makes it impossible to predict very low travel times of individual vehicles (see section 10.3).
- The Overtaking rule has failed to remove erroneous travel time records with very low and very high actual travel times (see section 2.7.5).

### 7.11.2 Optimal MAPE settings

The previous section identified the settings that resulted in the k-NN model having good RMSE and MAPE properties. The performance index used was a function of the RMSE, MAPE, and computation time. The optimisation process was repeated, but this time, it was decided to optimise solely the MAPE, albeit taking the computation time into
account. Table 7.4 suggested that the LEM and DM should be the median and SE-VAR respectively. The unit map distance metric was discarded as settings using this were approximately three times slower than those using the median distance metric. A Fibonacci line search was undertaken to find the value of $k$ which minimised the MAPE. This was found to be 225. Thus the optimal settings to reduce MAPE are given below:

- Distance Metric = SE-VAR
- Local Estimation Method = Median
- Number of nearest Neighbors = 225

These settings will also be used when the k-NN method is compared against other ULTT models in chapter 8. Using two k-NN ULTT models which have been optimised in different ways in the work of chapter 8 will show how sensitive the k-NN model is to the methodology by which it has been optimised.

![Comparison of Estimated vs Actual Travel Times – Russell Square 2003](image)

Figure 7.9: Actual versus Estimated travel times for the model which minimised MAPE

As before a graph was plotted of actual versus estimated travel times using these k-NN model settings when used with the Russell Square 2003 data set. A total of 2500
validation records were used to produce the graph which is shown in figure 7.9. From this figure it can be seen that in the main part this k-NN model has characteristics similar to the Lowess model described above. At very low actual travel times the k-NN model overestimates the travel time. This k-NN model never made an estimate less than 141 (the minimum actual travel time was 72). Similarly at very high actual travel times this k-NN model underestimates the travel time. The k-NN model never made an estimate greater than 577 (the maximum actual travel time was 2381 secs).

7.11.3 Optimal RMSE settings

The optimisation process was repeated once again, but this time, it was decided to optimise solely the RMSE, albeit taking the computation time into account. Table 7.4 suggested that the LEM and DM should be the regression and SE-VAR respectively. The unit map distance metric was discarded as settings using this were approximately three times slower than those using the median distance metric. The SE-STD and regression combination was also dismissed since this combination was also much slower than the SE-VAR and regression combination. From the Fibonacci Line Search the optimal settings to minimise the overall RMSE were found to be:

- Distance Metric = SE-VAR
- Local Estimation Method = Regression
- Number of nearest Neighbors = 2316

As before a graph, shown in figure 7.10, was plotted of actual versus estimated travel times using these k-NN model settings when used with the Russell Square 2003 data set. From this figure it can be seen that in the main part this k-NN model has characteristics similar to the previous two models described above, overestimating travel times at very low actual travel times and underestimating the travel time at very high actual travel times. However, the operating region of this model appeared to be different to that of the model which minimised MAPE. The model which minimised RMSE gave travel time estimates between 173 and 692 seconds. However, the model which minimised MAPE gave travel time estimates between 141 and 577 seconds. This suggests that the model which minimised RMSE will perform better at very high actual travel times, but worse at very low actual travel times, than the model which minimised MAPE. The model deemed to be best overall, gave travel time estimates between 134 and 736 seconds - i.e. a better range than the other two models.

7.12 Effect of the size of the underlying database on the performance of the k-NN model

One important parameter that has yet to be considered in detail is the effect of the size of the underlying database on the performance of the k-NN model. In section 6.3.2 one of the
7.12. Effect of the size of the underlying database on the performance of the k-NN model

Figure 7.10: Actual versus Estimated travel times for the model which minimised RMSE

key assumptions for the k-NN method was to have a large dataset of historical records. However, there are many costs in obtaining a large database, including computational costs. Firstly large databases require skilled technicians to manage the data. Secondly, as stated in section 6.5 the amount of computation required by the k-NN method in finding the $k$ nearest neighbors increases as the number of records in the historical database increases. It is likely that increasing the size of the database will have diminishing returns on the increase in performance of the k-NN model. This section presents an experiment that explores and investigates whether it is possible to use a k-NN ULTT method without having to have an excessively large historical dataset.

7.12.1 Methodology and Data

An experiment was undertaken that explored the effect of the size of the database and the value of $k$ on the performance of the k-NN model. Data for this experiment came from the Russell Square 2003 dataset described in section 7.4.2.

Two experiments were carried out. In the first experiment the median LEM and SE-VAR distance metric were used. In the second experiment the Lowess LEM and SE-VAR
7.12. Effect of the size of the underlying database on the performance of the k-NN model were used. These LEM-DM combinations were chosen since the work of section 7.11 showed that these settings performed well.

In each experiment the LEM and DM were kept constant whilst the value of $k$ and the size of the training dataset was varied in a manner similar to that conducted in the full-factorial grid search.

When using a median LEM and a SE-VAR distance metric a total of 9 different values of $k$ were considered. These were: 20, 50, 100, 200, 300, 400, 500, 750, and 1000. A total of 11 different sizes of training dataset were tried. These were: 1000, 2000, 3000, 4000, 5000, 7500, 10000, 12500, 15000, 17500, 20000 records. The maximum allowable value of $k$ was the minimum size of the training dataset. For each combination of size of database and value of $k$, the performance of the k-NN model, measured using MAPE, was calculated using 250 validation records. Independence between the training data and validation data was ensured by selecting validation data from a single day and excluding all other records from this day from the training data set. Due to the random nature of selecting both the historical and validation database, the whole procedure was repeated 50 times for each combination.

When using a Lowess LEM and a SE-VAR distance metric a total of 12 different values of $k$ were considered. These were: 20, 50, 100, 200, 300, 400, 500, 750, 1000, 1500, 2000, and 3000. A total of 9 different sizes of training dataset were tried. These were: 3000, 4000, 5000, 7500, 10000, 12500, 15000, 17500, 20000 records. A large possible value of $k$ was used since results presented in section 7.11 suggest that when using a Lowess LEM larger values of $k$ should be used. The results from both experiments are presented below.

7.12.2 Results

In the first instance a median LEM and a SE-VAR distance metric were used. The results are presented in figure 7.11. Each line represents one size of historical database.

From figure 7.11 it can be seen that there is a clear relationship between the size of the database, the value of $k$, and the performance of the k-NN method. The figure suggests that as the size of the database increases, the best possible performance of the model approaches a limit asymptotically. The optimal value of $k$ is also seen to increase as the size of the underlying database increases. Finally it is seen that when very large databases are used, the performance of the k-NN model becomes less sensitive to the value of $k$ used. For example, the gradient of the curve for the line representing a database of size 15,000 is far smaller than the gradient of the curve for the line representing a database of size 1,000.

In the second experiment a Lowess LEM and a SE-VAR distance metric were used. The results are shown in figure 7.12. The results share the same basic characteristics as the above graph. That is the best possible performance is seen to increase asymptotically as the size of the database increases, and the optimal value of $k$ increases as the size of the database increases. One difference between figure 7.11 and figure 7.12 is that the Lowess
7.13 Conclusions

This chapter has studied the characteristics of the k-NN method. It firstly verified the primary assumption of the k-NN model, namely that records which are close together in feature space come from the same underlying output distribution. This experiment justified the consideration of the k-NN method for use as a ULTT model. To this authors’ best knowledge no other literature which has applied the k-NN model has explicitly tested this basic assumption of the k-NN method.

Figure 7.11: Effect of the size of the database and the value of k on the performance of the k-NN model. LEM = median, DM = SE-VAR

LEM is less sensitive to the value of $k$ chosen than the median LEM. However, the Lowess LEM performs very badly when a low value of $k$ is used.

The results from both combinations of LEM and distance metric suggest that there is little extra benefit to be gained in the performance of the k-NN ULTT model by using a training dataset over 10,000. The work in this section thus justifies the use of a training set of 15,000 records in the research in this and the next chapter. Although it could have been possible to use a larger training set in the research presented in these chapters there would have been minimal reduction in the error of the travel time estimate but increased computational costs.

7.13 Conclusions

This chapter has studied the characteristics of the k-NN method. It firstly verified the primary assumption of the k-NN model, namely that records which are close together in feature space come from the same underlying output distribution. This experiment justified the consideration of the k-NN method for use as a ULTT model. To this authors’ best knowledge no other literature which has applied the k-NN model has explicitly tested this basic assumption of the k-NN method.
The second part of this chapter developed a methodology by which the architecture of the k-NN method could be optimised for use as a ULTT model. This was achieved by following a three stage strategy to identify the optimal combination of the three key design parameters. In the first stage all levels for each design parameter which were obviously much worse than the others were excluded. This was undertaken by looking at the parameters separately from one another. In reality this approach was only possible for the distance metric since it was found that there was a very large correlation between the optimal LEM and optimal value of $k$. In the second stage a full factorial grid search was undertaken to determine the optimal settings of the LEM and DM. It was noted that there are various ways to quantify the accuracy of the ULTT model: MAPE, RMSE, or a hybrid performance index which was a function of MAPE, RMSE, and computational time. In the third stage, a Fibonacci line search was used to determine the optimal value of $k$ to a desired precision.

This methodology was used to identify the ideal settings for the key design parameters for a k-NN ULTT model using the Russell Square 2003 dataset. These are repeated below:

- Distance Metric = SE-VAR
- Local Estimation Method = Lowess
• Number of nearest neighbors = 2163

The final part of the chapter gave a caveat to the above findings. It was shown that an additional parameter in the optimisation of the k-NN method is the size of the underlying historical dataset. An experiment was undertaken which showed that the optimal number of nearest neighbors will increase as the size of the historical database increases. This experiment also justified the use of a training data-set of 15,000.

This chapter has thus presented a methodology that can be used to optimise the k-NN ULTT model. It is recognised that there may be other methods of optimising the k-NN ULTT model. In particular the use of a genetic algorithm as suggested by Bajwa et al. (2003b,a) could be an alternative approach. However a comparison of the optimisation approach developed in this chapter and that of a genetic algorithm is outside the scope of this thesis. The next chapter will compare the performance of other ULTT models with the optimal k-NN ULTT model identified in this chapter.
Chapter 8

Comparison of the k-NN method against other ULTT models

The first chapter in part B described existing ULTT models, whilst the second chapter proposed a new ULTT model based on the k-NN method. The third chapter in part B presented a methodology to optimise the k-NN method as a ULTT model. This chapter brings this work together by outlining a series of experiments that were undertaken to assess the performance of the k-NN method against other ULTT models. The first section of this chapter outlines the eight other ULTT models used in the comparison. Section 8.2 presents an experiment to identify the optimal ULTT model for use on the Russell Square link in 2003. Sections 8.3 and 8.4 repeat this experiment for the Tottenham Court Road link in 2003 and Russell Square link in 2004 respectively. These two additional experiments highlight how sensitive the performance of the ULTT models are to road topography (i.e. single or multi lane), as well as to changes over time.

Section 8.5 describes an experiment which was undertaken to quantify the effect of the ILD data cleaning method on the performance of various ULTT models. This experiment justifies the use of filtering the raw ILD data using the DSA cleaning treatment.

The chapter closes in section 8.6 by considering how the k-NN method could be used to obtain better estimates of travel time from GPS data rather than ANPR data.

8.1 Other ULTT models

One criticism of the literature on ULTT models is the failure to compare the performance of different models. The overall accuracy of the k-NN method was thus tested against various other models. For all ILD based ULTT methods the same data set and methodology was used. The ILD based ULTT methods were chosen since they were deemed to be representative of the current state of the art (Zhang & He 1998; Xie et al. 2004). In addition three ULTT models which used solely time information (i.e. 15-minute time period in day, day-of-week) and two naïve ULTT estimators were included. These are outlined below:
8.1.1 Basic Regression

The first model is a basic regression model. The form of the model is given by equation 8.1.

\[ TT_i = \beta_0 + \sum_{d=1}^{D} \left( \beta_{1,d} \cdot o_{d,i} + \beta_{2,d} \cdot \frac{o_{d,i}}{q_{d,i}} \right) + \epsilon_i \]  \hspace{1cm} (8.1)

\( TT_i \) = Link Travel Time for vehicle \( i \)
\( d \) = Detector identifier
\( D \) = total number of detectors on the link
\( o_{d,i} \) = 15-minute aggregated occupancy measured by detector \( d \) for vehicle record \( i \).
\( q_{d,i} \) = 15 minute aggregated flow measured by detector \( d \) for vehicle record \( i \).
\( \beta_0, \beta_{j,d} \) = Parameters to be estimated using OLS regression.

The first component of the equation recognises research by Sisiopiku et al. (1994) that travel time increases as occupancy increases. The second component uses the ratio of the occupancy to flow as a proxy measure of the inverse of the spot speed. The summation allows for this model to incorporate data from more than one ILD.

8.1.2 Gault and Taylor

The second model is based upon the Gault and Taylor model referred to in section 5.7. It is given by equation 8.2. As with equation 8.1 the parameters \( \beta_{j,d} \) are estimated using OLS regression.

\[ TT_i = \beta_0 + \sum_{d=1}^{D} \left( \beta_{1,d} + \beta_{2,d} \cdot o_{d,i} \cdot s_{d,i} + \beta_{3,d} \cdot s_{d,i} \right) + \epsilon_i \]  \hspace{1cm} (8.2)

\( s_{d,i} \) = 15 minute aggregated degree of saturation measured by detector \( d \) for vehicle record \( i \).

\[ s_{d,i} = \frac{q_{d,i}}{q_{d-sat}} \]  \hspace{1cm} (8.3)

\( q_{d-sat} \) = saturation flow on the road at the location of the detector \( d \).

Various modifications have been made from the original Gault and Taylor model of equation 5.44. Firstly the model has been adapted to allow for input from more than one detector. This involved ensuring only a single constant term remained in the equation. Secondly, in order to estimate the degree of saturation the saturation flow, \( q_{d-sat} \), had to be estimated. The saturation flow was estimated as the maximum flow recorded at the
8.1. Other ULTT models

detector. This is stated in equation 8.4.

\[ q_{d-sat} = \max(q_{d,i}) \] (8.4)

8.1.3 Artificial Neural Network

An artificial neural network was implemented using the in-built functions of the Matlab Artificial Neural Network toolbox. A feed-forward backpropagation network was chosen using a log-sigmoid transfer function from the input to the hidden layer and a linear transfer function from the hidden layer to the output layer. Some experimentation suggested that 12 neurons in the hidden layer, and Levenberg-Marquardt backpropagation training, gave near optimal results. When using the ANN in the experiments below, training was carried out for 120 seconds on a Pentium 4 2.8 GHz computer. Experimentation suggested that 120 seconds was sufficient for the ANN to find the optimal solution and that there was negligible if any increase in performance by allowing the ANN model more training time.

8.1.4 Naïve Mean

The naïve mean estimator was defined as the sample mean of the travel times from all records in the historical dataset.

8.1.5 Naïve Median

The naïve median estimator was defined as the sample median of the travel times from all records in the historical dataset.

8.1.6 Time: day of week

All historical records coming from the same day-of-week as the validation record were first identified. The estimator was defined as the median travel time of all these records.

8.1.7 Time: period 15

All historical records coming from the same 15-minute period of the day as the validation record were first identified. The estimator was defined as the median travel time of all these records.

8.1.8 Time: day of week & period 15

All historical records coming from the same day-of-week and the same 15-minute period of the day as the validation record were first identified. The estimator was defined as the median travel time of all these records.
8.2 Russell Square Link - 2003

This section describes an experiment that was undertaken to compare the performance of each of the ULTT models outlined in the previous section on the single-lane Russell Square Link in 2003.

8.2.1 Data and Methodology

The validation and training data-set used is described in section 4.7. The quality of both travel time and ILD measurements had been improved by screening using the overtaking rule and DSA test, resulting in an overall data-set of 24,718 records. A single day was selected at random from the 37 available days in the data set. From this data set 200 records were chosen to act as the validation data. The training data set was chosen by randomly selecting 15,000 records from the remaining 36 days of data. In this way independence between the validation and training data set was assured.

Each of the 10 ULTT methods was calibrated using this training data. Once trained, the model was then used to estimate the travel time of the 200 validation records. Due to the random nature of selecting both the validation and historical data sets, this whole process was repeated 60 times using different validation and training data sets each time. The performance of each model was then quantified by calculating the MAPE and RMSE between the estimated travel time from the ULTT model and the actual travel time (see equations 2.17 and 2.18).

All 8 ULTT models outlined above were included in the experiment alongside two k-NN models. The first k-NN model, the k-NN Lowess, was the model that was deemed to be the optimal overall ULTT model with good MAPE and RMSE properties (refer to section 7.11.1). Its settings are given below:

- Distance Metric = SE-VAR
- Local Estimation Method = Lowess
- Number of nearest Neighbors = 2163

The second k-NN model, the k-NN median, was the model that optimised solely the overall MAPE (refer to section 7.11.2). Its settings are given below:

- Distance Metric = SE-VAR
- Local Estimation Method = median
- Number of nearest Neighbors = 225

The results and analysis from the comparison are presented below.
8.2.2 Results

The results comparing the performance of the 10 ULTT models is given in table 8.1. From this table it can be seen that when the MAPE is considered, the k-NN methods clearly outperform all other methods (column 1). Although the MAPE for all models look quite high, it must be remembered that due to the short-term variability of travel times (some vehicles get stopped by the traffic lights and others do not), there is an inherent limit to the performance of any ULTT model (Robinson & Polak 2004).

<table>
<thead>
<tr>
<th>ULTT model</th>
<th>(1) MAPE, %</th>
<th>(2) RMSE, secs</th>
<th>(3) % worse than k-NN Lowess</th>
<th>(4) corr p</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>21.33</td>
<td>122.13</td>
<td>58.1</td>
<td>-0.846</td>
</tr>
<tr>
<td>Basic Regression</td>
<td>21.16</td>
<td>121.51</td>
<td>59.1</td>
<td>-0.854</td>
</tr>
<tr>
<td>Gault &amp; Taylor</td>
<td>20.84</td>
<td>121.05</td>
<td>57.6</td>
<td>-0.848</td>
</tr>
<tr>
<td>k-NN lowess</td>
<td>18.24</td>
<td>120.84</td>
<td>NaN</td>
<td>-0.817</td>
</tr>
<tr>
<td>k-NN median</td>
<td>17.83</td>
<td>123.56</td>
<td>51.6</td>
<td>-0.877</td>
</tr>
<tr>
<td>Naïve Mean</td>
<td>30.04</td>
<td>140.67</td>
<td>63.8</td>
<td>-1.000</td>
</tr>
<tr>
<td>Naïve Median</td>
<td>27.47</td>
<td>142.74</td>
<td>63.1</td>
<td>-1.000</td>
</tr>
<tr>
<td>Time: day of week</td>
<td>27.82</td>
<td>143.40</td>
<td>63.7</td>
<td>-0.989</td>
</tr>
<tr>
<td>Time: day of week &amp; period 15</td>
<td>29.77</td>
<td>168.79</td>
<td>59.7</td>
<td>-0.709</td>
</tr>
<tr>
<td>Time: period 15</td>
<td>20.44</td>
<td>135.83</td>
<td>54.9</td>
<td>-0.930</td>
</tr>
</tbody>
</table>

Table 8.1: Comparison of 10 ULTT models on the Russell Square Link - 2003 dataset.

Since the ANN and regression methods aim to minimise the total squared error, it is also useful to consider the RMSE (column 2). Using this metric, the k-NN Lowess method is still the best ULTT model. For each of the 12,000 validation records, a count was made of the times that the k-NN Lowess gave a more accurate estimate than each other ULTT models. From this the percentage of times that the k-NN Lowess gave a more accurate estimate than the other ULTT models was calculated. This is shown in column 3.

To better visualise the performance of the models a graph of the actual travel time against estimated travel time for each of the 12,000 validation records was plotted for each ULTT model. In addition a graph was plotted of the actual travel time against residual (i.e. estimated travel time - actual travel time) for each of the 12,000 validation records. Due to the high number of records the graphs show the density of records rather than the individual records (dark blue refers to a large number of records and light blue refers to a low number. White refers to no records.) The graphs for the best regression model, best time-information model, best naïve model, ANN, and two k-NN ULTT models are shown in figure 8.1. The graphs on the left hand side are the plots of actual versus estimated travel times, whereas the graphs on the right hand side are the plots of actual travel time versus residuals.

A common feature can be seen from all ULTT models. This is that all ULTT models
Figure 8.1: Performance of the various ULTT models for the Russell Square 2003 data. a. Actual vs Estimated Travel Time, b. Actual vs Residuals
8.2. Russell Square Link - 2003

Figure 8.1: Performance of the various ULTT models for the Russell Square 2003 data. a. Actual vs Estimated Travel Time, b. Actual vs Residuals
8.2. Russell Square Link - 2003

(c) top: Naive - mean travel time, bottom: 15-minute time period

Figure 8.1: Performance of the various ULTT models for the Russell Square 2003 data. a. Actual vs Estimated Travel Time, b. Actual vs Residuals
tend to overestimate the travel-time at low actual travel times, and that they tend to underestimate the travel-time at high actual travel time. This is clearly seen in the plot of the residuals. The coefficient of correlation between the residuals and actual travel time was drawn. This is shown on the plot as well as given in column 4 of table 8.1. The correlation was lowest with the k-NN Lowess ULTT model. The failure of all models to accurately estimate travel time at high actual travel times suggests that the models may be poorly specified. In particular it may be necessary to increase the concentration of ILDs on the road network to better model ULTT. For example it may be necessary to have at least one ILD between every junction since the majority of the delay on a link could occur at a single junction. Other ILDs on the link may not be affected by delay which occurs at such a junction. Figure 4.6 suggests that input from an additional 2 ILDs would be needed to ensure that there was an ILD between every junction on the Russell Square link. Other reasons for the poor performance of all ULTT models at very low and very high actual travel times have been given in section 7.11.1.

One apparently strange result in table 8.1 came from the ULTT model, ‘Time: day of week and period 15’. This does worse than both of the other time-based models, even though this ULTT model appears to be more selective than the other models. The reason why this model performs worse than the other two models is because the training data-set contains only 15,000 records. Since there are a total of 96 * 7 = 672 possible combinations of day-of-week and period-of-day, then within the 15,000 dataset there will on average be approximately 22 records from each combination. However, at night there will tend to be considerably less than this resulting in a very small sample from which to estimate the expected travel time resulting in a poor estimate. In some cases there will be no other travel time record from the same day-of-week and period-of-day combination and thus the median travel time of all the data will be used. This will also lead to a poor estimate.

It was hypothesised that due to its inherent nature of using only local data, the k-NN model may perform well at the extremes of the operation of the model - i.e. at low and high travel times. For example, the k-NN model optimised for the Russell Square 2003 dataset will use only the 2163 closest records of the 15,000 available when training the Lowess model. Other ILD based ULTT models will use all 15,000 historical records when being trained. Thus for these models the overall pattern across all data may hide any localised feature in the data. Since the k-NN model is trained using only local data, any localised feature or pattern should be more apparent and more accurately modelled. To see whether this hypothesis was correct, the 12,000 validation records were firstly sorted by travel-time and grouped into bins of 2 percentiles each (i.e. each bin contained 240 records). For each bin the MAPE and RMSE of each ULTT model was calculated. These results are shown in figure 8.2.

The top plot shows the performance of each ULTT method measured using MAPE for each percentile of travel time. This plot shows that although the k-NN Lowess method (red line) is never the best ULTT model, it performs consistently well at all percentiles of actual travel time. Other models may perform well at one extreme but not the other. For
Performance of ULTT models under different traffic conditions

- **MAPE**
- **RMSE**

**Uncongested** low actual travel time

**Congested** high actual travel time

GaultTaylor
kNN−median
kNN−lowess
ANN
NaiveMean
Time−Period

Figure 8.2: Performance of the k-NN Lowess based ULTT model against other ULTT models over the whole range of operation on the Russell Square Link - 2003 dataset.
example the Gault and Taylor method (blue line) performs well at very high actual travel
times but not very low actual travel times, whilst the k-NN median method (green line)
performs well at low travel times but not high actual travel times. This last point may
explain why in table 8.1 the k-NN median has a lower overall MAPE but a higher overall
RMSE than the k-NN Lowess model. Similar plots are seen on the bottom plot which
shows the performance of each ULTT method measured using RMSE for each percentile
of travel time.

The only model which clearly outperforms the k-NN Lowess at certain percentiles is
the naïve average method (purple line) which, by definition, performs very well around
the average actual travel time.

In keeping with the best practices suggested by Gujarati (1995) checks were made
on the significance of the parameters of the regression based models. In both the Basic
Regression and the Gault and Taylor ULTT methods, all parameters tended to be sig-
nificant (due to the multiple runnings of this experiment there were times when some
parameters were not significant). A check was also made on the collinearity between the
independent variables of the regression based models. For the Gault and Taylor method
there was little correlation between the independent variables related to the same detec-
tor. However there was often correlation (greater than 0.5) between the corresponding
variables of different detectors. For the Basic Regression method there was correlation
greater than 0.5 between the occupancy and the ratio of occupancy to flow. In addition
there was also some collinearity greater than 0.5 between the corresponding variables of
different detectors. These results suggest that multicollinearity may be a problem with the
regression based models. However Gujarati (1995, p.325) states that even in the case of
high multicollinearity, the OLS estimator retains its property as the best linear unbiased
estimator. Therefore, there is little reason to believe that the collinearity affected the
overall performance of the regression based ULTT models. However care should be taken
with the interpretation of the values of the estimated parameters.

This section has thus shown that on the Russell Square 2003 link, the k-NN method
outperforms all other ULTT models. Russell Square is a single-lane link. It is important
to verify that the k-NN method performs well on other types of link. This is undertaken
in the next section.

8.3 Tottenham Court Road Link - 2003

The previous section showed the performance of the various ULTT models on a single
lane road. This section repeats the experiment carried out in the above section but for a
multi-lane link. The multi-lane link studied was the Tottenham Court Road link described
in section 2.12. Data from 2003 was used.
8.3.1 Data and Methodology

The same methodology as that described in section 8.2.1 was used. The 2003 Tottenham Court Road dataset described in section 4.8.1 was used. This data had been filtered using the Overtaking Rule and the DSA treatment and contained 104,810 records. The optimisation procedure carried out in chapter 7 for the Russell Square 2003 data was repeated for this dataset. The optimal k-NN ULTT settings were found to be:

- Distance Metric = SE-STD
- Local Estimation Method = Lowess.
- Number of nearest Neighbors = 3231.

This result is surprising since the best distance metric was the SE-STD. As stated in section 6.4.2 this distance metric is not dimensionless. This result provides further evidence that the use of a regression based LEM reduces the sensitivity of the k-NN model to the distance metric. This model will be referred to as the ‘k-NN TCR’ model, whereas the optimal Russell Square settings outlined in section 7.11.1 will be referred to as the ‘k-NN RS’ model. The two models were included in the comparison along with the ULTT models outlined in section 8.1.

8.3.2 Results

The results showing the performance of the various ULTT models in estimating ULTT on the Tottenham Court Road link in 2003 are shown in table 8.2.

<table>
<thead>
<tr>
<th>ULTT model</th>
<th>Overall</th>
<th>Estimated vs Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) MAPE, %</td>
<td>(2) RMSE, secs</td>
</tr>
<tr>
<td>Gault &amp; Taylor</td>
<td>23.99</td>
<td>43.68</td>
</tr>
<tr>
<td>Basic Regression</td>
<td>25.73</td>
<td>49.71</td>
</tr>
<tr>
<td>k-NN TCR</td>
<td>23.72</td>
<td>42.59</td>
</tr>
<tr>
<td>k-NN RS</td>
<td>23.69</td>
<td>42.75</td>
</tr>
<tr>
<td>ANN</td>
<td>24.22</td>
<td>45.27</td>
</tr>
<tr>
<td>Naïve Mean</td>
<td>40.52</td>
<td>74.71</td>
</tr>
<tr>
<td>Naïve Median</td>
<td>34.39</td>
<td>76.97</td>
</tr>
<tr>
<td>Time: day of week</td>
<td>28.61</td>
<td>65.40</td>
</tr>
<tr>
<td>&amp; period 15</td>
<td>26.34</td>
<td>63.84</td>
</tr>
<tr>
<td>Time: period 15</td>
<td>34.87</td>
<td>75.94</td>
</tr>
<tr>
<td>Time: day of week</td>
<td>34.39</td>
<td>76.97</td>
</tr>
</tbody>
</table>

Table 8.2: Comparison of 10 ULTT models on the Tottenham Court Road Link - 2003 dataset.

The results show that the k-NN TCR model provided the most accurate estimates when measured by RMSE. This model had only a slightly worse overall MAPE than the
Figure 8.3: Performance of the k-NN TCR model against other ULTT models over the whole range of operation on the Tottenham Court Road link - 2003 dataset.
k-NN RS model. The k-NN TCR model was also seen to provide results with greater accuracy more than 50% of the time when compared against all other models (column 3). Nevertheless there was still correlation between the residuals from the k-NN estimate and the actual travel time (column 4), implying that at low actual travel times the k-NN model tended to overestimate the travel time, and at high actual travel times, the k-NN model tended to underestimate the travel time. Both k-NN based models outperformed all other models when performance was measured by both the MAPE and RMSE. These results show that the k-NN method is robust, since the optimal k-NN settings on the Russell Square single-lane link performed well on the Tottenham Court Road multi-lane link. This result suggests that it may not be necessary to optimise the k-NN settings for each individual link. Instead a single k-NN setting could be used for all links. This would facilitate the implementation of a k-NN based ULTT model for use in the commercial environment.

Figure 8.3 shows the performance of the k-NN TCR model (on the figure denoted by kNN lowess-1) against the other models (n.b. the k-NN RS model is denoted by kNN-lowess-2). It is seen that the k-NN models perform better than other ULTT models primarily at very low and very high actual travel times. It is also seen that the k-NN models perform very similar to one another at all values of actual travel time. This further suggests that the k-NN method is not particularly sensitive to the choice of the key-design parameters, and is hence easily transferable from one link to another.

8.4 Russell Square Link - 2004

The experiment was repeated on one final dataset. This was the Russell Square 2004 dataset described in section 4.8.2. The purpose of this experiment was to see how the performance of the ULTT models changed over time.

8.4.1 Data and Methodology

The same methodology as outlined in section 8.2.1 was used to measure the performance of each ULTT model. There was one important difference concerning the ILD data. Although data from three ILDs was used, data from the detector N02/052a1 was used instead of the detector N02/063e1. This was because the latter detector was faulty for the duration of the period. The k-NN model was optimised for this dataset using the approach outlined in chapter 7. The optimal settings are given below:

- Distance Metric = SE-VAR
- Local Estimation Method = Lowess.
- Number of nearest Neighbors = 2508

These settings are referred to as kNN_lowess_1. The optimal RMSE settings of the 2003 dataset (see section 8.2.1) is referred to as the kNN_lowess_2 dataset.
8.4.2 Results

The results showing the performance of the various ULTT models in estimating ULTT on the Russell Square 2004 link are shown in Table 8.3.

<table>
<thead>
<tr>
<th>ULTT model</th>
<th>(1) MAPE, %</th>
<th>(2) RMSE, secs</th>
<th>(3) % worse than k-NN Lowess 1</th>
<th>(4) corr p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gault &amp; Taylor</td>
<td>21.74</td>
<td>130.18</td>
<td>56.7</td>
<td>-0.864</td>
</tr>
<tr>
<td>Basic Regression</td>
<td>21.43</td>
<td>130.38</td>
<td>58.4</td>
<td>-0.871</td>
</tr>
<tr>
<td>kNN_lowess_1</td>
<td>19.42</td>
<td>130.23</td>
<td>NaN</td>
<td>-0.847</td>
</tr>
<tr>
<td>kNN_lowess_2</td>
<td>19.43</td>
<td>130.28</td>
<td>51.4</td>
<td>-0.846</td>
</tr>
<tr>
<td>ANN</td>
<td>22.35</td>
<td>130.69</td>
<td>58.0</td>
<td>-0.870</td>
</tr>
<tr>
<td>Naïve Mean</td>
<td>34.78</td>
<td>148.29</td>
<td>68.2</td>
<td>-1.000</td>
</tr>
<tr>
<td>Naïve Median</td>
<td>31.54</td>
<td>150.59</td>
<td>67.8</td>
<td>-1.000</td>
</tr>
<tr>
<td>Time: day of week &amp; period 15</td>
<td>24.99</td>
<td>149.13</td>
<td>56.4</td>
<td>-0.846</td>
</tr>
<tr>
<td>Time: period 15</td>
<td>22.69</td>
<td>139.82</td>
<td>56.1</td>
<td>-0.932</td>
</tr>
<tr>
<td>Time: day of week</td>
<td>31.55</td>
<td>148.26</td>
<td>65.7</td>
<td>-0.986</td>
</tr>
</tbody>
</table>

Table 8.3: Comparison of 10 ULTT models on the Russell Square Link - 2004 dataset.

From the table it can be seen that both the k-NN settings outperform all other models when measured by MAPE. However for the RMSE setting the Gault and Taylor method is found to give marginally better results. However, the Gault and Taylor method performed far poorer than the k-NN settings when measured by MAPE. Thus the k-NN methods are seen once again to be the best ULTT model.

The optimal settings for the Russell Square 2003 and 2004 data sets are very similar; the k-value is approximately 15% larger for the 2004 dataset than the 2003 dataset. This similarity explains why the optimal settings from 2003 have a performance more-or-less the same as the optimal 2004 settings. This experiment thus demonstrates the robustness of the k-NN method over time and to minor changes in the input data (i.e. the replacement of data from one detector to another detector). The k-NN method has thus been shown to be transferable between links and robust to minor changes in input data and to time.

The data used in sections 8.2, 8.3, and 8.4 all used ILD data which had been filtered using the DSA method. The next section provides the justification of having used this cleaning treatment.

8.5 Effect of the ILD data cleaning method

Sections 3.4 and 3.5 discussed several methods to ‘clean’ the raw ILD data. Having presented the ULTT models, it is now possible to quantify the effect of each of the ILD data cleaning methods on the performance of the ULTT models. This section describes an experiment that was undertaken to examine this.
8.5. Effect of the ILD data cleaning method

8.5.1 Data for experiment

The data for this experiment was based upon the 2003 Russell Square dataset described in section 4.7. The raw travel times from the ANPR system were first pre-screened using the overtaking rule. After this filtering there were a total of 40,775 records. However some of these records contained missing ILD data. After these records were removed there were a total of 32,544 records. Four additional ILD data cleaning treatments were applied to this data to filter out all records with suspect ILD data. The treatments used and the number of records left after the filtering are shown in table 8.4.

<table>
<thead>
<tr>
<th>ILD data cleaning treatment</th>
<th>No of records after filtering</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unfiltered</td>
<td>40,775</td>
<td>No ILD filtering applied</td>
</tr>
<tr>
<td>NoMissingData</td>
<td>32,544</td>
<td>All records with missing data removed.</td>
</tr>
</tbody>
</table>
| Univariate                  | 32,487                      | Records with the following attributes removed:  
  i. Any flow < 0 vehicles / hr.  
  ii. Any flow > 2000 vehicles / hr.  
  iii. Any occupancy < 0%  
  iv. Any occupancy > 95% |
| DSA                         | 24,718                      | Records not passing all 4 DSA tests removed. Criteria for passing each test outlined in section 3.4.2 given below:  
  i. Test #1 < 10  
  ii. Test #2 < 10  
  iii. Test #3 < 60  
  iv. Test #4 > 1 |
| Adjacent Detectors          | 31,416                      | Records not meeting the criteria given by equation 3.17:  
  i. \( P_{critical} = 5\% \) |
| Univariate Adjacent Detectors | 31,416                    | Univariate screening followed by the Adjacent Detectors cleaning. |

Table 8.4: Number of valid records after filtering by the various methods

8.5.2 Experiment Methodology

Each ULTT model was trained and calibrated using each of the 6 data sets outlined in table 8.4. The validation data used was the ‘NoMissingData’ data-set. A single day was chosen at random to provide the validation data. From this single day, 250 travel time records were chosen at random to act as the validation data set. To ensure independence between the training and validation data sets, all records from the same day as the validation data were then removed. In addition, to ensure that the training data-sets were of the same size, a subset of 15,000 records was chosen from each of the original data-sets at random. The calibrated ULTT model was then used to estimate the travel times of the validation data set. Comparing the estimated and actual travel times, it was possible to calculate
the Mean Absolute Percentage Error (MAPE), and the Root Mean Square Error (RMSE). Due to the random nature of selecting both the validation and historical data sets, this whole process was repeated 50 times using different validation and training sets each time.

To compare whether one ILD filter method was statistically better than another ILD filter method, a paired difference experiment was performed. To achieve this the difference between the absolute percentage error of the two filter methods for each of the 12,500 validation records was calculated. A statistical test was then undertaken to calculate the probability that the mean difference was zero. The test statistic is given by equation 8.5.

\[ t = \frac{x_D - 0}{s_D / \sqrt{n_D}} \]  

(8.5)

Where:

- \( n_D \) = Number of validation records.
- \( x_D \) = mean difference in performance of the two filter methods.
- \( s_D \) = standard deviation of the difference in performance of the two filter methods.

### 8.5.3 Results

This section presents the results of the experiment. It firstly presents results when the MAPE is considered to be the performance index.

**MAPE**

Table 8.5 shows the MAPE for each combination of ULTT model and ILD data cleaning treatment.

<table>
<thead>
<tr>
<th>ULTT model</th>
<th>NoMissing</th>
<th>Univariate</th>
<th>DSA</th>
<th>Adjacent ILDs</th>
<th>Univariate AdjILDs</th>
<th>Unfiltered</th>
</tr>
</thead>
<tbody>
<tr>
<td>BasicRegression</td>
<td>22.41</td>
<td>22.29</td>
<td>22.01</td>
<td>22.28</td>
<td>22.29</td>
<td>22.36</td>
</tr>
<tr>
<td>NaiveAverage</td>
<td>32.05</td>
<td>32.04</td>
<td>31.97</td>
<td>31.72</td>
<td>31.69</td>
<td>32.42</td>
</tr>
<tr>
<td>ANN</td>
<td>22.65</td>
<td>22.48</td>
<td>22.57</td>
<td>22.66</td>
<td>22.49</td>
<td>NaN</td>
</tr>
</tbody>
</table>

Table 8.5: MAPE for each combination of ULTT model and ILD data cleaning treatment.

Table 8.5 shows that the Adjacent ILDs treatment proposed in section 3.5 does not appear to be better than other treatments. Instead the DSA treatment is shown as providing the most accurate estimates for the Gault and Taylor, basic regression, and k-NN ULTT models. This treatment is also seen to be comparable to the best treatments for the ANN.

Further analysis was undertaken on the performance of each combination of ULTT model and ILD data cleaning treatment. Firstly, a paired difference test outlined in equation 8.5, was carried out to test whether the DSA treatment was statistically significantly
better than other treatments for the first 3 ULTT models, and whether it was significantly worse than the other treatments with the ANN model. Secondly the percentage of times out of the 12,500 validation records that the estimate from the DSA treatment was better than the other treatment was calculated (i.e. the percentage of times that the absolute percentage error using the DSA estimate was less than the absolute percentage error using the other treatment estimate). The results of this additional analysis for each ULTT model are presented in tables 8.6 to 8.9. The Naïve Average model was not considered due to its poor overall performance.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>NoMissing</th>
<th>Univariate</th>
<th>DSA</th>
<th>Adjacent ILDs</th>
<th>Univariate AdjILDs</th>
<th>Unfiltered</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2) t-statistic</td>
<td>9.7877*</td>
<td>7.595*</td>
<td>NaN</td>
<td>7.7587*</td>
<td>8.7379*</td>
<td>6.8113*</td>
</tr>
<tr>
<td>(3) % times worse than DSA</td>
<td>51.6</td>
<td>50.8</td>
<td>NaN</td>
<td>50.9</td>
<td>51.3</td>
<td>51.7</td>
</tr>
</tbody>
</table>

Table 8.6: Performance of DSA method against other methods for Gault & Taylor ULTT

<table>
<thead>
<tr>
<th>Treatment</th>
<th>NoMissing</th>
<th>Univariate</th>
<th>DSA</th>
<th>Adjacent ILDs</th>
<th>Univariate AdjILDs</th>
<th>Unfiltered</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) mean MAPE</td>
<td>22.41</td>
<td>22.29</td>
<td>22.01</td>
<td>22.28</td>
<td>22.29</td>
<td>22.36</td>
</tr>
<tr>
<td>(2) t-statistic</td>
<td>14.5576</td>
<td>9.8399*</td>
<td>NaN</td>
<td>6.0317*</td>
<td>6.0878*</td>
<td>12.7166*</td>
</tr>
<tr>
<td>(3) % times worse than DSA</td>
<td>50.0</td>
<td>48.5</td>
<td>NaN</td>
<td>45.9</td>
<td>46.4</td>
<td>49.7</td>
</tr>
</tbody>
</table>

Table 8.7: Performance of DSA method against other methods for basic regression ULTT

<table>
<thead>
<tr>
<th>Treatment</th>
<th>NoMissing</th>
<th>Univariate</th>
<th>DSA</th>
<th>Adjacent ILDs</th>
<th>Univariate AdjILDs</th>
<th>Unfiltered</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) mean MAPE</td>
<td>19.46</td>
<td>19.44</td>
<td>18.93</td>
<td>19.62</td>
<td>19.62</td>
<td>29.10</td>
</tr>
<tr>
<td>(2) t-statistic</td>
<td>14.4226</td>
<td>13.3918*</td>
<td>NaN</td>
<td>16.1432*</td>
<td>16.5025*</td>
<td>47.0673*</td>
</tr>
<tr>
<td>(3) % times worse than DSA</td>
<td>52.0</td>
<td>52.1</td>
<td>NaN</td>
<td>52.0</td>
<td>51.9</td>
<td>63.8</td>
</tr>
</tbody>
</table>

Table 8.8: Performance of DSA method against other methods for k-NN ULTT

Tables 8.6 through to 8.8 clearly show that the DSA method is significantly better than the other treatments when used in the Gault and Taylor, basic regression, and k-NN ULTT models. Although the DSA method is not the best treatment for use by the ANN ULTT model, table 8.9 suggests that it is not significantly worse than any other treatment. It should be noted that the high number of validation records, \( n_D \), in equation 8.5 makes it more likely that any difference will be significant. Nevertheless, the DSA treatment was not significantly worse than any other treatment with the ANN ULTT model, even though it had a worse MAPE reading than two other cleaning treatments.
Figure 8.4: pdf of the error (estimated - actual travel time) for all ILD treatments with the Basic Regression ULTT model
### Table 8.9: Performance of DSA method against other methods for ANN ULTT

The results presented in tables 8.6 to 8.9 show an interesting feature. Even in the cases where the DSA model is significantly better than the other treatment, this treatment does not consistently provide better estimates for all validation records. The probability that a travel time estimate for an individual validation record estimated by a ULTT model filtered using the DSA treatment is better than the estimate from the same ULTT model filtered using a different ILD treatment is only around 50%. For example table 8.7 shows that the DSA method on average gives a less accurate estimate than other treatments when used to clean data for the Basic Regression ULTT model. The table shows that the DSA treatment only provides a more accurate estimate than the Univariate treatment 48.5% of the time. Overall however, the DSA treatment still has a significantly lower error than the Univariate treatment. To determine why this was the case, it was decided to study the residuals of the error (estimated travel time - actual travel time) for both treatments. Figure 8.4 shows the probability density function (pdf) of the residuals for both the DSA and Univariate treatments when the basic-regression ULTT model is used. It shows the pdf of the residuals for the DSA and Univariate treatments are more-or-less the same. This explains why the results from the third row of table 8.7 which shows that the DSA treatment is only better than the Univariate treatment methods around 50% of the time.

Nevertheless, the DSA treatment has been shown in row (2) to provide statistically significantly better results than the Univariate treatment. This apparent paradox can be explained by looking at the pdf of the pairwise differences between the absolute residuals of the estimates from the two treatments, $\Delta_i$, given by equation 8.6. Its pdf is shown in figure 8.5 (upper).

$$
\Delta_i = |TT_{i,actual} - TT_{i,DSA-estimate}| - |TT_{i,actual} - TT_{i,univariate-estimate}|
$$

$\Delta_i =$ pairwise difference in absolute error of two treatments for validation record $i$.

$TT_{i,actual} =$ actual travel time for validation record $i$.

$TT_{DSA-estimate} =$ estimated travel time for validation record $i$ using the DSA treatment.

$TT_{i,univariate} =$ estimated travel time for validation record $i$ using the DSA treatment.

The top graph shows the relative frequency of the pairwise difference of the residuals.
8.5. Effect of the ILD data cleaning method

PDF of difference between absolute residuals of two distributions. No validation Data = 12500
Dist #1 – ULTT = BasicRegression, Filter = Univariate. Dist #2 – ULTT = BasicRegression, Filter = DSA

**Figure 8.5:** Top graph: Plot of the pdf of $\Delta_i$. Bottom graph: LHS of top diagram (red line) projected onto RHS of top diagram.
Figure 8.6: Difference between the pairwise residuals from the estimates using the DSA and Univariate data set. 12,500 validation records were used in the analysis.
between the DSA and univariate treatments. Points on the RHS indicate that the DSA is worse than the Univariate treatment, and vice-versa on the LHS. Although the mode is on the right hand side, the left hand tail is a bit fatter than the right hand tail. This is seen in the bottom graph of figure 8.5. This graph also shows the pdf of $\Delta_i$. However, in this diagram the LHS of the plot has been 'folded over' and is displayed on the RHS of the plot (in effect a look at the symmetry of the graph). For values of $\Delta_i$ where the blue line lies above the red line, the DSA treatment is worse than the Univariate treatment. It can be seen that beyond $\Delta_i = 10$ seconds, the red line lies on top of the blue line, meaning that the univariate treatment is far likelier to produce bad results than the DSA treatment.

Figure 8.6 clarifies the bottom graph of figure 8.5 by plotting the difference between the blue and red line against the value of $\Delta_i$. This figure clearly shows that when the Univariate treatment is performing better than the DSA treatment, the DSA treatment is only slightly worse. However, at points where the DSA treatment performs better than the Univariate treatment, the Univariate treatment is doing considerably worse. It is this feature that results in the paradox of the Univariate treatment providing a more accurate result a greater percentage of times than the DSA treatment, whilst at the same time the DSA treatment is seen to be significantly better than the Univariate treatment overall.

**RMSE**

Table 8.10 shows the RMSE for each combination of ULTT model and ILD data cleaning treatment.

<table>
<thead>
<tr>
<th>ULTT model</th>
<th>NoMissing</th>
<th>Univariate</th>
<th>DSA</th>
<th>Adjacent ILDs</th>
<th>Univariate AdjILDs</th>
<th>Unfiltered</th>
</tr>
</thead>
<tbody>
<tr>
<td>GaultTaylor</td>
<td>110.87</td>
<td>110.81</td>
<td>110.67</td>
<td>110.87</td>
<td>111.05</td>
<td>110.61</td>
</tr>
<tr>
<td>BasicRegression</td>
<td>112.99</td>
<td>112.96</td>
<td>112.59</td>
<td>113.33</td>
<td>113.44</td>
<td>112.70</td>
</tr>
<tr>
<td>kNN</td>
<td>110.81</td>
<td>110.77</td>
<td>110.48</td>
<td>110.87</td>
<td>110.85</td>
<td>131.50</td>
</tr>
<tr>
<td>NaiveAverage</td>
<td>129.85</td>
<td>129.85</td>
<td>129.89</td>
<td>129.83</td>
<td>129.87</td>
<td>129.83</td>
</tr>
<tr>
<td>ANN</td>
<td>112.40</td>
<td>111.70</td>
<td>113.20</td>
<td>112.08</td>
<td>111.97</td>
<td>308.82</td>
</tr>
</tbody>
</table>

Table 8.10: RMSE for each combination of ULTT model and ILD filtering method.

Similarly as with the MAPE results, the DSA is seen to be the best filter method for the Gault and Taylor, Basic Regression, and k-NN methods. It is surprising that the unfiltered dataset is seen to perform best for the first of these ULTT models. As with the MAPE results, the Adjacent Detector Test treatment, proposed in section 3.5, performs poorly.

One other observation to make from tables 8.5 and 8.10 is that the Gault and Taylor and Basic Regression ULTT models are not particular sensitive to the use of unfiltered data. On the other hand the k-NN and ANN models are seen to deteriorate markedly when an unfiltered data set is used.
8.6 Potential application of the k-NN method to GPS data

The purpose of using the ANPR travel time data set was to allow the k-NN ULTT model to be compared and evaluated against other ULTT models and the true value of travel time. In effect the k-NN method allowed travel time records from different times but the same underlying travel time distribution to be aggregated together to allow for an accurate estimate of the ULTT given the current flows and occupancies measured on the link. It could be possible to use travel time records from GPS probe vehicles rather than travel time records from ANPR systems. This would help overcome one of the limitations of measuring ULTT using data from GPS probe vehicles, namely an insufficient number of probe vehicles. Travel time records from GPS probe vehicles would not be used explicitly to estimate the travel time in the period but would be added, along with the appropriate flow and occupancy readings, to the historical database. Thus if it were desired to estimate ULTT the only input to the k-NN model would need to be the current flow and occupancy measurements at the time of interest. Unlike the model proposed by Xie et al. (2004), probe vehicle records from the current period would not be used directly to estimate the ULTT of the current period. The research in this chapter suggests that this approach would be successful. Certainly the results in tables 8.1, 8.2, and 8.3 suggest that using travel time records with similar ILD data rather than from a similar day-of-week or time-of-day would result in more accurate estimates of ULTT.

8.7 Conclusions

This chapter has reported the results of a series of experiments undertaken to assess the performance of the k-NN method compared to other ULTT models. It has been shown that the k-NN method outperforms all other ULTT models on both single-lane and multi-lane roads. It was also shown that the k-NN method is not particularly sensitive to its settings, and it may be possible to have a single k-NN architecture for all links. Section 8.5 then investigated the effect of the ILD data cleaning treatment on the performance of ULTT models. It was shown that the DSA treatment is the optimum treatment to use for most ULTT models including the k-NN method. This section also showed that the k-NN method is sensitive to the inclusion of unfiltered ILD data. It is thus imperative that if the k-NN method is used, the historical database must firstly be filtered for records with erroneous ILD data. Nevertheless, the k-NN method has been shown to be superior in terms of accuracy to other ULTT methods. Part C of this thesis extends this work by considering not only whether the k-NN is the best ULTT model, but whether the k-NN method can be used to characterise the variability of travel times on a link.
Part C - Application to the study of TTV

Part B of this thesis has developed an urban link travel time model, based on the k-NN method, that uses data from inductive loop detectors to estimate the travel time on a link. Part C furthers this work, researching whether the k-NN ULTT model can also be used to study travel time variability (TTV). Chapter 9 provides a comprehensive introduction to the topic of TTV. It identifies how TTV can be disaggregated and quantified. It provides a review of the empirical research into TTV. Chapter 10 examines how effective the k-NN method is at modelling TTV. In particular the ability of the k-NN method to model day-to-day and vehicle-to-vehicle TTV is studied. Chapter 11 presents the conclusions of the research undertaken in this thesis. It also presents recommendations for future work.
Chapter 9

Travel Time Variability: Literature Review

The concept of Travel Time Variability (TTV) has been recognised within the literature for a long time. One of the earliest considerations of TTV was by Wardrop (1952). He noted that journey times often have very skew distributions, such that there tends to be a minimum journey time, but it is possible to have very long journey times. Thomson (1968) is one of the earliest papers to consider TTV (or predictability as he terms it) as an important characteristic of the road network. He notes an important feature about most journeys:

‘Thus on the majority of occasions he [the traveller] will arrive early and his time will often be wasted just as much as if he were sitting in a traffic jam. There can be no doubt that in London the unpredictability of journey times is an important source of time losses. In addition it plays such havoc with bus schedules that it is one of the greatest problems affecting the management of London bus services.’

More recently the Department for Transport in the UK has realised the importance of TTV, and has commissioned various research (Frith 2000; Mott MacDonald 2000a; Sutch 2001; Porter 2001). One output from this research is the inclusion of TTV as a parameter in cost-benefit analysis software of new road schemes (Mott MacDonald 2000a).

This chapter reviews the travel time variability (TTV) literature. This provides the background to the proposed research in TTV undertaken using the k-NN model in the next chapter. The chapter begins by asking the question ‘why does travel time vary?’ It identifies various travel time drivers which when changed, affect the travel time. It is the changing of these drivers that cause travel time variability. Sections 9.2 and 9.3 introduce concepts and metrics to allow TTV to be characterised. Section 9.2 describes various ways by which it is possible to disaggregate TTV into its component parts, whilst section 9.3 introduces ways of quantifying TTV. Section 9.4 then provides an overview of much of the empirical work that has been undertaken in modelling TTV. Section 9.5 then briefly introduces other important concepts in TTV which are not used directly within this thesis but should be mentioned in any review of TTV. The chapter concludes in section 9.6 by
9.1 What causes travel time to vary

9.1.1 Drivers of Travel Time

Before attempting to model TTV and even before giving a full literature review, it is helpful to review why travel time varies. The answer to this question has already been studied in part B. In section 5.2 it was stated that travel time can be disaggregated into three component parts:

- Turning movement intersection delay time
- Constant-speed cruise time
- Mid-link delay time

Consider the first of these; turning movement intersection delay time. The queuing theory outlined in appendix A.2 provides much insight into what causes this to vary. It suggests that as the flow (i.e. demand) on a link increases, so the travel time increases. It also suggests that as saturation flow decreases (in effect, as the capacity of the link decreases) so the travel time also increases.

Equation 5.36 shows that the constant-speed cruise time is affected by the choice of free-flow speed taken by the driver. Queuing theory also suggests that the capacity flow of a non-signalised junction is related to how much risk a driver is prepared to take in order to undertake a turning manoeuvre (TRB 1975). In general another factor in influencing travel times is the vehicle and driver behaviour of the individual vehicles on the road.

Figures 2.8 and 2.9 illustrate that even in the short term travel time can vary significantly due to the presence of traffic lights and zebra crossings. Such control objects can increase the travel time of some vehicles but not others.

The above examples have all suggested that travel time, and hence travel time variability, depends on a few key generic drivers. These travel time factors can thus be categorised into four classes. They are listed below:

- Demand - increasing demand increases travel time.
- Capacity - increasing capacity decreases travel time.
- Vehicle Performance - can increase or decrease travel time (this class also includes driver behaviour).
- Control - can increase or decrease travel time.

Many variables such as flow, or traffic signal settings, have an effect on link travel time. Such variables can be called travel time factors and can be classified under one or more
of the travel time drivers listed above. Each driver is explained in greater detail below, and several examples of travel time factors falling under this driver are given. It is the variation of these travel time factors which cause TTV.

9.1.2 Demand

As the demand for travel on a network increases so the travel time through that network tends also to increase monotonically. Table 9.1 presents a list of several travel time factors that affect demand on a link. This list is certainly not exhaustive, but gives a flavour of the factors at work.

<table>
<thead>
<tr>
<th>Travel Time Factor</th>
<th>Description.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commuting</td>
<td>Increases demand on the network at certain times of the day.</td>
</tr>
<tr>
<td>Sports Event</td>
<td>Increases demand on the network in the proximity of the stadium.</td>
</tr>
<tr>
<td>School Term</td>
<td>Increases the demand on the network at the start and end of the school day as parents shuttle their kids to school on the 'school run.'</td>
</tr>
<tr>
<td>Adverse Weather</td>
<td>Adverse weather may encourage people to take private transport as opposed to walking or cycling, thus increasing demand on the road network. However, adverse weather may also dissuade people from undertaking certain activities such as going to the beach, thus reducing demand on the road link that leads to the beach.</td>
</tr>
<tr>
<td>Activities</td>
<td>An individual may have an activity that they need to attend. They will have to use the transport network to get from their present location to their desired location.</td>
</tr>
<tr>
<td>Economy</td>
<td>As a country gets richer, its populace tends to obtain a higher disposable income, which can be used for travelling.</td>
</tr>
<tr>
<td>Availability of vehicle storage facilities (e.g., parking spaces)</td>
<td>If there is a lack of parking spaces at a popular location, then the traveller will be forced to drive around the locality seeking such a space. This will increase demand on the road network.</td>
</tr>
</tbody>
</table>

Table 9.1: Travel Time Factors affecting the demand driver

9.1.3 Capacity

A valid question may be asked as to why increasing demand on the network should tend to increase travel time through that network. The reason is because critical points in the network can only service traffic at a certain rate. These critical points tend to be at the nodes of the network, be they road junctions, train stations, or landing slots at an airport. However the critical point need not be at a node - a steep hill on a motorway causing a sudden reduction in vehicle speed may cause the flow to break down and traffic jams to appear. Thus each facility on a network will have a certain capacity - the amount of traffic it can handle in a certain time period.
9.1. What causes travel time to vary

Once demand exceeds capacity, queues will occur, and the traveller will have to wait to be ‘serviced’ by the network facility. In general there is a monotonic relationship between increasing capacity of a network facility (i.e. a link or node) and reducing the travel time through that facility. This relationship may not necessarily be linear in nature.

Much of the literature assumes that capacity is constant. Indeed many calculations of network performance use the ratio of volume to capacity, assuming the capacity of the facility to be constant. This however may not be the case, even in the short term. Table 9.2 presents a list of travel time factors that cause capacity to change. This list is certainly not exhaustive, but gives a flavour of the factors at work. As capacity of the facility is reduced, the travel time through that facility will tend to increase.

<table>
<thead>
<tr>
<th>Travel Time Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidents</td>
<td>An accident may partially or fully block a transport link. The capacity of this link will thus be reduced or even nullified altogether.</td>
</tr>
<tr>
<td>Parked Vehicles</td>
<td>Parked vehicles may partially block the road or inhibit visibility, forcing drivers to drive more slowly, resulting in a reduction of the capacity of the link.</td>
</tr>
<tr>
<td>Poor Maintenance</td>
<td>A train, plane, or bus may break down and have to be taken out of service. There may not be a replacement vehicle, and thus the capacity of the scheduled-service will be reduced.</td>
</tr>
<tr>
<td>Adverse Weather</td>
<td>Adverse weather may reduce visibility, forcing vehicles to drive slower in order that they can drive within their 'comfort zone' or safety limit. The reduction in travel speed will result in a reduction of capacity.</td>
</tr>
<tr>
<td>Broken Signals</td>
<td>Broken signals on a transport network will force the closure of a node or a reduction of capacity through that node.</td>
</tr>
<tr>
<td>Flows in other lanes</td>
<td>The flow on a main road may affect the capacity of a minor road when there is a non-signalised junction connecting the two roads. As the flow on the main road increases or becomes less platooned so the amount of gaps that could be used by vehicles in the minor road to enter the main road reduces. Thus the capacity of the minor road will be reduced.</td>
</tr>
</tbody>
</table>

Table 9.2: Travel Time Factors affecting the capacity driver

9.1.4 Vehicle Performance

The third important generic factor that dictates what the travel time on a particular journey will be is the performance of the vehicle and behaviour of the driver. For example some vehicles will only be able to drive up a steep hill at very low speeds due to the engine having insufficient power.

Vehicle performance should also include the characteristics of the driver of the vehicle, and in particular the desired travelling speed of a driver. As an example, on a single lane road where overtaking is prohibited, the speed of any vehicle will be limited to the speed...
9.1. What causes travel time to vary 240 of the leading vehicle in the platoon. This speed has been chosen by the driver of the leading vehicle in the platoon. Thus some of the short-term travel time variability can be explained by heterogeneous speeds adopted by individual drivers. As another example, consider a vehicle in a minor road turning into a major road. There will be gaps in the traffic-stream on the major road that will allow a vehicle from the minor road into the major road. Different drivers have different gap-acceptance criteria - some drivers will even force the mainstream traffic to slow abruptly to allow them in. Other drivers will reject smaller gaps in the traffic-stream of the major road, and wait for a sizable gap to appear before they are confident enough to make the turning manoeuvre.

Table 9.3 presents a list of travel time factors that cause vehicle-performance to vary. This list is certainly not exhaustive, but gives a flavour of the factors at work. Several of the factors listed have already appeared in the list of factors affecting capacity and demand.

<table>
<thead>
<tr>
<th>Travel Time Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle Speed</td>
<td>The vehicle has a maximum speed at which it can comfortably travel.</td>
</tr>
<tr>
<td>Driver Speed</td>
<td>The driver or pilot of the vehicle may not necessarily travel at the maximum speed of the vehicle, but will choose the speed at which the vehicle travels.</td>
</tr>
<tr>
<td>Parked Vehicles</td>
<td>Parked vehicles may partially block the road or inhibit visibility, forcing drivers to drive more slowly to stay within their ‘comfort zone.’</td>
</tr>
<tr>
<td>Adverse Weather</td>
<td>Adverse weather may reduce visibility, forcing vehicles to drive slower in order that they can drive within their ‘comfort zone’ or safety limits.</td>
</tr>
<tr>
<td>Poor Maintenance</td>
<td>A train, plane, or bus may break down and have to be taken out of service. There may not be a replacement vehicle, and thus the capacity of the scheduled-service will be reduced.</td>
</tr>
<tr>
<td>Risk Acceptance</td>
<td>Some drivers will accept more risk than others - for example, in the ‘gap’ that they will accept before manoeuvring from a minor road to a major road. Riskier drivers will tend to increase the capacity of a junction.</td>
</tr>
<tr>
<td>Human factors</td>
<td>Humans are often unpredictable, and their behaviour varies between humans, and over time within humans. The mood of a driver may affect the way they drive. A driver in a ‘bad-mood’ may drive more aggressively and accept more risk. A tired driver is likely to drive more slowly and react slower to events around them.</td>
</tr>
</tbody>
</table>

Table 9.3: Travel Time Factors affecting the vehicle behaviour driver

9.1.5 Control

The fourth and final important generic factor that dictates what the travel time on a particular journey will be is ‘control’ - that is of direct management of the traffic network.
The most obvious example of a control factor is traffic lights. Traffic lights effectively dictate the capacity of a link; if more green time is provided, then more vehicles can flow out of the link, and hence more vehicles can use the link without causing traffic queues. Control provides the ability to manage the transport network, and directly change the capacity of a link, or even the demand into a location.

A good example of control being used to control demand is the use of the Electronic Road Pricing system (ERP) in Singapore. (Goh 2002). During rush hour the cost of using roads is increased, whilst in off-peak hours, it is free to use these same roads. In this way, demand is pushed from peak hours to off-peak hours. By reducing the demand at peak hours the travel time in these hours is reduced. Adding this demand to off-peak hours has little effect on travel time since at these times there is excess capacity.

The table below gives a comprehensive list of ‘control’ travel time factors that can affect the travel time of a journey.

<table>
<thead>
<tr>
<th>Travel Time Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed Limit</td>
<td>The travel time on a long link will be dictated by the maximum legal speed limit that a driver or pilot is allowed to travel at on that link. The majority of drivers will use this speed limit to dictate what speed they drive at.</td>
</tr>
<tr>
<td>Railway Crossing</td>
<td>Trains are given priority over vehicles. A vehicle may have to wait several minutes for the passage of a train or several trains across a road.</td>
</tr>
<tr>
<td>Traffic Lights</td>
<td>The most common form of control seen on road networks, these control the amount of vehicles that can flow out of a link, and thus control the capacity of these links.</td>
</tr>
<tr>
<td>Vehicle restrictions</td>
<td>Certain vehicles may be banned from using certain routes at certain times. More travel time will be needed to use alternative routes.</td>
</tr>
<tr>
<td>Tolls</td>
<td>A charge can be made for the use of a transport facility. This charge can be varied to control demand. A good example of this is the Electronic Road Pricing system (ERP) in Singapore, or the increase in airplane tickets at festivals. Increasing the cost of a facility helps to reduce demand for that transport facility.</td>
</tr>
<tr>
<td>Toll booths on a toll-paying highway</td>
<td>Each tollbooth can only service a certain amount of vehicles per hour. The capacity of the toll-road can be controlled by limiting the amount of tollbooths that let vehicles onto the toll-road.</td>
</tr>
</tbody>
</table>

Table 9.4: Travel Time Factors affecting the control driver

9.1.6 Interrelationship between the 4 generic drivers

The 4 drivers of travel time variability are not wholly independent of one another. Examples of the interrelationship of the drivers are given below:
• **Capacity and demand** - There may be relationships between the capacity of a network facility and the demand of a network facility. A good example of such a relationship is vehicle parking spaces. If there is not enough capacity to store vehicles at a car-park, then these excess cars will be forced to drive in the locality searching for car spaces. These extra vehicles are of course adding extra demand and congestion on to the road network.

• **Capacity and Control** - Control can influence capacity. As has already been mentioned, Urban Traffic Control (UTC) systems such as SCOOT and SCATS which manage traffic lights, can control the capacity of a link by increasing or decreasing the split time given to it.

• **Demand and Control** - Demand can influence control devices. A good example of this are Urban Traffic Control systems which will detect high demand on roads approaching a signalised junction, and increase the proportion of green time given to these high roads in high demand. Indeed, there is actually a three-way relationship between demand, capacity, and control. Increased demand will lead to more green time being given to a link, which leads to increased capacity on that link, attracting more demand! In reverse, control devices can influence demand. An example of this, Electronic Road Pricing has already been mentioned.

• **Vehicle performance and capacity** - Vehicle performance can also affect capacity. For example, a highly risk-averse driver turning from a minor road to a major road may only accept a large headway before making the turning movement. This driver may reject other possible headways that other more risk-accepting drivers would take. Thus risk-averse drivers will reduce the capacity of the minor road.

### 9.2 Disaggregation of travel time variability

The previous section has introduced the concept of travel time variability and provided reasons to explain why the travel time may vary. Four key drivers of TTV were identified. There are other ways in which TTV can be disaggregated. This section reviews the literature on breaking-down TTV into various components.

#### 9.2.1 Disaggregation by travel time driver

One way of disaggregating travel time variability has already been proposed in the previous section, i.e. by the travel time driver. Some recent research has recognised this. For example, Ove Arup And Partners International Ltd (2002) undertook work in the related field of journey time variability (i.e. path time variability - refer to section 2.14) and disaggregated TTV (they call it journey time variability) into the following three components:

• Demand related effects.
• Capacity related effects.
• Incident related effects.

To this author’s best knowledge the literature has no empirical study on the explicit
effect of demand on TTV. Certainly flow (a proxy indicator of demand) is a very important
factor. From the work on queuing theory it can be seen that delay time rises exponentially
as the ratio of the arrival flow to the capacity flow increases (see equation A.14). Thus
if the flow is already near the capacity flow, a minor incident that temporally causes a
reduction in capacity will have a large impact on the delay. However, at lower flows this
same capacity reducing incident would cause only a negligible increase in delay.

9.2.2 Disaggregation by predictability

Wong & Sussman (1973) hypothesised that TTV can be disaggregated depending on how
predictable the source of variability was. They identified three levels of predictability:

• **Regular condition-dependent variations**: these are predictable variations re-
sulting from differences between rush-hour and slack-hour traffic, differences between
summer and winter conditions, days of the week etc.

• **Irregular condition-dependent variations**: variations are also introduced by
irregular (and hence unpredictable) changes in network conditions. Examples here
would be a sudden rain storm, an accident, a fire, etc.

• **Random variations**: Two vehicles starting at the same moment from a given point
to another on the same route may in all likelihood take different times to complete
the trip ... due to hitting a traffic light at different points in its cycle... a dog darting
out from the sidewalk...

Wong & Sussman (1973) note the arbitrariness of the distinction between irregular
condition-dependent variations and random variations. The distinction between the two
is that the latter variation is only very short term and will not affect all vehicles on the
same route in the same time period. Nevertheless, the difficulty in assigning variability to
the correct category makes this model of TTV disaggregation difficult to use.

9.2.3 Disaggregation by location

TTV can be disaggregated by where on the road network it occurs. For example, Dale
et al. (1996) disaggregated TTV into variability on the link and variability at a junction.
Disaggregating it in such a fashion is the basis of the link-junction model. Such a model
allows the variance of travel time along a path to be calculated by summing the individual
variances on all the links and junctions that constitute that path. This assumes there is
no correlation of travel-time between links. Their analysis of data from major arterials
in London showed that on average there was a correlation coefficient of 0.2 between the
9.2. Disaggregation of travel time variability

travel times on adjacent links. Issues concerning the link-junction model have already been highlighted in section 2.14.

9.2.4 Disaggregation by temporal scale

Another way to disaggregate TTV is by temporal scale of the variability. In particular the time periods outlined in table 9.5 are often considered explicitly (e.g. Montgomery & May 1987, Park et al. 1999, Grant-Muller et al. 2001). These will be mathematically defined in section 10.1.1.

<table>
<thead>
<tr>
<th>Temporal Scale</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle-to-vehicle (v2v)</td>
<td>Variability in travel-time between vehicles that travel over the same link in the same short period of time, say 5 to 15 minutes. This variability could thus be described as the short-term travel time variability. This variability is the ‘random variability’ referred to by Wong &amp; Sussman (1973) in section 9.2.2. The reader should not confuse v2v TTV with vehicle travel time variability, which is the variability in travel times from the same vehicle.</td>
</tr>
<tr>
<td>Period-to-period (Within-day) (p2p)</td>
<td>Variability in travel-time that occurs over the course of a single day. Within this thesis a single day is split up into 96 periods, each of duration 15 minutes.</td>
</tr>
<tr>
<td>Day-to-day (d2d)</td>
<td>Variability in travel-time that occurs between the same period of different days. For example there is expected to be variation in the travel time patterns between a Monday and a Sunday.</td>
</tr>
<tr>
<td>Seasonal (i.e. between months) (s2s)</td>
<td>Variability in travel-time in the long run. For example, in summer demand for travel to a beach, and hence the travel time, will be greater than in the winter.</td>
</tr>
</tbody>
</table>

Table 9.5: Disaggregation of TTV by time period. Bold terms in brackets are the nomenclature used in the rest of this thesis.

Each component of TTV can be illustrated on a graph. Figure 9.1 illustrates vehicle-to-vehicle (v2v) and period-to-period (p2p) TTV. It displays a scatterplot of all vehicle travel time records plotted by time of day. The solid red line represents the mean travel time at each 15-minute period of time. It can be seen that at any one time of day there is variation in the travel time of each individual vehicle. For example, at around 16:45 travel time is seen to vary between 222 and 412 seconds. This variation in travel time is termed vehicle-to-vehicle (v2v) TTV. The red line showing the mean travel time at each 15-minute period of time is also seen to vary. For example between around 5:30 and 10:00 the mean travel time varies from just over 100 seconds to over 400 seconds. This component of variability is called period-to-period (p2p) variability.

Day-to-day variability is illustrated in appendix figures C to C.4. For example it is seen that there is a big difference in travel time at 9am on a typical Sunday compared to 9am on a typical Monday. Within this research only v2v, p2p, and d2d TTV will be
9.2. Disaggregation of travel time variability

Figure 9.1: Scatterplot of individual vehicle travel times by time of day. The v2v TTV accounts for the ‘width’ of the plot.
9.3. Quantifying TTV

The previous sections have outlined various components of TTV but have not proposed a way to quantify these. This section briefly outlines metrics which have been proposed and used to measure TTV. The most common metric to measure variability is the standard deviation, $\sigma$. There are derivations of this metric, one such measure being the reliability measure, $R$, proposed by Polus (1979), which is the inverse of standard deviation, and is given in equation 9.1.

$$ R = \frac{1}{\sigma} \quad (9.1) $$

One problem with using the standard deviation is that it is not normalised and thus not dimensionless. Thus the most widely used metric in the literature to quantify TTV, is the coefficient of variation, $CV$. This is the ratio of the standard deviation of travel time to the mean of travel time, $\mu$, and is given by equation 9.2.

$$ CV = \frac{\sigma}{\mu} \quad (9.2) $$

One criticism of the use of the Coefficient of Variation is that it gives little information...
about the underlying distribution. In particular it is quite likely that journeys which are quicker than usual cause less frustration than journeys which are longer than usual. Ove Arup And Partners International Ltd (2002) thus advocate the use of values indicating the expected earliness and lateness. The advantage of using such measurement is that they are directly applicable to many scheduling models which calculate the cost of TTV (see section 9.5.2).

Dandy & McBean (1984) suggest that a more representative travel time is one that is exceeded only once in every 20 trips - i.e. the 95th percentile travel time. Lomax et al. (2001) use this in the definition of the Buffer Index, which can be used as a measure of TTV.

\[
BI = W \left[ \frac{TR95 - ATR}{ATR} \right] \cdot 100
\]

\(BI = \) Buffer Index
\(ATR = \) Average travel rate (in minutes per mile)
\(TR95 = \) 95th percentile travel rate (in minutes per mile)
\(W = \) Appropriate weighting

This section has suggested that there is no one single metric which suitably characterises TTV, and that several such metrics should be used. This is considered further in chapter 10.

9.4 Empirical research into TTV on road networks

This section reviews empirical research that has been undertaken on TTV on road networks for private passenger vehicles. The first part considers the research that has been undertaken when TTV is disaggregated by temporal scale. The majority of research falls in this category. The second part then considers other empirical research on private vehicle TTV on road networks.

9.4.1 Temporal Scale TTV

This section presents a review of the research that has been undertaken on temporal scale TTV. It will be shown that the majority of research suffers from lack of data, which often results in the research failing to properly control for each of the various temporal scales of TTV (i.e. varying only one, but keeping all others constant).

Smeed & Jeffcoate (1971) undertook a study of the day-to-day variability of journey times over a particular route from home in Bray to central London. Data was collected in two different periods. One reading was taken for the journey into work and one on the way from work. The dataset was relatively small. The first period from June 1969 to May 1970 contained 124 morning and 75 evening commutes. The second period, from October 1970 to February 1971, contained only 47 and 41 records for the morning and evening commute.
They attempted to observe the variation of journey times with starting times (i.e. p2p TTV) and the variation between days of journey times which started at particular clock times (i.e. d2d TTV). Due to the lack of data they were unable to properly disaggregate the effect of p2p TTV on their study of d2d variability and vice-versa. Due to the nature of their data collection they were also unable to consider and control for the v2v variability. One conclusion they did draw was that travel times on a Monday were significantly higher than on other days. They also suggested that the distribution of travel times was normal. However, since a normal distribution implies a small probability of having a negative travel time, it is theoretically more pleasing to use another distribution such as the log-normal or Gamma. However, these two distributions were not considered.

Mogridge & Fry (1984) performed a similar study to Smeed & Jeffcoate (1971), measuring TTV over a route from south London to central London. Once again the sample size was small (only 94 morning and 78 evening journeys), and no proper attempt was made to control between the effects of v2v, p2p and d2d variability. They concluded that the coefficient of variation was about 0.20 in the morning peak and 0.15 in the evening peak. They also suggested that the travel times were distributed log-normally.

Herman & Lam (1974) studied the travel time on 25 routes into a common workplace in Detroit. Although only just above 100 data points were collected for each route, they at least recognised the limitations caused by this lack of data. They found that the coefficient of variation of travel time decreased with increasing mean travel time, and stabilised at around 6% for trips to work and 10% for trips from work. This result that there tends to be more TTV in the evening peak contradicts one of the conclusions of Smeed & Jeffcoate (1971) who found that there was less variability in the evening peak. This suggests that the characteristics of TTV differ from link to link. Herman & Lam (1974) also found that travel times on a certain route were more or less uncorrelated from one section to another. However such correlations were only investigated for one individual route. One final conclusion was that the standard deviation of travel times was seen to increase as the mean travel time increased. However, they could not fit the data adequately to a model. At high average travel times a linear relationship between mean travel time and standard deviation held, whilst at lower mean travel times a power formula was seen to provide the best fit - indeed they suggested that the standard deviation was proportional to the root of the mean travel time. Research was also undertaken into the distribution of the travel time. They found that the normal distribution was only a good fit for the lowest 60% of all travel times. There was a ‘fatter tail’ above this point. This suggests that a log-normal or gamma distribution would give a better fit.

There are several other examples in the literature of small travel time surveys. Richardson & Taylor (1978) performed research on trips taken in the morning and evening peak using six floating car vehicles. They concluded that travel times were distributed log-normally and not normally. Polus (1979) undertook similar research to that above using data from 14 arterial routes into Chicago. He found that link travel time fitted a gamma distribution well. However, the sample size used was smaller than in other research, and
only one route had more than 30 travel time records. Hendrickson & Plank (1984) conducted a small study of TTV on 4 links in Pittsburgh. They found that the average coefficient of variation was 13%. However, less than 8 travel times were recorded for each link diminishing the confidence in the results. Dandy & McBean (1984) undertook research into TTV for various modes of transport in Cambridge, Massachusetts, USA. They recorded a total of 87 travel times for a morning commute and 78 travel times for the evening commute. To determine what distribution best fitted the travel time data they first performed a Kolmogorov-Smirnov test. Due to the low sample size there was no significant difference between the observed and fitted distribution for the normal, log-normal, and gamma distributions. They then undertook a more stringent test of normality, testing whether the sample skewness was significantly different from 0. For three of the 8 subsets of data, the sample skewness was seen to be significantly different from 0 (at the 5% level) indicating that a log-normal fit was better than a normal fit. In research conducted in Leeds, May & Montgomery (1983) also reported a log-normal fit of short term travel time data.

More recently Black et al. (2004) describe work undertaken to measure TTV of urban journeys in Leeds and north London. They created a model to relate the coefficient of variation (CV) of link travel times that started at the same time of day, with the mean congestion index (CI), measured at that same time of day, and the journey length, d. The congestion index was defined as the link travel time divided by the free-flow travel times, and is given in equation 9.4. They concluded that equation 9.5 could model the relationship between these variables.

\[
CI = \frac{\text{travel time in current 15-minute period}}{\text{free flow travel time}} \quad (9.4)
\]

\[
CV = 0.148 \cdot CI^{0.781} \cdot d^{0.285} \quad (9.5)
\]

Their research was hampered by the use of very small datasets. Their London data sets comprised of only 8 different readings for each time period, and the Leeds dataset only 10 different readings for the same time period. Since only one vehicle was used, they also did not control for the v2v TTV.

Much of the above research has been limited by lack of data. However, technologies such as ANPR have allowed a vast volume of travel time data to be collected. Thus in recent years there has been renewed interest in the research of travel time variability in the academic community, and by government. For example, Sutch (2001) undertook research on day-to-day variability using data collected from ANPR cameras on the M6 motorway in the UK. Sutch (2001) attempted to model the behaviour of day-to-day variability. Using visual analysis of diagrams Sutch (2001) firstly determined that weekdays can be split into 3 classes: Monday, Tuesday to Thursday, and Friday. He then compared the mean travel time at a certain time of day for several days from the same day-group. He found that there was more day-to-day variability in the periods after the period with the highest
mean travel time than in those before. In other words there is less variability in the build up to the peak, than in the recovery from the peak.

Frith (2000) used a similar categorisation of days as Sutch (2001). She noted that roads with a morning peak generally carry more traffic on Mondays and roads with an evening peak generally carry more traffic on Fridays. Using six days of ANPR data from the A12 road in the UK she fitted a quadratic equation between the ratio of the journey time to free-flow travel time with the coefficient of variation of day-to-day variability. She found that this d2d variability increased as the mean travel time increased. This result was supported by research from Porter (2001) using the same M60 dataset. He suggested that between days variability (d2d) increased as the mean travel time increased. However it is not known how well the researchers managed to filter out the effect of any incidents. This is important to consider since an incident will cause the mean travel time to increase. Because the travel time on a day with an incident will be different to the travel time on days without an incident, then a single incident will cause both the mean travel time and the d2d TTV for that particular period to increase. This relationship between mean travel time and standard deviation could thus have been caused solely by the presence of several incidents.

Very little work has been undertaken in apportioning the amount of variability caused by each temporal scale of TTV (refer to Table 9.5). One notable exception is the work of Park et al. (1999). As an application to his short-term ANN model, Park et al. (1999) attempted to apportion the total TTV to each of the four temporal periods stated in Table 9.5. His results are repeated in Table 9.6.

<table>
<thead>
<tr>
<th>% variance</th>
<th>% variance on link 3</th>
<th>% variance on link 4</th>
<th>% variance on link 5</th>
<th>% variance on link 6</th>
<th>% variance average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within Intervals (v2v)</td>
<td>6.6</td>
<td>6.2</td>
<td>4.7</td>
<td>10.3</td>
<td>7.0</td>
</tr>
<tr>
<td>Between Intervals (p2p)</td>
<td>56.7</td>
<td>52.4</td>
<td>52.9</td>
<td>57.9</td>
<td>55.0</td>
</tr>
<tr>
<td>Between Days (d2d)</td>
<td>29.3</td>
<td>33.9</td>
<td>33.9</td>
<td>29.4</td>
<td>31.6</td>
</tr>
<tr>
<td>Between Months (m2m)</td>
<td>7.4</td>
<td>7.5</td>
<td>7.5</td>
<td>2.5</td>
<td>6.2</td>
</tr>
</tbody>
</table>

Table 9.6: TTV sources using Nested ANOVA (adapted from Park et al. (1999)).

The work shows that on an urban highway in Houston USA, the within-day (p2p) and day-to-day (d2d) variability accounted for the most TTV. It would be interesting to repeat similar work in the urban context. It would be expected that within the urban context vehicle-to-vehicle (v2v) variability would account for a higher proportion of the
9.5. Other aspects of TTV

Total TTV, due to features such as traffic lights. This is considered further in section 10.1.

9.4.2 Other empirical work on TTV

Most empirical research on measuring TTV has attempted to model TTV by first disaggregating it by temporal scale. Relatively little research into the other modes of disaggregation have been undertaken. Pearce (2001) reports of work undertaken by the Highways Agency which found that recurrent congestion accounted for 65% of all congestion on its networks, with a further 25% accounted by incident congestion. The last 10% was caused by roadworks.

Another area in which there has not been great research is in the relationship between demand (or flow) and the amount of TTV. This failing in the literature has been noted by several authors (e.g. Frith 2000). One reason why this research has not been undertaken is due to the lack of suitable data.

9.5 Other aspects of TTV

The above review has concentrated on the effect of TTV at the link and network level. However, there are other aspects of TTV, which although not researched in this thesis, should be mentioned in passing to give the reader a broader view of TTV. This includes research on TTV experienced by other modes of transport, traveller behaviour towards TTV, and the predictability of travel time.

9.5.1 Other Modes of transport

The literature review given has concentrated on TTV experienced by private vehicles on road networks. A sizeable amount of research has been undertaken on TTV experienced on other forms of transport. (e.g. Rietveld 2001; Bates et al. 2001)

9.5.2 Traveller behaviour towards travel time variability

Much research has been carried out to understand how TTV seen at the network level affects the travelling behaviour of individual drivers. A brief summary of the major areas in this field of study is given below. For comprehensive literature reviews of this area, the reader is referred to papers by Polak (1987), Bates et al. (2001), and Noland & Polak (2002).

Safety Margin

The idea of a safety margin and risk aversion was first used by Thomson (1968) who noted that a driver will allow a margin of time to allow for the fact that a journey time varies.
The safety margin, $m$, is given by Polak (1987):

$$m = (t_W - t_H) - E[r]$$

$E[r] = $Expected travel time

Polak (1987) was able to demonstrate that there is a monotonic relationship between the spread of journey times (a measure of TTV) and the safety margin. The safety-margin is also referred to as the ‘head start’ (Noland & Small 1995).

**Mean-Variance Models**

The first travel-behaviour models tried to incorporate the variance in travel times to the total cost of the journey. Jackson & Jucker (1981) proposed that the objective of each traveller was to minimise a cost function that was a function of both mean travel time and the variance of travel time. It is given by equation 9.7. Such a model is commonly termed a ‘mean-variance’ model.

$$c(T_p) = E(T_p) + \lambda Var(T_p)$$

$E(T_p) = $expected travel time - i.e. the mean.

$Var(T_p) = $Variance of the travel time

$\lambda = $A nonnegative parameter which represents the degree to which the variance of travel time is undesirable to a traveller. This is also referred to as the risk-averse parameter.

To determine the value of the parameter, $\lambda$, Jackson & Jucker (1981) performed a survey. They posed a series of questions asking the respondent which of two alternative routes they would choose to take. Route 1 had a mean travel time of 30 minutes with no delay on any days. Route 2 had a mean travel time of 20 minutes, but with the chance of a delay of $x$ minutes on one day each week. By varying the value of $x$ in 5-minute steps they sought the value of $x$ that persuaded the respondent to choose route 1 rather than route 2. This enabled them to calculate the value of $\lambda$.

An alternative to the equation 9.7 is to use the standard deviation rather than the variance as an explanatory variable of the ‘utility’ of travel (Bates et al. 2001).

$$U = \beta_t T + \beta_\sigma \sigma$$

$T =$mean travel time

$\sigma =$standard deviation of travel time

$\beta_t =$cost of a minute of expected travel time
\( \beta_\sigma = \text{cost of a minute of standard deviation of travel time.} \)

The ratio of the two parameters above is termed the reliability ratio, \( R \) (Bates et al. 2001). It states how individual travellers value reliability in travel time with respect to travel time itself. It is given by equation 9.9.

\[
R = \frac{\beta_\sigma}{\beta_t}
\]  (9.9)

Bates et al. (2001) found that for private vehicles this value tended to vary between 1.1 and 2.2, although a value of 0.79 was found in one study in London. They suggest that a reliability ratio of 1.3 is the most appropriate value for private vehicles. As an example, given a reliability ratio, \( R \), of 2, then journey A with a mean travel time of 40 minutes and a standard deviation of 4 minutes (i.e. a coefficient of variation of 10%) is as attractive as journey B with a mean travel time of 42 minutes but a standard deviation of only 3 minutes (i.e. a reduction in one minute of standard deviation is valued the same as having two minutes extra mean travel time).

### Scheduling models

Mean-Variance models do not allow for any differentiation between early and late arrivals. Equation 9.7 implies that a traveller is equally distressed by arriving 10 minutes early as they are to arriving 10 minutes late. Scheduling models overcome these limitations. An example of a scheduling model is given by Noland & Small (1995) and is given in equation 9.10.

\[
E[U] = \alpha E[T] + \beta E[SDE] + \gamma E[SDL] + \theta P_L
\]  (9.10)

\( E[U] = \text{Expected utility} \)
\( E[T] = \text{Expected travel time} \)
\( E[SDE] = \text{Expected schedule delay early} \)
\( E[SDL] = \text{Expected schedule delay late} \)
\( P_L = \text{probability of late arrival} \)
\( \alpha, \beta, \gamma, \theta = \text{parameters to be estimated} \)

### 9.5.3 Perception and Predictability

Another area of research that is related to traveller behaviour towards TTV, is the predictability of TTV. Some travel time is predictable, whereas others is not. It is this unpredictable element of travel time that has a cost, and which must be incorporated into a safety margin. For example, although a journey will take far longer during rush hour than at the middle of the night, the traveller will have predicted the amount of extra time that the journey will take. This extra time is not included in the safety-margin, but rather
9.6 Limitations of existing research

The above literature review has revealed several fundamental problems and limitations with the existing literature on TTV. Firstly, there is confusion over the use of terminology. Fundamentally there is often a failure to realise that TTV is comprised of several components. The failure to disaggregate TTV into these components reduces the usefulness of the results in the literature. As an example the failure to account for v2v variability when studying d2d variability leads to misleading results. Several authors have been aware of the importance of disaggregating TTV into its respective components, however, their research has been hampered by the second fundamental problem with much existing TTV research; namely lack of data. Much research undertaken in the literature involved recording a single link travel time on one particular journey day over several days. Such a dataset would not allow the user to take account of the v2v variability when studying the d2d variability. Only with modern technology such as ANPR cameras has it become possible to record multiple journeys at the same time of day over several days. This approach would allow for the d2d and v2v TTV to be disaggregated from one another. The inability to capture a large set of journey times in the same short time period, also explains a third failure of current research: why there is no literature (to this authors’ knowledge) on characterising v2v variability.

The datasets obtained as part of this thesis allow for all three criticisms of the existing literature to be overcome. Chapter 10 will present some empirical research on the characteristics of each temporal scale of TTV. Its main objective, however, is to consider
whether travel time estimates from the k-NN model developed in part B allow the TTV to be characterised correctly.
Chapter 10

Application of the k-NN method to the characterisation of TTV

The previous chapter has presented a brief introduction to travel time variability (TTV). It has been shown that much of the literature has suffered from a lack of sufficient data. This lack of sufficient data has prohibited researchers from properly disaggregating the different temporal scales of TTV. Even today only roads fitted with ANPR cameras are likely to have sufficient travel time records for a thorough study of TTV to be undertaken. This chapter explores whether the estimates of travel time obtained from the k-NN method allows the TTV to be characterised accurately. This is achieved by comparing the patterns of TTV measured using k-NN derived travel times with the patterns of TTV measured using actual ANPR travel time records. If it is shown that the characterisation of TTV using the k-NN data is accurate, then the k-NN method could be used to model TTV on links where sufficient ANPR data is unavailable.

Section 9.2.4 suggested four temporal scales of TTV. This chapter will focus on the ability of the k-NN method to model the vehicle-to-vehicle (v2v), and day-to-day (d2d) TTV. These components of TTV tend to be less predictable than p2p variability. Before this is undertaken, section 10.1 determines how much of the TTV each temporal scale causes. Section 10.2 describes empirical research undertaken in modelling d2d TTV and section 10.3 work describes empirical work in the modelling of v2v TTV.

10.1 Apportioning the sources of TTV

10.1.1 Disaggregating TTV into various temporal scales

The various temporal scales of TTV have been identified in section 9.2.4. It is informative to quantify the absolute and relative amount of each scale of TTV. This section presents a method to disaggregate TTV into its component d2d, p2p, and v2v parts.

This approach is based on the concept of analysis of variance (Gujarati 1995; Kottegoda & Rosso 1998). Let $X_{ijk}$ be the $k$-th individual vehicle link travel time record in the $j$-th
10.1. Apportioning the sources of TTV

period of the $i$-th day. Let there be $n_{ij}$ travel time records in the $j$-th period of the $i$-th day, and $n_i$ travel time records in the $i$-th day. Let there be a total of $D$ days, and each period is 15-minutes long - i.e. there are 96 periods in a day. If the overall mean travel time is given by $\mu$, then the total sum-of-squares, $V$, is given by equation 10.1.

$$V = \sum_{i=1}^{D} \sum_{j=1}^{96} \sum_{k=1}^{n_{ij}} (X_{ijk} - \mu)^2$$

(10.1)

Let $\bar{\alpha}_i$ be the mean travel time on day $i$, and for convenience define $F$ as 10.2:

$$F(\cdot) = \sum_{i=1}^{D} \sum_{j=1}^{96} \sum_{k=1}^{n_{ij}} \cdot$$

(10.2)

$$V = F \left( [(\bar{\alpha}_i - \mu) + (X_{ijk} - \bar{\alpha}_i)]^2 \right)$$

(10.3)

$$V = F \left( (\bar{\alpha}_i - \mu)^2 \right) + 2F \left( (\bar{\alpha}_i - \mu)(X_{ijk} - \bar{\alpha}_i) \right) + F \left( (X_{ijk} - \bar{\alpha}_i)^2 \right)$$

(10.4)

Now, the average travel time on day $i$, $\bar{\alpha}_i$, is given by equation 10.5.

$$\bar{\alpha}_i = \frac{1}{\sum_{j=1}^{96} \sum_{k=1}^{n_{ij}}} \sum_{j=1}^{96} \sum_{k=1}^{n_{ij}} X_{ijk}$$

(10.5)

Equation 10.5 can be rearranged to give 10.6.

$$\sum_{j=1}^{96} \sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{\alpha}_i) = 0$$

(10.6)

The middle term above in equation 10.4 can then be cancelled out due to equation 10.6. Thus 10.4 can be rewritten as equation 10.7.

$$V = F \left( (\bar{\alpha}_i - \mu)^2 \right) + F \left( (X_{ijk} - \bar{\alpha}_i)^2 \right)$$

(10.7)

An indication of the variability caused by the day-to-day variation is thus given by the first term. For convenience this term is called $TTV_{d2d}$ and is stated formerly in equation 10.8. It represents the travel time variability caused by day-to-day variability. Thus 10.7 can be rewritten as equation 10.9.

$$TTV_{d2d} = \sum_{i=1}^{D} \sum_{j=1}^{96} \sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{\alpha}_i)^2$$

(10.8)

$$V = TTV_{d2d} + F \left( (X_{ijk} - \bar{\alpha}_i)^2 \right)$$

(10.9)
Let $\bar{\beta}_{ij}$ be the mean travel time in period $j$ of day $i$. Then

$$V = TTV_{d2d} + F \left( (X_{ijk} - \bar{\beta}_{ij}) + (\bar{\beta}_{ij} - \bar{\alpha}_i)^2 \right)$$ (10.10)

$$V = TTV_{d2d} + F \left( (X_{ijk} - \bar{\beta}_{ij})^2 \right) + 2F \left( (\bar{\beta}_{ij} - \bar{\alpha}_i)(X_{ijk} - \bar{\beta}_{ij}) \right) + F \left( (\bar{\beta}_{ij} - \bar{\alpha}_i)^2 \right)$$ (10.11)

Once again the crossover term is zero due to equation 10.12, allowing equation 10.11 to be rewritten as equation 10.13.

$$\sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{\beta}_{ij}) = 0$$ (10.12)

$$V = TTV_{d2d} + F \left( (\bar{\beta}_{ij} - \bar{\alpha}_i)^2 \right) + F \left( (X_{ijk} - \bar{\beta}_{ij})^2 \right)$$ (10.13)

The 2nd term of the right hand side is a measure of the travel time variability due to period-to-period variability, $TTV_{p2p}$, and is given by equation 10.14.

$$TTV_{p2p} = \sum_{i=1}^{D} \sum_{j=1}^{96} ((\bar{\beta}_{ij} - \bar{\alpha}_i)^2)$$ (10.14)

The third term of equation 10.14 is a measure of the vehicle-to-vehicle variability, $TTV_{v2v}$, and is given by equation 10.15.

$$TTV_{v2v} = \sum_{i=1}^{D} \sum_{j=1}^{96} \sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{\beta}_{ij})^2$$ (10.15)

Replacing equations 10.14 and 10.15 into 10.13 gives equation 10.16.

$$V = TTV_{d2d} + TTV_{p2p} + TTV_{v2v}$$ (10.16)

That is, the total variability has been disaggregated into the variability caused by day-to-day, period-to-period, and vehicle-to-vehicle variability.

**10.1.2 Experiment and Methodology to decompose TTV into its various components**

An experiment was carried out to decompose TTV into its component parts using the 4 data-sets from Russell Square and Tottenham Court Road in both 2003 and 2004 (refer to sections 4.7 and 4.8). TTV was estimated using both k-NN and ANPR datasets. ANPR data was filtered using the Overtaking Rule as outlined in sections 4.7 and 4.8.

Estimates of travel-time were also obtained using the k-NN method. For each record in the original dataset whose ILD data passed the DSA test, the flow and occupancy data were used to estimate the travel time using the k-NN method. The relevant optimised
10.1. Apportioning the sources of TTV

k-NN settings given in chapter 8 were used. As before, all data from the same day as the record whose travel time was being estimated were excluded temporarily from the historical dataset. This ensured independence between the validation and training data.

Using equations 10.8, 10.14, 10.15, and 10.16 it was then possible to decompose the TTV into its component parts. The results of this are presented in sections 10.1.3 and 10.1.4.

10.1.3 Results - ANPR data

This section presents the results of the decomposition of TTV into its component parts for the ANPR data set. The results are presented in table 10.1.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>No records in dataset</th>
<th>mean variance (secs²)</th>
<th>day-to-day (d2d)</th>
<th>period-to-period (p2p)</th>
<th>vehicle-to-vehicle (v2v)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean (secs²)</td>
<td>Mean (secs²)</td>
<td>Mean (secs²)</td>
</tr>
<tr>
<td>Russell Square 2003</td>
<td>40,775</td>
<td>17,800</td>
<td>1,550</td>
<td>8.7</td>
<td>12,200</td>
</tr>
<tr>
<td>Russell Square 2004</td>
<td>124,636</td>
<td>13,600</td>
<td>1,590</td>
<td>11.7</td>
<td>8,840</td>
</tr>
<tr>
<td>Tottenham Court Road 2003</td>
<td>146,186</td>
<td>4,620</td>
<td>536</td>
<td>11.6</td>
<td>3,010</td>
</tr>
<tr>
<td>Tottenham Court Road 2004</td>
<td>309,530</td>
<td>2,210</td>
<td>97</td>
<td>4.4</td>
<td>989</td>
</tr>
</tbody>
</table>

Table 10.1: Sources of TTV by temporal scale. ANPR data used.

The table clearly shows that the primary source of TTV is that which occurs within the same day (i.e. p2p TTV). This source accounted for more than 50% of the total variability in most cases. The secondary major source of TTV was seen to be vehicle-to-vehicle (v2v). There appears to be at least 3 times more v2v TTV on urban roads than on highways (see table 9.6). This is because there are more control devices on urban road networks which will affect some vehicles but not others. Vehicle-to-vehicle variability will be studied in greater detail in section 10.3. Day-to-day variability was seen to account for around 10% of TTV.

The breakdown of sources of TTV for both Russell Square datasets and the Tottenham Court Road dataset of 2003 were seen to be similar to one another. One interesting observation was a dramatic decrease in p2p and d2d variability between 2003 and 2004 on the Tottenham Court Road link. Mean p2p variance per vehicle in 2004 was seen to fall to less than 1/3rd that of the 2004 value. This suggests that a change in OD patterns or simply better congestion measures have reduced the amount of p2p variability on this link. In the same period the amount of v2v TTV was seen to remain approximately the same.
10.1.4 Results - k-NN

The analysis of section 10.1.3 used the ANPR derived travel time of each record. Since the objective of part C is to see how well k-NN derived travel time data can model the TTV, this analysis was repeated using estimates of travel time obtained from the k-NN method. The results for this analysis are given in table 10.2. Due to reasons outlined in section 4.8.3 the Tottenham Court Road 2004 dataset was not included in the analysis.

From this table it can be seen that in all datasets p2p TTV accounted for over 80% of all TTV. The remainder was accounted for by d2d variability. The k-NN datasets exhibited almost no v2v TTV. This is unsurprising, since the estimate of the k-NN gives the expected travel time of a vehicle given certain flow and occupancy inputs. If these values remain constant over the same 15 minute period, the estimate will also remain constant.

For all categories of TTV, the absolute amount of TTV was less than with the ANPR dataset. Only the absolute values of p2p TTV approached those calculated using the ANPR dataset. The k-NN tended to underestimate the amount of p2p TTV by approximately 25%. One failing of the k-NN model is that it will tend to underestimate the travel time at high actual travel times, and overestimate the travel time at low travel times. Examples of this can be seen in the graphs of ANPR and k-NN travel times estimates for a three month period for the Russell Square dataset in 2004 presented in appendix C. In particular the data on Thursday 10th June 2004 shows an incident which resulted in very high travel times in the ANPR dataset. The k-NN method underestimated the effect of this incident. This reduction in performance in modelling the extreme values explains why the p2p TTV from the k-NN estimates is less than that seen with the ANPR data.

The amount of d2d TTV in the k-NN dataset was seen to be approximately half that in the ANPR dataset.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>No records in dataset</th>
<th>mean variance (secs²)</th>
<th>% day-to-day (d2d)</th>
<th>% period-to-period (p2p)</th>
<th>% vehicle-to-vehicle (v2v)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean (secs²)</td>
<td>%</td>
<td>Mean (secs²)</td>
</tr>
<tr>
<td>Russell Square 2003</td>
<td>24,718</td>
<td>7,130</td>
<td>1,170</td>
<td>16.4</td>
<td>5,810</td>
</tr>
<tr>
<td>Russell Square 2004</td>
<td>98,973</td>
<td>6,000</td>
<td>1,080</td>
<td>17.9</td>
<td>4,800</td>
</tr>
<tr>
<td>Tottenham Court Road 2003</td>
<td>104810</td>
<td>3,280</td>
<td>410</td>
<td>12.5</td>
<td>2,820</td>
</tr>
</tbody>
</table>

Table 10.2: Sources of TTV by temporal scale. k-NN estimates used.

10.1.5 Conclusion: Ability of the k-NN method to disaggregate TTV

Comparing tables 10.1 and 10.2 it can be seen that the k-NN model estimates provide an approximate estimate of the absolute amount of p2p and d2d TTV. However, the k-NN
model estimates give a poor indication of v2v TTV. An alternative way of using the k-NN method to estimate the amount of v2v TTV is presented in section 10.3. This method is based on looking at the distribution of travel time of the nearest neighbors. Due to the poor estimate of absolute v2v TTV, the k-NN model apportioned a higher percentage of TTV to both the d2d and p2p TTV than was the case seen using the ANPR dataset.

The next two sections will study in greater detail how the k-NN method characterises d2d and v2v TTV.

10.2 Modelling day-to-day TTV using the k-NN method

Day-to-day (d2d) TTV refers to the phenomenon that the mean travel time for vehicles travelling in the same 15-minute period of the day will vary from day-to-day. For example, the mean travel-time on a link between 08:45 and 09:00 measured on a Monday (workday) is likely to be different to the mean travel-time on a link between 08:45 and 09:00 measured on a Sunday (weekend). As outlined in section 9.4.1 this component of TTV has been widely studied, albeit with a common failing to properly control for p2p and v2v TTV. This section undertakes three analyses of d2d TTV; the identification of similar days of week, variation of d2d TTV over different periods of the day, and modelling the relationship between d2d TTV and the mean travel time. This analysis is undertaken using both the actual ANPR travel time records and the estimates of travel time obtained from the k-NN model. The work is undertaken in this fashion since the main objective of Part C is not to study travel-time variability per-se, but to determine whether k-NN derived estimates of travel time could be used to model TTV.

10.2.1 Identification of similar days-of-the-week

The first analysis undertaken was to identify which days-of-the-week were similar to one another. This was undertaken though the use of cross-correlation analysis (refer to section 5.5). The research was undertaken with both the ANPR dataset and the k-NN estimated dataset.

For this experiment the 2004 Russell Square dataset, explained in section 4.8.2 was used. This data contained 91 days of data (i.e. 13 weeks). There were thus a total of 8736 15-minute periods. For each of these 15-minute periods the expected travel time was calculated using both the raw ANPR data and the k-NN method (see section 4.8.2 for the settings used). To calculate the expected travel time from the ANPR dataset, the arithmetic mean of all the individual travel time records which fell in the period were used. The number of travel time records used to calculate this was also recorded. The results of this can be seen in appendix C which shows the ANPR and k-NN estimated travel time for each of the 8,736 periods between the 1st April and 30th June 2004. It should be noted that both measures have missing values.

Having calculated the expected travel time for each period, the ANPR period data was
10.2. Modelling day-to-day TTV using the k-NN method

Figure 10.1: Variation of mean travel time with the period of the day (Sunday through to Wednesday). Both ANPR and k-NN results are shown for the Russell Square 2004 dataset.

then grouped into bins whose data came from the same day-of-week and period-of-day. Only ANPR period data which had been calculated using more than 7 individual travel time records was put into these bins. A minimum value of 7 was used to reduce the effect of outliers and to control for v2v TTV. Ideally a higher value would be used, but this would have excluded many more 15-minute time periods from the analysis. There were a total of 672 bins (7 days * 96 periods). For each bin with more than 2 records, the mean travel time was calculated. Once again this criteria to have a minimum of 2 records was stipulated to reduce the effect of outliers. This process was then repeated using the k-NN dataset. The results of this are given in figure 10.1 and figure 10.2. Data is plotted by day-of-week. These figures suggest that the k-NN estimates of mean travel time match the ANPR data well.

The next stage was to perform cross correlation between the mean travel time profile of each day of the week. The correlation coefficient was calculated using equation 10.17.

\[
r_{x,y} = \frac{1}{96 \cdot s_x \cdot s_y} \sum_{i=1}^{96} (x_i - \mu_x)(y_i - \mu_y)
\]  

(10.17)
Figure 10.2: Variation of mean travel time with the period of the day (Thursday through to Saturday). Both ANPR and k-NN results are shown for the Russell Square 2004 dataset.
10.2. Modelling day-to-day TTV using the k-NN method

\[ r_{x,y} = \text{Correlation coefficient between data-series } x \text{ and data-series } y. \]

\[ x, y = \text{data-series containing mean travel time for each of the 96 15-minute periods of a certain day-of-the-week.} \]

\[ i = 15\text{-minute period of the day} \]

\[ \mu_x, \mu_y = \text{mean travel time for days } x \text{ and } y \text{ respectively.} \]

\[ s_x, s_y = \text{standard deviation of travel times for days } x \text{ and } y \text{ respectively.} \]

It should be noted that \(-1 \leq r_{x,y} \leq 1\). A value of \(r_{x,y} = 1\) implies that data series \(x\) and \(y\) are perfectly linearly correlated with one another.

This procedure was undertaken using the ANPR and k-NN datasets. The results from the ANPR dataset are presented in table 10.3, and the k-NN dataset in table 10.4.

<table>
<thead>
<tr>
<th></th>
<th>sun</th>
<th>mon</th>
<th>tue</th>
<th>wed</th>
<th>thu</th>
<th>fri</th>
<th>sat</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>1.00</td>
<td>0.46</td>
<td>0.50</td>
<td>0.43</td>
<td>0.64</td>
<td>0.67</td>
<td>0.85</td>
</tr>
<tr>
<td>mon</td>
<td>1.00</td>
<td>0.93</td>
<td>0.94</td>
<td>0.90</td>
<td>0.93</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>tue</td>
<td>1.00</td>
<td>0.91</td>
<td>0.93</td>
<td>0.94</td>
<td>0.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wed</td>
<td></td>
<td>1.00</td>
<td>0.89</td>
<td>0.91</td>
<td>0.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thu</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.95</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fri</td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 10.3: Correlation between mean travel times from different days of the week. ANPR data from the Russell Square 2004 dataset was used.

<table>
<thead>
<tr>
<th></th>
<th>sun</th>
<th>mon</th>
<th>tue</th>
<th>wed</th>
<th>thu</th>
<th>fri</th>
<th>sat</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>1.00</td>
<td>0.51</td>
<td>0.57</td>
<td>0.61</td>
<td>0.63</td>
<td>0.60</td>
<td>0.89</td>
</tr>
<tr>
<td>mon</td>
<td>1.00</td>
<td>0.96</td>
<td>0.97</td>
<td>0.95</td>
<td>0.95</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>tue</td>
<td></td>
<td>1.00</td>
<td>0.95</td>
<td>0.95</td>
<td>0.97</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>wed</td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.94</td>
<td>0.94</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>thu</td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.95</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>fri</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>sat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

Table 10.4: Correlation between mean travel times from different days of the week. k-NN data from the Russell Square 2004 dataset was used.

Analysis from both the ANPR and k-NN datasets suggest the same underlying pattern. That is all weekdays (Monday through to Friday) are similar with each other, with a correlation coefficient greater than 0.9. In addition there is a high correlation between Saturday and Sunday. The major difference between the k-NN and ANPR dataset concerns the estimates from correlation between Saturday and Monday, Tuesday, and Wednesday. The reason for this is unknown. Nevertheless, it can still be concluded that the results from the k-NN dataset are very similar to those from the ANPR dataset. Thus k-NN derived estimates of travel time can be successfully used to determine those days-of-the-week which are similar to one another.
10.2. Modelling day-to-day TTV using the k-NN method

10.2.2 Variation of d2d TTV over the course of the day

The second analysis undertaken was to investigate how the amount of d2d TTV varied through the course of the day. This section outlines a small study into this. Similar to other research in the chapter, this section presents the results of both the ANPR dataset and the k-NN estimated dataset.

For this experiment the same dataset and methodology as described in section 10.2.1 was used. In addition to calculating the mean travel time, the standard deviation of travel time for each bin was also calculated. This standard deviation represented the amount of d2d TTV. The results of this are given in figure 10.3 and figure 10.4. Data is plotted by day-of-week.

From the results it can be seen that the standard deviation varies over the day with a small amount of variation during the night and a large amount of variation occurring in the late afternoon and early evening. The important thing to note is that the estimate of standard deviation using the k-NN method is seen to be similar to the estimate of standard deviation using the ANPR data set. This suggests that k-NN derived data could be used.
10.2. Modelling day-to-day TTV using the k-NN method

Figure 10.4: Variation of day-to-day TTV with the period of the day (Thursday through to Saturday). Both ANPR and k-NN results are shown for the Russell Square 2004 dataset.
to study this aspect of d2d TTV.

10.2.3 Relationship between d2d TTV and the mean travel time

Research by Black et al. (2004), reviewed in section 9.4.1, suggested a relationship between the day-to-day coefficient of variation and the congestion index. An experiment was thus undertaken to see whether such a relationship existed for the Russell Square 2004 dataset. As before, this research was undertaken using both ANPR and k-NN data. The coefficient of variation and congestion index was calculated for each of the 692 possible combinations of day-of-week and period-of-day, in a fashion similar to that outlined in section 10.2.2. In order to calculate the congestion index it was first necessary to estimate the free-flow travel time. This was estimated as the 5th lowest percentile of travel times. This value was chosen after review of the cumulative distribution function of the travel times for the Russell Square 2004 dataset shown in figure D.5. For the ANPR dataset the free flow travel time was calculated as 127 seconds, and for the k-NN data set it was estimated at 137.1 seconds. The fact that the k-NN method overestimated the free-flow travel time is consistent with the results shown in figures 7.8 and 7.9 that the k-NN method tends to overestimate the travel times at very low actual travel times. Equation 10.18 was then fitted to each of the datasets using OLS regression (albeit a log transformation was first required).

\[ CV = \alpha_1 \cdot CI^{\alpha_2} + e \]  

(10.18)

\( CV \) = Coefficient of Variation
\( CI \) = Congestion Index - see equation 9.4.
\( e \) = error term
\( \alpha_1, \alpha_2 \) = parameters to be estimated.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>( \alpha_1 )</th>
<th>t-stat, ( \alpha_1 )</th>
<th>( \alpha_2 )</th>
<th>t-stat, ( \alpha_2 )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANPR</td>
<td>0.068512</td>
<td>-77.5731</td>
<td>1.2211</td>
<td>24.2118</td>
<td>0.50094</td>
</tr>
<tr>
<td>k-NN</td>
<td>0.050271</td>
<td>-80.866</td>
<td>2.0652</td>
<td>32.4213</td>
<td>0.61072</td>
</tr>
</tbody>
</table>

Table 10.5: Relationship between the congestion index and the d2d coefficient of variation.

The results of this are shown in table 10.5 and figures 10.5 and 10.6 for the ANPR and k-NN datasets respectively. All parameters are seen to be significant. The figures suggest that equation 10.18 fits the data poorly. In addition the values of \( \alpha_1 \) and \( \alpha_2 \) obtained from the ANPR data differ from those of the k-NN data. Interestingly, the model fit to the k-NN data has a higher \( R^2 \) value than the model fit to the ANPR data. One reason for this is that as shown in section 10.1, the data derived from the k-NN method tends to have lower levels of absolute d2d TTV than the ANPR data. This would reduce the value of \( CV \), reducing the amount of scatter in the plot, and hence resulting in a better model.
10.2. Modelling day-to-day TTV using the k-NN method

Equation 10.18 was fit to the data. The ANPR dataset of Russell Square 2004 was used fit. From this research it can be concluded that the model of Black et al. (2004) does not appear to be appropriate for the Russell Square link.

10.2.4 Conclusion: Ability of the k-NN method to model d2d TTV

The results presented in this section suggest that travel times derived from the k-NN model could be used to accurately model day-to-day TTV. Analysis using the k-NN dataset was able to identify those days of the week which, according to analysis of the ANPR dataset, were similar to one another. Likewise the patterns of standard deviation observed from the k-NN and ANPR dataset were similar to one another. As suggested in section 8.6 the k-NN method could be used to control for v2v TTV where there are insufficient records in any one time period.
Figure 10.6: Relationship between the congestion index and the d2d coefficient of variation. Equation 10.18 was fit to the data. The k-NN dataset of Russell Square 2004 was used.
10.3 Modelling vehicle-to-vehicle TTV using the k-NN method

Another important source of TTV is the vehicle-to-vehicle variability. This type of variability is the ‘random variation’ proposed by Wong & Sussman (1973). It accounts for vehicles starting within a couple of minutes of one another taking different journey times. Causes of such variability include: the effect of traffic lights (does a vehicle get a red or green light), stopping buses, vehicles turning mid-link, pedestrians crossing at zebra crossings, and differing cruise speeds of different drivers. Very little empirical research has been undertaken into this field. The primary reason for this is that many previous studies in TTV used a single vehicle to obtain travel time records, and it was thus not possible to collect travel times from vehicles that started the journey in the same short time period on the same day. With GPS equipped vehicles and ANPR cameras it is feasible to obtain many travel time records within the same 15-minute time period. However in practice, these technologies, and in particular GPS, may at present be unable to capture a sufficient proportion of the vehicle population. It is thus necessary to determine how the sample size affects the accuracy of any characterisation of v2v variability. Section 10.3.1 firstly defines some measures by which the vehicle-to-vehicle distribution can be characterised. Section 10.3.2 then outlines an experiment undertaken to assess the effect of the sample size on the accuracy of measuring the vehicle-to-vehicle distribution.

An approach to overcoming the limitation of a small sample size is then proposed. This is based upon the k-NN model developed in part B. Although in its existing form the k-NN model only outputs the expected travel time for a vehicle given the flows and occupancies measured by the ILDs, it is possible to obtain estimates of the vehicle-to-vehicle distribution using an extension of the k-NN method. The approach used is outlined in section 10.3.3. The ability of this estimator to model vehicle-to-vehicle variability will be investigated using two experiments outlined in sections 10.3.4 and 10.3.5.

Section 10.3.6 looks at a different aspect of v2v TTV. In particular it outlines a small experiment which investigates the variation of v2v TTV with the mean travel time in the period. This is undertaken using both the ANPR and k-NN datasets. Once again the primary objective of this research is to identify whether the k-NN derived results can be used to model v2v TTV.

10.3.1 Characterising v2v TTV

Ideally the k-NN method should be able to estimate the same distribution of travel-times in a 15-minute time period as is obtained from ANPR data. It is possible to check whether the distribution of travel times matches that from the ANPR data by performing a Kolmogorov-Smirnov test. This is undertaken in section 10.3.4. However, the Kolmogorov-Smirnov test requires at least 25 records from each data-source in any time period (Kottegoda & Rosso 1998). Thus it may not be possible to perform this test. It is thus necessary to define other ways to characterise the travel time distribution, and hence the vehicle-to-vehicle TTV, in any 15-minute period. It is proposed to characterise
10.3. Modelling vehicle-to-vehicle TTV using the k-NN method

the v2v TTV using the following five properties of the travel time distribution in the 15-minute period:

- Mean
- Standard deviation
- Coefficient of Variation - i.e. normalised standard deviation.
- 95th percentile travel time - i.e. 95% off all travel time records are smaller than this.
- Buffer Index - see equation 9.3. This is a normalised measure of the 95th percentile travel time.

These five metrics will be known collectively as the v2v metrics. Several estimators are proposed to estimate each of the v2v metrics. These are outlined in table 10.6.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>v2v metric to be estimated</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - k-NN estimate</td>
<td>mean</td>
<td>Optimised model used as defined in section 7.4.</td>
</tr>
<tr>
<td>2 - Sample mean (mean)</td>
<td>mean</td>
<td>The mean of the travel times in the sample.</td>
</tr>
<tr>
<td>3 - Sample median (median)</td>
<td>mean</td>
<td>The median of the travel times in the sample.</td>
</tr>
<tr>
<td>4 - Sample standard deviation (STD)</td>
<td>Standard deviation</td>
<td>The standard deviation of the travel times in the sample.</td>
</tr>
<tr>
<td>5 - Percentile standard deviation of the sample (r-STD)</td>
<td>Standard deviation</td>
<td>The 84th (p84) and 16th (p16) percentile travel time of the sample. The percentile travel time is defined as (p84 - p16) /2. See section 2.9.2.</td>
</tr>
<tr>
<td>6 - estimator 4 / estimator 2 (cov)</td>
<td>Coefficient of variation</td>
<td>See equation 9.2.</td>
</tr>
<tr>
<td>7 - 95th percentile (p95)</td>
<td>95th percentile</td>
<td>The 95th percentile travel time of the sample.</td>
</tr>
<tr>
<td>8 - (estimator 7 - estimator 2) (BI) / estimator 2</td>
<td>Buffer Index</td>
<td>Buffer index - see equation 9.3</td>
</tr>
</tbody>
</table>

Table 10.6: Estimators used to calculate the v2v metrics

10.3.2 Effect of the sample size on the estimation of the v2v metrics

In section 2.3.3 it was shown that the accuracy of the estimate of the population mean was affected by the sample size. In a similar fashion the accuracy in estimating the v2v metrics will be affected by the sample size. An experiment was thus undertaken to assess this.
Data

As before, the 2004 Russell Square dataset was used. In order to ensure that the actual v2v metrics could be calculated, any 15-minute period which had less than 25 travel time records was excluded from the data set. In addition all days whose ILD data failed the DSA test were also excluded. This filtering excluded 84% of all the time periods leaving a total of 1395 15-minute time periods for the study.

Methodology

For each 15-minute period, the actual v2v metrics were calculated using all $N$ travel time records available in the period. A total of $n$ records were then taken from the $N$ original records and the v2v metrics calculated using the estimators of table 10.6. This was repeated for each of the 1395 periods. Due to the random nature of selecting the sample, this whole process was repeated 50 times. After this the overall accuracy, measured by MAPE and Mean Percentage Error (MPE), of the estimates were calculated for this sample size. Using the MPE in addition to MAPE allowed biases in the estimates to be readily seen. The whole process was then repeated for different sample sizes. Since some time-periods contained only 25 travel time records, the maximum value of $n$ used was 20.

<table>
<thead>
<tr>
<th>sample size</th>
<th>MAPE</th>
<th>v2v metric</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>P95</th>
<th>COV</th>
<th>Buffer Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td></td>
<td>mean</td>
<td>1.73</td>
<td>3.35</td>
<td>14.0</td>
<td>7.4</td>
<td>7.5</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>median</td>
<td>2.47</td>
<td>4.12</td>
<td>17.2</td>
<td>10.8</td>
<td>10.9</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>r-STD</td>
<td>3.41</td>
<td>4.71</td>
<td>21.0</td>
<td>15.5</td>
<td>15.6</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>STD</td>
<td>4.44</td>
<td>6.08</td>
<td>22.7</td>
<td>20.6</td>
<td>20.8</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>cov</td>
<td>5.52</td>
<td>7.12</td>
<td>26.8</td>
<td>26.2</td>
<td>26.5</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>p95</td>
<td>6.17</td>
<td>7.14</td>
<td>30.9</td>
<td>30.7</td>
<td>31.0</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>BI</td>
<td>7.35</td>
<td>8.95</td>
<td>37.9</td>
<td>38.4</td>
<td>38.7</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>9.12</td>
<td>9.12</td>
<td>53.2</td>
<td>54.3</td>
<td>54.6</td>
</tr>
</tbody>
</table>

Refer to table 10.6 for explanation of estimators.

Table 10.7: Effect of the sample size on the MAPE of the estimates of the characteristics of the vehicle-to-vehicle travel time distribution. The ANPR data from the Russell Square 2004 dataset was used.

Results

Tables 10.7 and 10.8 show the results of the above experiment. They show how the accuracy of the estimate of the v2v metrics is affected by the size of the sample used. Table 10.7 shows that as the size of the sample decreases so the accuracy of the v2v metrics decreases. This relationship appears to be monotonic. Interestingly the robust estimators (median and r-STD) perform less well than the non-robust estimators. Table 10.8 shows
10.3. Modelling vehicle-to-vehicle TTV using the k-NN method

Refer to table 10.6 for explanation of estimators.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Mean</th>
<th>Median</th>
<th>r-STD</th>
<th>STD</th>
<th>Cov</th>
<th>P95</th>
<th>COV</th>
<th>Buffer Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.01</td>
<td>0.38</td>
<td>-0.12</td>
<td>0.47</td>
<td>0.43</td>
<td>-0.11</td>
<td>-1.36</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.02</td>
<td>0.35</td>
<td>0.75</td>
<td>0.96</td>
<td>0.88</td>
<td>0.18</td>
<td>-0.31</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.04</td>
<td>0.37</td>
<td>1.82</td>
<td>1.83</td>
<td>1.66</td>
<td>1.12</td>
<td>3.87</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-0.04</td>
<td>0.29</td>
<td>1.35</td>
<td>3.22</td>
<td>3.05</td>
<td>3.36</td>
<td>14.82</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>0.34</td>
<td>4.02</td>
<td>5.44</td>
<td>5.09</td>
<td>5.93</td>
<td>26.48</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.07</td>
<td>0.29</td>
<td>7.60</td>
<td>7.38</td>
<td>6.94</td>
<td>7.64</td>
<td>34.55</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.07</td>
<td>0.25</td>
<td>15.46</td>
<td>10.97</td>
<td>10.42</td>
<td>10.03</td>
<td>46.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.08</td>
<td>0.08</td>
<td>43.45</td>
<td>20.02</td>
<td>19.29</td>
<td>13.71</td>
<td>63.74</td>
<td></td>
</tr>
</tbody>
</table>

Table 10.8: Effect of the sample size on the MPE of the estimates of the characteristics of the vehicle-to-vehicle travel time distribution. The ANPR data from the Russell Square 2004 dataset was used.
10.3. Modelling vehicle-to-vehicle TTV using the k-NN method

The MPE of all estimates. Since the MPE of the estimates of all v2v metrics tends to be positive, this suggests that the estimators are all biased towards underestimating the value of the actual metrics.

10.3.3 Estimating the vehicle to vehicle TTV using the k-NN method

The k-NN model developed in part B estimates the expected travel time of a vehicle given a feature vector of flows and occupancies. In its existing form, the model does not give any indication of the v2v TTV. However, an estimate of v2v TTV can be easily obtained from the k-NN method. This is because it is assumed that the records selected by the k-nearest neighbors come from the same underlying travel time distribution (see section 6.3). If this is truly the case, then any variation in the travel times of those records with a similar input feature vector is due solely to v2v TTV. Thus the travel times of the k nearest neighbors can be used not only to estimate the expected travel time, but also to estimate other characteristics of the travel time distribution such as the standard deviation and CoV (i.e. measures of v2v TTV).

It is accepted however, that in reality not all travel time records of the k-nearest neighbors may come from the same underlying distribution. Because of this the top and bottom 5% of all travel times were discarded from the data set. The remaining 90% of all travel times were used to estimate the v2v metrics. It is realised that the choice of the threshold percentile at which to cut off the travel time records may affect the estimates of the v2v metrics. A figure of 5% was chosen since it was thought that such a value would allow most outliers to be omitted from the dataset yet still ensure a wide range of travel times in the sample.

10.3.4 Ability of the k-NN method to model the v2v distribution

An experiment was devised to assess how accurately the k-NN method could model the actual v2v variability. This was undertaken by seeing whether the distribution of travel times in a 15-minute period estimated by the k-NN method was similar to that of the actual ANPR data.

Data

As before, the 2004 Russell Square dataset was used. It was necessary to ensure that the actual characteristics of the v2v TTV could be calculated, to enable the performance of the k-NN derived estimates of the v2v characteristics to be determined. Thus any 15-minute period which had less than 25 travel time records was excluded from the research. In addition all days whose ILD data failed the DSA test was excluded. This filtering excluded 84% of all the time periods leaving a total of 1395 15-minute time periods for the study. As reported in section 4.8.2 there was a total of 98,973 travel time records in the historical data set after screening using the Overtaking Rule and DSA treatment.
Methodology

For each of the 1395 15-minute time periods the distribution of travel times was estimated using the k-NN method. This was achieved by firstly identifying the 100 historical records which had flows and occupancies closest to those of the relevant time-period. The SE-VAR distance metric was used as recommended for this data set in section 8.4. The Kolmogorov-Smirnov two sample test (Kottegoda & Rosso 1998, p.288) was then used to determine whether the distribution of travel times in a 15-minute period estimated by the k-NN method (i.e. the distribution of the 100 closest historical neighbors) was similar to that of the actual ANPR data.

The test statistic is given by the maximum absolute difference between the cumulative distribution function of both distributions, \( D_{m,n} \). The two samples are said to come from the same distribution if the condition of equation 10.19 is met.

\[
\sqrt{\frac{mn}{m+n}} D_{m,n} \leq z
\]  

(10.19)

\( m \) = number of observations in first sample
\( n \) = number of observations in second sample

For the 95% intervals, \( z \) is given by 1.3581 (Kottegoda & Rosso 1998).

Results

As an example of the Kolmogorov-Smirnov test, figure 10.7 shows the tests undertaken for two 15-minute time periods. In the top example from Saturday 3rd April 2004 between 00:00 and 00:15, there are 28 ANPR records in the 15-minute period and after filtering the top and bottom 5 percentile travel times there were 91 records from the k-NN method. The maximum difference was 0.3022 which is slightly greater than the permissible 0.2935 at the 95% confidence interval. Thus the two sample distributions are said to come from different distributions. In the bottom example from Saturday 3rd April 2004 between 13:00 and 13:15, there are 32 ANPR records in the 15-minute period and after filtering the top and bottom 5 percentile travel times there were 91 records from the k-NN method. The maximum difference was 0.178 which is less than the permissible 0.279 at the 95% confidence interval. Thus the two sample distributions are said to come from same distributions.

The Kolmogorov-Smirnov test was performed on all 1395 15-minute time periods. The set of travel times selected by the k-NN method was said to come from the same underlying distribution as the ANPR travel times for 624 time-periods. Thus the k-NN method is able to model the 15-minute travel time distribution in approximately half of all 15-minute time periods (44.8%).
Figure 10.7: Kolmogorov-Smirnov test to determine if the distribution of travel times from the k-NN method is the same as the actual distribution.
10.3.5 Ability of the k-NN method to estimate the v2v metrics

The results from the previous section suggested that the k-NN method could be used to model the 15-minute travel time distribution to a certain degree. This section presents an experiment that was undertaken to determine what explicit characteristics of the v2v variability the k-NN method was able to model.

Data and Methodology

The data used for this experiment was the same as that outlined in section 10.3.4. For each of the 1395 15-minute time period with more than 25 records, the v2v metrics were calculated using the estimators described in table 10.6. An additional estimator was used to estimate the population mean travel time. This was the ‘k-NN estimate’ which used the optimised k-NN model described in section 8.4.

In order to calculate the v2v metrics for a 15-minute time period, the travel time records from the same day were removed from the historical data set. This ensured independence between the validation and training data sets. As was the practice in chapters 8 and 9 a subset of 15,000 records were then chosen to form the training data set.

The 15-minute aggregated flow and occupancy for all three detectors on the link (see figure 4.8) were then obtained from the ASTRID records. These 6 fields were used to constitute the input feature-vector. The 5000 nearest neighbors to this feature vector were then identified from the training set. The distance metric used to do this was the SE-VAR (see equation 6.9). The highest 5% and lowest 5% of travel time records were then excluded from the data set. The rationale for doing this has already been given in section 10.3.3. It is recognised however, that the estimates of v2v metrics could be sensitive to the choice to remove the upper and lower 5-th percentile of travel time records. Using the remaining 90% of travel time records the v2v metrics were then calculated using the estimators outlined in table 10.6. The latter part of the above process was repeated, but using different values of nearest-neighbors, $k$, in the calculation of the v2v metrics. The v2v metrics were thus calculated for 18 different values of $k$. These were: 10, 20, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 3000, 4000, and 5000.

The whole process outlined above was repeated for all 1395 15-minute periods. It was then possible to calculate the accuracy of the k-NN method in estimating the v2v metrics for each value of $k$ used.

Results

Tables 10.9 and 10.10 show the results of the above experiment. They show how the MAPE and MPE of the k-NN method in estimating the v2v metrics is affected by the number of nearest neighbors used. It should be noted that in the 4th column of these tables, the optimal k-NN estimator of the population mean was not a function of the value of $k$. This estimator used the Lowess LEM, SE-VAR distance metric, and a fixed
10.3. Modelling vehicle-to-vehicle TTV using the k-NN method

Figure 10.8: Effect of the number of nearest neighbors used on the ability of the k-NN method to estimate the v2v metric. The Russell Square 2004 dataset was used.
### Table 10.9: Effect of the number of nearest neighbours chosen on the ability of the k-NN model to estimate the characteristics of the vehicle-to-vehicle distribution. Accuracy measured by MAPE. The Russell Square 2004 dataset was used.

<table>
<thead>
<tr>
<th>no nearest neighbours</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>P95</th>
<th>COV</th>
<th>Buffer Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>13.85</td>
<td>12.87</td>
<td>10.51</td>
<td>51.3</td>
<td>72.3</td>
</tr>
<tr>
<td>20</td>
<td>11.25</td>
<td>11.27</td>
<td>10.51</td>
<td>39.7</td>
<td>38.5</td>
</tr>
<tr>
<td>50</td>
<td>10.99</td>
<td>10.78</td>
<td>10.51</td>
<td>33.6</td>
<td>32.7</td>
</tr>
<tr>
<td>100</td>
<td>10.56</td>
<td>10.36</td>
<td>10.51</td>
<td>31.4</td>
<td>28.0</td>
</tr>
<tr>
<td>200</td>
<td>10.58</td>
<td>10.38</td>
<td>10.51</td>
<td>30.9</td>
<td>27.0</td>
</tr>
<tr>
<td>300</td>
<td>10.65</td>
<td>10.47</td>
<td>10.51</td>
<td>30.9</td>
<td>26.7</td>
</tr>
<tr>
<td>400</td>
<td>10.74</td>
<td>10.57</td>
<td>10.51</td>
<td>31.0</td>
<td>26.6</td>
</tr>
<tr>
<td>500</td>
<td>10.80</td>
<td>10.63</td>
<td>10.51</td>
<td>31.1</td>
<td>26.7</td>
</tr>
<tr>
<td>600</td>
<td>10.85</td>
<td>10.69</td>
<td>10.51</td>
<td>31.4</td>
<td>26.9</td>
</tr>
<tr>
<td>700</td>
<td>10.92</td>
<td>10.75</td>
<td>10.51</td>
<td>31.7</td>
<td>27.0</td>
</tr>
<tr>
<td>800</td>
<td>10.97</td>
<td>10.81</td>
<td>10.51</td>
<td>32.0</td>
<td>27.3</td>
</tr>
<tr>
<td>900</td>
<td>11.03</td>
<td>10.86</td>
<td>10.51</td>
<td>32.3</td>
<td>27.6</td>
</tr>
<tr>
<td>1000</td>
<td>11.09</td>
<td>10.92</td>
<td>10.51</td>
<td>32.7</td>
<td>27.9</td>
</tr>
<tr>
<td>1500</td>
<td>11.36</td>
<td>11.16</td>
<td>10.51</td>
<td>34.2</td>
<td>29.1</td>
</tr>
<tr>
<td>2000</td>
<td>11.72</td>
<td>11.48</td>
<td>10.51</td>
<td>35.7</td>
<td>30.4</td>
</tr>
<tr>
<td>3000</td>
<td>12.58</td>
<td>12.24</td>
<td>10.51</td>
<td>38.8</td>
<td>33.2</td>
</tr>
<tr>
<td>4000</td>
<td>13.47</td>
<td>13.05</td>
<td>10.51</td>
<td>41.7</td>
<td>35.6</td>
</tr>
<tr>
<td>5000</td>
<td>14.31</td>
<td>13.85</td>
<td>10.51</td>
<td>44.1</td>
<td>37.7</td>
</tr>
</tbody>
</table>
10.3. Modelling vehicle-to-vehicle TTV using the k-NN method

<table>
<thead>
<tr>
<th>no nearest neighbours</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>P95</th>
<th>COV</th>
<th>Buffer Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>v2v metric</td>
<td>mean</td>
<td>median</td>
<td>optimal k-NN</td>
<td>r-STD</td>
</tr>
<tr>
<td>10</td>
<td>-4.04</td>
<td>-1.06</td>
<td>-1.99</td>
<td>-34.6</td>
<td>-59.4</td>
</tr>
<tr>
<td>20</td>
<td>-0.39</td>
<td>0.76</td>
<td>-1.99</td>
<td>-25.4</td>
<td>-25.3</td>
</tr>
<tr>
<td>50</td>
<td>-2.18</td>
<td>-0.41</td>
<td>-1.99</td>
<td>-22.5</td>
<td>-20.5</td>
</tr>
<tr>
<td>100</td>
<td>-1.30</td>
<td>0.19</td>
<td>-1.99</td>
<td>-21.3</td>
<td>-15.9</td>
</tr>
<tr>
<td>200</td>
<td>-1.19</td>
<td>0.37</td>
<td>-1.99</td>
<td>-22.0</td>
<td>-15.6</td>
</tr>
<tr>
<td>300</td>
<td>-1.02</td>
<td>0.59</td>
<td>-1.99</td>
<td>-22.6</td>
<td>-15.7</td>
</tr>
<tr>
<td>400</td>
<td>-0.89</td>
<td>0.74</td>
<td>-1.99</td>
<td>-23.1</td>
<td>-16.1</td>
</tr>
<tr>
<td>500</td>
<td>-0.78</td>
<td>0.89</td>
<td>-1.99</td>
<td>-23.6</td>
<td>-16.5</td>
</tr>
<tr>
<td>600</td>
<td>-0.70</td>
<td>1.00</td>
<td>-1.99</td>
<td>-24.1</td>
<td>-17.0</td>
</tr>
<tr>
<td>700</td>
<td>-0.63</td>
<td>1.11</td>
<td>-1.99</td>
<td>-24.5</td>
<td>-17.4</td>
</tr>
<tr>
<td>800</td>
<td>-0.58</td>
<td>1.18</td>
<td>-1.99</td>
<td>-25.0</td>
<td>-17.9</td>
</tr>
<tr>
<td>900</td>
<td>-0.54</td>
<td>1.25</td>
<td>-1.99</td>
<td>-25.4</td>
<td>-18.3</td>
</tr>
<tr>
<td>1000</td>
<td>-0.52</td>
<td>1.31</td>
<td>-1.99</td>
<td>-26.0</td>
<td>-18.8</td>
</tr>
<tr>
<td>1500</td>
<td>-0.50</td>
<td>1.46</td>
<td>-1.99</td>
<td>-28.1</td>
<td>-21.0</td>
</tr>
<tr>
<td>2000</td>
<td>-0.62</td>
<td>1.44</td>
<td>-1.99</td>
<td>-30.3</td>
<td>-23.0</td>
</tr>
<tr>
<td>3000</td>
<td>-1.00</td>
<td>1.27</td>
<td>-1.99</td>
<td>-34.1</td>
<td>-26.8</td>
</tr>
<tr>
<td>4000</td>
<td>-1.40</td>
<td>0.99</td>
<td>-1.99</td>
<td>-37.4</td>
<td>-29.9</td>
</tr>
<tr>
<td>5000</td>
<td>-1.76</td>
<td>0.72</td>
<td>-1.99</td>
<td>-40.1</td>
<td>-32.3</td>
</tr>
</tbody>
</table>

Table 10.10: Effect of the number of nearest neighbors chosen on the ability of the k-NN model to estimate the characteristics of the vehicle-to-vehicle distribution. Accuracy measured by MPE. The Russell Square 2004 dataset was used.
10.3. Modelling vehicle-to-vehicle TTV using the k-NN method

value of $k$ of 2508 to estimate the mean travel time as stated in section 8.4. The MAPE results are shown graphically, one plot for each of the v2v metrics in figure 10.8.

Tables 10.9 and 10.10 show that the characteristics of the estimates of the the v2v metrics calculated using the k-NN method differ from those calculated directly from ANPR travel time data:

- Firstly, there is no monotonic relationship between the value of $k$ and the accuracy of v2v metrics. Figure 10.8 suggests that the optimal number of nearest neighbors to use to estimate all five v2v metrics is around 400. From table 10.9 it can be seen that when $k = 400$, the k-NN ULTT model estimate is the best estimator for the population mean, and the sample standard deviation of the $k$ travel times is the best estimator for the population standard deviation.

- Secondly, the robust estimator of the population mean (i.e. the sample median) outperforms the non-robust estimator of population mean (i.e. the sample mean). This is the opposite to that found when using the ANPR records to calculate the v2v metrics.

- Table 10.10 suggests that estimates of the v2v metrics are biased. The negative sign suggests that estimates for all estimators are biased towards overestimating the actual value. This is the opposite to the case when the ANPR records are used to calculate the v2v metrics (see table 10.8).

Comparisons between table 10.7 and table 10.9 (number of nearest neighbors, $k = 400$), suggest that if at least 2 travel-time records from a 15-minute time period can be measured, then a more accurate estimate of the population mean travel time can be obtained than is obtained from the k-NN derived estimate.

However, if information about the v2v TTV is desired, then the k-NN method (where $k = 400$) will provide more accurate estimates of the standard deviation than can be obtained from direct measurement of vehicle travel times if less than 5 such readings in any 15-minute time period are obtained. In effect the k-NN method allows travel time records from different 15-minute time periods but the same underlying travel time distribution to be aggregated together to allow for a more accurate estimate of v2v TTV for that 15-minute time period to be obtained.

10.3.6 Variation of v2v TTV with mean travel time

The first four sub-sections of section 10.3 have led to the conclusion, albeit with caveats, that the k-NN method can be successfully used to estimate the amount of v2v TTV. As an application of the ability of the k-NN method to estimate v2v TTV, an experiment was undertaken to model the relationship between the amount of v2v TTV and the mean travel time.
10.3. Modelling vehicle-to-vehicle TTV using the k-NN method

Methodology

Using the data outlined in section 10.3.5 a scatter plot of the mean travel time of a period, \( \mu \), and the standard deviation within that period, \( \sigma \), was drawn. There were a total of 1395 points. Two scatter plots were drawn. The first scatter-plot used the actual values of population mean and standard deviation obtained from the actual ANPR data. The second scatter-plot used data derived from the k-NN model. The estimator of population mean was the k-NN estimate (Lowess LEM, SE-VAR distance metric, \( k = 2508 \)), whilst the estimator of standard deviation was the sample standard deviation of the 400 nearest neighbors (as justified in section 10.3.5).

For each scatter plot the correlation coefficient, \( p \), between the mean and standard deviation was calculated. In addition, two curves were fitted to the data. The form of these equations is given by equations 10.20 and 10.21. The first of these is based on the ARUP model proposed by Black et al. (2004). The second curve is a 2nd order polynomial.

\[
\sigma = \alpha_1 \cdot \mu^{\alpha_2} \quad \text{(10.20)}
\]

\[
\sigma = \beta_0 + \beta_1 \cdot \mu + \beta_2 \cdot \mu^2 \quad \text{(10.21)}
\]

Results

The results of the experiment when the ANPR data was used is presented in figure 10.9. The equation for the ARUP model is given by equation 10.22 and the polynomial is given by equation 10.23 (t-statistics are shown in parentheses).

\[
\begin{align*}
\sigma &= 0.727 \cdot \mu^{0.725} \\
&= (-2.47)(30.89) & R^2 = 0.41
\end{align*}
\]

\[
\begin{align*}
\sigma &= 4.92 + 0.162\mu - 7.18 \times 10^{-5}\mu^2 \\
&= (1.96) \quad (8.06) \quad (-2.19) & R^2 = 0.39
\end{align*}
\]

The results of the experiment when the k-NN data was used is presented in figure 10.10. The equation for the ARUP model is given by equation 10.24 and the polynomial is given by equation 10.25.

\[
\begin{align*}
\sigma &= 0.29 \cdot \mu^{0.909} \\
&= (-12.96)(52.44) & R^2 = 0.66
\end{align*}
\]
Variation of v2v std with mean TT
ANPR data: $p = 0.621$

ARUP: $\sigma = 0.727 \times \mu^{0.725}$ ($R^2 = 0.41$)

Polynomial: $\sigma = 4.92 + 0.162\mu - 7.18 \times 10^{-5}\mu^2$, ($R^2 = 0.39$)

Figure 10.9: Relationship between the mean travel time and the v2v standard deviation of travel times. ANPR data from the Russell Square 2004 dataset was used.
10.3. Modelling vehicle-to-vehicle TTV using the k-NN method

Variation of v2v std with mean TT

- ARUP: $\sigma = 0.29 \times \mu^{0.909}$ ($R^2 = 0.66$)
- Polynomial: $\sigma = 18.5 + 0.0476\mu + 0.000211\mu^2$ ($R^2 = 0.71$)

Figure 10.10: Relationship between the mean travel time and the v2v standard deviation of travel times. k-NN data from the Russell Square 2004 dataset was used.
10.3. Modelling vehicle-to-vehicle TTV using the k-NN method

\[ \sigma = 18.5 + 0.0476\mu + 2.11 \times 10^{-4}\mu^2 \]  
\[ (7.34) \quad (2.65) \quad (6.89) \quad R^2 = 0.71 \]  

Figure 10.9 suggests that as the mean travel time increases so the standard deviation of travel times also increases. The correlation coefficient, \( p \), of 0.621 also points to this. Since all periods chosen for this exercise contained more than 25 records, it is unlikely for the data to have been affected by incidents which affected some vehicles in the period and not others. Thus this source of TTV has been taken into account. There do not appear to be a sufficient number of records with a high mean travel time to determine whether the standard deviation tends towards a limit asymptotically, although the polynomial curve in figure 10.9 suggests that this may be the case.

Figure 10.10 presents the results where k-NN derived data has been used. The relationship between the standard deviation and the mean travel time is seen to be more pronounced when the k-NN data is used. The correlation coefficient using this data is 0.837. This dataset also suggests that the standard deviation does not fall off asymptotically with increasing travel time. However, this could be because of the lack of points at a high mean value of travel time. As has already been noted in figure 8.1, the k-NN estimator tends to underestimate travel time at high actual travel time, resulting in a lack of data-points with a high travel time. This characteristic of the k-NN ULLT model thus limits the ability of k-NN derived data to model all aspects of v2v TTV.

10.3.7 Conclusion: Ability of the k-NN method to model v2v TTV

Section 10.3 has presented research assessing how well v2v TTV can be modelled using data derived from the k-NN model. To achieve this it was first necessary to define some metrics to quantify v2v TTV. These metrics are collectively called the v2v metrics. A methodology which allows the v2v metrics to be estimated from the k-NN method was suggested.

It was shown that the estimates of the v2v metrics using the k-NN method were not extremely accurate if a large number of actual travel time records in the 15-minute time period were available. However, it was found that the k-NN methodology could provide relatively more accurate estimates of the standard deviation of the travel times in the 15-minute time period if fewer than 5 actual travel time records from the same 15-minute time period were available. A Kolmogorov-Smirnov test also showed that the k-NN model was successfully able to model the 15-minute travel time distribution in approximately half of all time periods.

Nevertheless there are problems with using k-NN derived data to model v2v TTV. This was illustrated in section 10.3.6. Although the k-NN method was able to identify the basic monotonic relationship between the standard deviation and mean travel time of any period, it was not able to model the shape of the curve accurately, and in particular the
asymptotic nature of the curve.

10.4 Conclusion: Ability of the k-NN method to model TTV

The primary objective of this chapter was to determine whether the k-NN model developed in part B of this thesis could be used to characterise travel time variability (TTV). It firstly disaggregated TTV into three components: day-to-day (d2d) TTV, period-to-period (p2p) TTV, and vehicle-to-vehicle (v2v) TTV. It was shown with the ANPR data that p2p TTV accounted for the largest variability followed by v2v TTV. However, when data derived from the k-NN method of the form of chapter 8 was used, the v2v TTV was reduced significantly.

Research undertaken in this chapter concluded that the k-NN method is able to model day-to-day (d2d) TTV very well. As an example, analysis of the k-NN derived data identified the same patterns of similar days of the week as was found using the ANPR data.

A modification was made to the k-NN method to allow it to better model v2v TTV. This was achieved by using the sample of 100 nearest neighbours selected by the k-NN method to represent a sample of realisations of travel-time records from the 15-minute period for which the v2v TTV was desired. Results from a Kolmogorov-Smirnov test showed that this modification allowed the k-NN method to successfully model the distribution of travel times in a 15-minute period in approximately half of all cases. It was also able to model metrics of the vehicle-to-vehicle (v2v) TTV to a certain degree of accuracy. Certainly the estimate of v2v TTV is more accurate than that from a small sample of actual measurements.

Further research could be undertaken to determine how well data derived from the k-NN method is able to model the relative change in TTV between two separate scenarios. For example, there may be more v2v TTV on days when there is heavy rain as opposed to days with dry weather. An experiment could be undertaken to see if the relative change in such TTV obtained from the k-NN derived data matches the relative change in TTV obtained from the true travel time records.

Nevertheless the research undertaken within this chapter suggests that the k-NN method could be used to model TTV on links where GPS probe vehicles are the primary source of travel time data. The k-NN method could be used to aggregate travel time records from different times but the same underlying travel time distribution, allowing the ULTT and TTV to be modelled accurately given the current flows and occupancies measured on the link. This has already been discussed in section 8.6. With the proliferation of GPS systems being incorporated into private vehicles, the k-NN method thus appears to be a very attractive way of modelling both urban link travel time, and travel time variability.
Chapter 11

Conclusions and Recommendations for Further Research

This chapter presents the conclusions of this thesis. It also presents recommendations for future research.

11.1 Conclusions

The primary aim of this thesis was to develop a methodology by which data from inductive loop detectors (ILDs) could be used to estimate urban link travel time (ULTT). The secondary aim of the research was to determine how effectively estimates of travel time output by the ULTT model developed within this thesis could be used to model travel time variability (TTV). Both these aims have been successfully achieved.

This thesis was divided in two 3 parts. Part A focused on the data that would be used in a ULTT model. Part B developed a suitable ULTT model. Part C then investigated whether the ULTT model developed in part B could be used to model TTV. A total of 11 objectives for all three sections were identified in chapter 1. These 11 objectives are revisited, and the conclusions from each summarised.

11.1.1 Define clearly the concept of an urban link and the travel time on an urban link.

The first objective stated in the introduction was to clearly define what a link and link travel time was. The definition of the link was defined in section 2.1. A nomenclature for defining each type of link, dependent on where the start and end points were located, was defined. Within this thesis, D-D links were used. That is, the link started and ended at the location of an ANPR camera detector.

Urban link travel time was defined in equation 2.1. It is defined as the mean travel
time of all valid vehicles in a certain time period. A valid vehicle was defined as a vehicle whose sole goal was to traverse over the link as quickly as safely and legally possible.

11.1.2 Develop a methodology to clean matched ANPR data

Data from ANPR cameras was used as the source of travel time data within this thesis. ANPR data was used due to its characteristics of detecting a large proportion of the total vehicle population and ability to identify accurately and precisely the time the vehicles passed the start and end of the link. A source of ANPR data was also made available for this research by Transport for London.

It is desired to include only travel time records from vehicles whose sole goal was to traverse the link as quickly as was safely and legally possible. Other vehicles may stop on the link, or take an alternative route; travel time records from these vehicles were excluded. A filtering method called the Overtaking Rule was proposed (Robinson & Polak 2006). The Overtaking Rule uses the temporal structure of the travel time records. It excludes vehicles which are overtaken and arrive at a certain time after the overtaking vehicle. Both simulation and real-world field experiments have identified the Overtaking Rule as being the best method to filter travel time records. The development of the Overtaking Rule is thus an important contribution to the body of knowledge offered by this research.

11.1.3 Develop a methodology to clean data from ILDs.

ILD data was also made available for this research by Transport for London. Tables 3.1 and 3.2 identified various problems with ILD data. Erroneous data can affect the performance of any model which uses it as input. A comprehensive literature review of ILD data cleaning treatments was provided, and a new multivariate treatment called the Adjacent Detector Test was proposed. The literature has not identified a way of quantifying the performance of an ILD data cleaning treatment. It was thus necessary to identify a way of doing this. This was achieved by measuring the effect of the ILD data cleaning treatment on the performance of the ULTT model. This experiment was undertaken in section 8.5. It was shown that the Daily Statistics Algorithm (DSA) treatment (Chen et al. 2003) outperformed all other treatments. It was also visibly highlighted that the failure to filter erroneous ILD data records from the training data set can severely affect the performance of any ULTT model. This methodology of measuring the performance of ILD cleaning treatment is very pragmatic and represents another contribution of this thesis.

11.1.4 Create an archive to store ILD data obtained from the London SCOOT system.

An objective of this thesis was to create an archive where raw ILD data from the 5000+ SCOOT ILDs could be stored. This archive is called LSAD. Due to reasons beyond the control of both this author and Transport for London this project was delayed and it was necessary to obtain ILD data from the TfL SCOOT ASTRID database. Nevertheless it
is envisaged that LSAD will come on-line by the end of 2005 as it is currently undergoing testing at TfL. LSAD will offer an invaluable resource to future transport researchers as it offers ILD data at temporal and spatial resolutions not available elsewhere in the UK.

11.1.5 **Undertake a comprehensive literature review and critique of ULTT models.**

The aim of part B was to develop an Urban Link Travel Time (ULTT) model to enable ILD data on a link to be used to estimate the travel time over that link. The first objective in this part was thus to undertake a comprehensive literature review and critique of existing ULTT models. Various techniques have been proposed to convert ILD data to estimates of travel time. These range from spot-speed methods to ANN models. The literature has many failings. With the exception of Bajwa et al. (2003b, a) the literature fails to filter the training data of erroneous data (both ILD and travel time records). Much of the literature with the notable exception of Zhang & He (1998), Cherrett et al. (2001), and Xie et al. (2004) also fails to compare each ULTT model with other ULTT models. There is also often a failure, either explicitly or implicitly, to account for the various turning movement delays. These failings have been identified and ways to overcome these determined in the ULTT model developed in part B.

11.1.6 **Develop and optimise a ULTT model based on the k-NN method**

A paper on traffic flow forecasting by Smith & Demetsky (1997) provided the motivation to study the use of the k-NN method as a ULTT model. A comprehensive study of the properties and characteristics of the k-NN method was undertaken. Following from this the key design factors of the k-NN method were identified. These were given as:

- **Number of nearest neighbors, \( k \)** - how many ‘closest’ historic records should be used in the estimation of the travel time.
- **Distance Metric** - how is distance between two points defined in multivariate space.
- **Local Estimation Method** - having identified the \( k \) closest neighbors, what function of the travel time of these records is used to estimate the travel time for the current feature vector.
- **Attributes to include in the feature vector** - what data should be presented as input data to the ULTT model.

A methodology was proposed to identify the optimal combination of these 4 design factors. This is presented in section 11.2. This method was used to optimise the k-NN model for use in the Tottenham Court Road 2003, Russell Square 2003, and Russell Square 2004 datasets. The proposed use of the k-NN method as a ULTT model and the methodology for its optimisation represent a significant contribution of this thesis.
11.1.7 Propose a set of criteria and performance indicators by which ULTT models can be evaluated

In order to determine ‘how good’ a ULTT model is, it is necessary to identify a set of criteria and performance indicators by which ULTT models can be evaluated. A list of desired properties of a ULTT model was presented in table 5.1. The most important criteria by which to assess a ULTT model is its accuracy. However, it is important to ensure that a ULTT model is consistently accurate over its whole operating range. To this author’s knowledge the research presented in this thesis is the first to assess the accuracy of a ULTT model over its whole operating range. Examples of this research are illustrated in figures 8.2 and 8.3. Another property of ULTT models which was assessed within this research was the robustness of the model to changes in both time and input data.

11.1.8 Evaluate the performance of various ULTT models

With the exception of Zhang & He (1998), Cherrett et al. (2001), and Xie et al. (2004) the literature fails to compare a newly proposed ULTT model with other existing ULTT models. The performance of the k-NN ULTT model developed in chapter 7 was thus compared with 8 other ULTT models. For all three datasets, the k-NN model was identified as outperforming all other ULTT models. In particular the k-NN model was shown to outperform other ULTT models at very low and very high actual travel times. This positive characteristic is due to the k-NN using only local data when making its estimation. In effect the k-NN method allows travel time records from the same underlying travel time distribution but different periods of time to be aggregated together to allow more accurate estimates of travel time. Thus the k-NN method offers an opportunity to allow GPS travel time and ILD data to be used to accurately estimate ULTT. The k-NN ULTT model was also shown to be robust to changes over time and input data.

11.1.9 Undertake a comprehensive literature review into TTV.

The aim of part C was to determine whether estimates of travel time obtained from the ULTT model developed in part B could be used to model TTV. The first objective of this part was to undertake a comprehensive literature review into TTV. The major conclusion of this review was that empirical research into TTV had been limited by the lack of travel time data, making it very difficult to control for the different types of temporal scales of travel time variability. The ANPR datasets provided to this research by TfL thus offered the opportunity to control for these various components of TTV.

11.1.10 Identify ways to characterise and quantify TTV.

In order to determine how effectively the estimates of travel time obtained from the ULTT model could model TTV, it was first necessary to identify how TTV could be characterised. This was successfully undertaken in chapter 10. TTV was disaggregated into three primary
components: vehicle-to-vehicle (v2v) TTV, period-to-period (p2p) TTV, and day-to-day (d2d) TTV. Equations for quantifying each component of TTV were given by equations 10.8, 10.14, and 10.15. A set of measures known as the v2v metrics was also proposed to allow v2v TTV to be characterised.

11.1.11 Determine whether travel time estimates obtained from the ULTT model developed on part B allows the TTV to be characterised accurately.

Having identified ways to characterise and quantify TTV, research was then undertaken to determine how effectively the estimates of travel time obtained from the ULTT model could model TTV. In particular work was undertaken to assess how well v2v and d2d TTV could be modelled. Travel time data derived from the k-NN ULTT model was shown to accurately model d2d TTV, allowing for example, similar days-of-the-week to be identified, and the accurate modelling of d2d over the course of the day. In order to model v2v TTV, an adaptation had to be made to the k-NN model. This adaptation used the distribution of travel time records of the k-nearest neighbors to model the v2v TTV. A Kolmogorov-Smirnov test showed that this adaptation allowed the distribution of travel times in any 15-minute period (i.e. v2v TTV) to be modelled correctly in approximately 50% of 15-minute time periods.

11.2 Summary of the methodology for turning data from ILD into estimates of ULTT

The primary objective of this thesis was to develop a methodology for turning data from ILDs on a link into estimates of ULTT over the same link. This has been successfully achieved. This section summarises the methodology for optimising this ULTT model. There are five stages in the methodology, outlined below. The methodology is illustrated in figure 11.1.

- **Stage #1: Filtering of the travel time data**
  The first stage filters the travel time data in the historical database (i.e. the training data), removing erroneous records. This is achieved using the overtaking rule developed in section 2.7.

- **Stage #2: Filtering of the ILD data**
  The second stage filters the ILD data in the historical database (i.e. the training data), removing erroneous records. This is achieved using the Daily Statistics Algorithm outlined in section 3.4.2.

- **Stage #3: k-NN model optimisation - pre-screening**
  The k-NN model is used to turn data from ILDs into estimates of ULTT. There
11.2. Summary of the methodology for turning data from ILD into estimates of ULTT

Figure 11.1: Five-step methodology for turning data from ILDs into estimates of ULTT

are three steps to optimising the k-NN ULTT model. The first step identifies those design parameters which are obviously poorer than the rest.

- **Stage #4: k-NN model optimisation - full-factorial**
The second step of optimising the k-NN ULTT model identifies the optimal combination of distance metric and local estimation method. It also identifies the approximate optimal number of nearest neighbors, \( k \), which should be used.

- **Stage #5: k-NN model optimisation - Fibonacci line search**
The final step in optimising the k-NN ULTT model uses a Fibonacci line search to identify the optimal value of \( k \) for the distance metric and local estimation method combination identified in the previous stage. On completion of this stage, the k-NN model has been optimised for use as a ULTT model.
The optimised ULTT model outlined above has been shown in chapter 8 to provide more accurate estimates of travel time than existing ULTT models. The development of this methodology represents the primary contribution of this research.

11.3 Recommendations for future research

During the course of this research various other areas of research that could be undertaken concerning the development of a methodology to estimate ULTT using ILD data were identified. Likewise, there is much scope for further research into TTV. This last section identifies areas of research that could be undertaken to extend the work outlined in this thesis.

11.3.1 Data Cleaning: future research

This first section proposes further research that could be undertaken to improve the data cleaning processes.

**ANPR cleaning: Further development of the Overtaking Rule**

The Overtaking Rule developed in section 2.7 was shown to outperform existing license plate cleaning methods. There are two areas where this method could be further improved. Firstly, this filter could be made more robust to the presence of overly quick vehicles such as emergency vehicles and buses in dedicated bus lanes. Secondly, as suggested in section 2.7.5, the performance of this method in the off-peak hours could be improved. In off-peak hours there may be very few valid vehicles to facilitate in the identification of erroneous travel time records. It could be assumed that at very low flows all vehicles are travelling at the free-flow travel time. Thus at low flows records from any vehicles with a travel time considerably greater than this free-flow travel time could be excluded.

**ILD cleaning: multivariate tests**

It is the belief of this author that there are many opportunities to improve ILD data cleaning methods. In particular there is much scope to develop cleaning treatments based on multivariate tests. One such multivariate test, the *Adjacent Detectors Test* was proposed within this thesis.

**ILD cleaning: Detection of parked vehicles.**

One important difference between link travel time models developed for urban environments and those for highways, is that there tends not to be parked vehicles on highways. The problem of vehicles parking over ILDs and causing erroneous occupancy data to be output is an area that has to this author’s knowledge not been explored in the literature. For example, on the Russell Square link, there was a hotel (the Russell Hotel) next to ILD
number n02/170e (refer to figure 4.5). Coaches and taxis were occasionally observed to park over this ILD, thus causing erroneous outputs of occupancy data from this ILD.

11.3.2 ULTT models: future research

This section proposes further research that could be undertaken to improve the ULTT model.

Aggregation period of ILD derived data

Within this thesis 15-minute aggregated data from the London SCOOT system was used. It had been envisaged that raw ILD data from the LSAD system would be used within this research. However, due to the reasons outlined in section 4.5 it was not possible to do this. LSAD would have allowed for research to determine the optimal aggregation period of ILD derived data such as flow and occupancy. It will be possible to undertake such work once LSAD is fully operational.

Input data for ULTT model

Due to the difficulty of collecting non ILD-based data, the input data to the model was fixed to the occupancy and flow from a set number of ILDs; 3 for the Russell Square link and 5 for the Tottenham Court Road link. Table 5.4 identified several other data inputs to a ULTT model. Research should be undertaken to identify what are the important data inputs to a ULTT model.

k-NN method - Distance Metrics

Research could be undertaken into identifying distance metrics which provide both increased model accuracy, as well as decreased computation time.

k-NN method - Local Estimation Method

One important contribution of this thesis was the identification of the Lowess local estimation method. To this author’s knowledge the Lowess method, a robust regression model, has not been used within the field of transportation research before. It may be that there is a LEM superior to the Lowess method. Research could be undertaken into this.

k-NN method - computation time

One criticism of the k-NN method is that it is computationally demanding. Research could be undertaken to optimise the speed of the k-NN model.
11.3.3 Travel Time Variability: future research

It was noted in chapter 9 that much existing research into TTV has been limited by the lack of available travel time data. New technologies such as ANPR allow sufficient travel time records to be captured in order to control for the various components of TTV such as v2v TTV. There remains scope to undertake empirical research into many areas of TTV. These include:

- Effect of the weather on TTV.
- The relationship between TTV on a link to TTV over a journey - for example, is the amount of variability on a journey the sum of variability over its constituent links.
- Performance of bus-lanes - do bus lanes reduce the TTV of buses and taxis.

Further research could also be undertaken in determining how well the k-NN method can model TTV. For example, as stated in section 10.4 further research could be undertaken to determine how well data derived from the k-NN method is able to model the relative change in TTV between two separate scenarios. For example, there may be more v2v TTV on days when there is heavy rain as opposed to days with dry weather.

11.4 Real world application

The primary aim of this thesis was to develop a methodology by which data from inductive loop detectors (ILDs) could be used to estimate urban link travel time (ULTT). This aim was successfully achieved. Within this thesis the link was defined as being between two ANPR cameras. Such a link was chosen since ANPR cameras have been shown in section 2.13 to allow the accurate measurement of the true link travel time. On such links there is unlikely to be a need to estimate ULTT using ILD data. However, such a link does allow ULTT models which use ILD data to be assessed and optimised. The two ANPR links used in this thesis have enabled the k-NN method developed in chapter 7 to be optimised and identified as the best ULTT model.

This k-NN ULTT model could be used to estimate ULTT on links that are not defined between two ANPR cameras (Robinson & Polak 2005). As stated in section 8.6, the ULTT model developed in this thesis could be used to aggregate GPS travel time records from different times but the same underlying travel time distribution. This would help to overcome one of the major short-falls of GPS derived travel time estimations, namely the small number of travel time records from suitably equipped GPS probe vehicles in any given 15-minute time period. It is recommended that further research should be undertaken to verify that the k-NN ULTT model developed in this thesis can indeed be used to estimate ULTT using travel time data derived from GPS probe vehicles. Such a ULTT model could help provide travel time data enabling network authorities to measure the performance of their road networks, and study the effects of travel time variability.
This ULTT model could also allow real time estimates of travel time over a link to be made since the only input required would be the current flow and occupancy measured by ILDs on the link in question.
Appendix A

Basic Traffic Theory

This appendix provides a brief introduction to the basics of traffic theory. In particular it gives a brief outline of traffic flow theory and queuing theory. It includes some material that is referred to in the main body of the text. This appendix is not intended to give a comprehensive introduction to traffic theory. Readers who require such a text are advised to refer to books such as TRB (1975), Leutzbach (1988), Daganzo (1997), and Roess et al. (1998).

A.1 Traffic Flow Theory

Traffic flow theory models the movement of vehicles on a link. This section introduces the fundamentals of this topic.

A.1.1 Flow

Flow is defined as the rate per unit time at which units of traffic pass a point. It is thus a point measurement measured over a period of time. It can be defined by equation A.1.

\[ q = \frac{N}{T} \]  
\( q = \text{flow} \)
\( N = \text{Number of vehicles passing over the point in the time period} \ T \)
\( T = \text{Time period over which the measurement is taken} \)

Flow is usually normalised to a period of time of one hour.

At a more general level, flow can be defined by equation A.2.

\[ q(A) = \frac{d(A)}{|A|} \]  
\( d(A) = \text{Total distance in time-space region (veh-km)} \)
\( |A| = \text{‘area’ of time-space region (km-hrs)} \)
Figure A.1: Vehicle trajectories and the intersection with a region of time-space. Adapted from Daganzo (1997).

This is illustrated by Figure A.1, which shows a time-distance plot of several vehicle trajectories. The numerator in the above expression is simply the sum of the distances travelled in the shaded region, A, by the vehicles whose trajectories run through it (in this case vehicles 2, 3, and 4). The denominator is given by the area of the shaded area.

It can be seen that the earlier definition of flow corresponds to the case of rectangle B. Within this thin rectangle, \( N \) vehicles pass in time \( T \). In the case of the above diagram \( N = 2 \) (vehicles 4 and 5).

### A.1.2 Density

Density is defined as the number of units of traffic per unit length of road. It is a measurement over space defined at a single point in time. It can be defined by equation A.3.

\[
    k = \frac{N}{L} \quad \text{(A.3)}
\]

- \( k \) = density
- \( N \) = Number of vehicles in the length of road \( L \) at a particular instant in time.
- \( L \) = Length of road over which measurement is recorded.

At a more general level, density can be defined by equation A.4.

\[
    k(A) = \frac{t(A)}{|A|} \quad \text{(A.4)}
\]

- \( t(A) \) = Total time in time-space region (veh-hrs)
- \( |A| \) = ‘area’ of time-space region (km-hrs)
This is illustrated by figure A.1. It can be seen that the earlier definition of density, recorded at an instant in time corresponds to the case of rectangle C. Within this thin rectangle $N$ vehicles are counted in distance $L$. In the case of the above diagram $N = 4$ (vehicles 1, 2, 3, and 4). Nb. In the case of a photograph, the width of the rectangle, $dt$, is given by the shutter speed - around $1/500$ seconds.

A.1.3 Speed

The concepts of time-mean-speed and space-mean-speed, and their relationship to one another have already been outlined in section 5.4.1.

A.1.4 Relationship between the basic traffic parameters

Conservation of Vehicles

Vehicles do not just appear or disappear into thin air. This important principle of the conservation of vehicles can be used to derive the following fundamental relationship between flow, density, and space mean speed, $v_s$.

$$q = kv_s \quad \text{(A.5)}$$

In addition this principle also leads to the following important equation, which is used in the development of wave propagation theory.

$$\frac{\delta p}{\delta t} + \frac{\delta q}{\delta x} = 0 \quad \text{(A.6)}$$

LWR Model

The Lighthill Whitham Richards, LWR, model is the simplest possible model relating speed to density. It assumes that the speed decreases linearly with density according to equation A.7.

$$v_s = v_0 \left( 1 - \frac{k}{k_j} \right) \quad \text{(A.7)}$$

$v_s$ = space mean speed
$v_0$ = Maximum speed
$k$ = density
$k_j$ = jam density - i.e. density of the vehicles in a traffic jam.

Using the conservation of vehicles, equations relating the speed and density to the flow can also be formulated. The relationships are shown in graphical format in figure A.2.
Figure A.2: Lighthill, Whitham and Richards, LWR model. K = density, V = velocity, Q = flow.

Extension to LWR Model

The LWR is able to model many features observed in the real world such as shock waves. However, the LWR is unable to model the following features (Wong & Wong 2002):

- Discontinuity in the fundamental diagram: in the real world, the congested and uncongested regimes tend to be separated by a discontinuity.

- Hysteresis: The relationship between speed and density differs depending on whether the traffic is becoming more congested or less congested. For more information on hysteresis refer to Zhang (1999).

- Platoon Dispersion: Vehicles travelling in platoons tend to diffuse over time. The LWR model is unable to model this phenomena.

Wong & Wong (2002) suggested that the failure to model these features was due to the assumption in the LWR model of homogenous traffic flow. Wong & Wong (2002) thus extended the LWR model to incorporate heterogenous traffic flow. Using simulated data, this extended LWR model was shown to exhibit the features of traffic flow outlined above.

An alternative approach to overcoming the limitations of the LWR model was proposed by Daganzo (1994) with the ‘cell transmission model’. This model splits the road up into a large number of small cells and then looks at the interaction between adjacent cells. Daganzo (1994) demonstrated that this method could model features such as ‘stop and go’ behaviour which the hydrodynamic based LWR model could not.

There are other criticisms of the LWR model. For example, the LWR model assumes infinite acceleration and deceleration (Papageorgiou 1998). Papageorgiou (1998) also notes the limited number of validation results for macroscopic traffic-flow models.

Car Following Models

Other models relating density with speed have been developed using car-following models. The stimulus in these models tends to be the difference of speeds between two adjacent vehicles. The sensitivity tends to be a function of the distance between the two vehicles.
and the speed of vehicle. The general formula for car following models is given by equation A.8.

\[ \ddot{x}_n = \frac{[\dot{x}_n(t + T)]^m}{[x_{n-1}(t) - x_n(t)]^l} [\dot{x}_{n-1}(t) - \dot{x}_n(t)] \]  

(A.8)

\( \ddot{x}_n \) = Acceleration of vehicle \( n \)

\( \dot{x}_n(t + T) \) = Speed of vehicle \( n \)

\( x_{n-1}(t) - x_n(t) \) = Distance between vehicle \( n \) and the next vehicle

\( \dot{x}_{n-1}(t) - \dot{x}_n(t) \) = Speed difference between vehicle \( n \) and the next vehicle

\( m, l \) = parameters to be chosen

### A.2 Queuing Theory

This section is not meant to be an exhaustive introduction to queuing theory. Instead, only aspects of queuing theory that are relevant for the PhD are given. For a thorough introduction to queuing theory, the reader is referred to Newell (1982), and Roess et al. (1998).

#### A.2.1 Definition of Traffic Intensity

In queuing theory the ratio of the vehicles entering the junction, \( v \), to the maximum number of vehicles that can be serviced by the junction, \( c \), is a key variable, and is known as the traffic intensity. It is formalised by equation A.9.

\[ \rho = \frac{v}{c} \]  

(A.9)

\( \rho \) = traffic intensity. This value is also known as the degree of saturation.

\( v \) = arrival rate

\( c \) = capacity rate

#### A.2.2 Flow below saturation

In the simplest case consider a single vehicle driving on a road with one traffic light. The time that the vehicle arrives is random but has a uniform distribution. The delay to the vehicle will depend on the time the vehicle arrived at the junction and is given by figure A.3. The pdf of the time taken through the link will look like figure A.3. The distribution of delay is described by equation A.10:

\[ f(t) = \begin{cases} 0 & 0 < t \leq sC \\ C - t & sC < t \leq sC \end{cases} \]  

(A.10)
The expected delay, $E[f(t)]$ is given by the equation A.11.

$$E[f(t)] = \frac{1}{C} \int_{t=0}^{C} f(t) = \frac{C(1-s)^2}{2} \quad (A.11)$$

$C = \text{cycle time}$
$s = \text{split - the fraction of green time in each cycle}$

This equation can be re-written in terms of the cycle time and the effective red time, $R$.

$$E[f(t)] = \frac{R^2}{2C} \quad (A.12)$$

This equation for expected delay assumes that vehicles have infinite acceleration and deceleration. The expected variance of delay is given by equation A.13.

$$Var(x) = \frac{r^3}{3} \left( \frac{1}{3} - \frac{R}{4C} \right) \quad (A.13)$$

The above work relies on the assumption that traffic flow is negligible when compared to the saturation flow of the junction. Webster (Roess et al. 1998) modeled the queue when flow was not negligible when compared to the flow on the junction, but the traffic intensity was still below 1. His theory is often termed steady state theory. Using the figure A.4 he was able to derive the total delay in a cycle in under-saturated traffic. This is represented by the shaded area in figure A.4. The average delay per vehicle, $UD$, is given by equation A.14.

$$UD = \frac{C\left[1 - \frac{G}{C}\right]^2}{2\left[1 - \frac{z}{s}\right]} \quad (A.14)$$

$C = \text{cycle time (secs)}$
$G = \text{green time (secs)}$
$v = \text{arrival flow rate (vehicles / sec)}$
$s = \text{saturation rate (vehicles / sec)}$
A.2. Queuing Theory

Equation A.14 assumed uniform arrival times, and is called the Uniform Delay. If the flow is made negligible to the saturation flow, i.e. \( v/s \) is set to zero, then equation A.14 simplifies to equation A.11 albeit with different notations. However, it is more accurate to model the arrivals as being Poisson in nature. Thus it is possible that several cars may arrive together just before the light turns red, forcing some of these cars to wait almost a full cycle time in order to get the next green stage. This delay is called the random delay. Webster modelled this by equation A.15.

\[
UD = \frac{c}{2} \left[ \frac{1 - \frac{c}{v}}{1 - \frac{s}{v}} \right]^2 + \frac{(\frac{v}{c})^2}{2v} - 0.65 \left( \frac{c}{v^2} \right)^{\frac{1}{3}} \left( \frac{v}{c} \right)^{\frac{2+\frac{a}{c}}{3}}
\]  

(A.15)

c = capacity of the system (vehicles / sec)

The first term is simply the uniform delay. The second term accounts for the randomness of arrivals due to a Poisson distribution being assumed - the random delay. The final term is a correction term based on simulation.

### A.2.3 Oversaturated flow

When the traffic intensity is greater than 1, Webster’s equation breaks down. In this traffic state the following theory, termed deterministic theory, is used.

From Figure A.5 the average overflow delay, \( OD \), in a time period \( T \) can be determined. It is given by equation A.16.

\[
OD = \frac{T}{2} \left[ \left( \frac{v}{c} \right) - 1s \right]
\]  

(A.16)

\( T \) = time period over which average time delay is calculated
A.2. Queuing Theory

Care must be taken when using this formula due to the following reasons:

- The above formula gives the average overflow delay in time $T$. Clearly from the diagram it will overestimate the delay to vehicles at the start of this period and underestimate the delay for these vehicles at the end of this period.

- It does not include the delay that occurs after time $T$ to vehicles in the queue at time $T$.

A.2.4 Kimber and Hollis model

The two models given for under-saturated and over-saturated flow do not converge together. Indeed when the traffic intensity is 1, then Webster’s model predicts an infinite delay whereas the model for oversaturated flow predicts a delay of 0. Webster’s model is found only to be satisfactory when the traffic intensity is below 0.85, whilst the oversaturated model gives good results only at traffic intensities above 1.15.

Most UTC systems try to keep the traffic intensity of the critical approach at around 0.9, and hence traffic intensities of 0.8 - 1.2 are very common in practice and thus important to model. (Robertson 1969; Hunt et al. 1981) To overcome the deficiency of the two models, Kimber and Hollis proposed a coordinate transformation method, whereby the two models for under and oversaturated models were blended together (Kimber & Hollis 1979).

The way the models are merged together is best demonstrated by figure A.6. Delay predicted by the under saturated model is given by the line $E1$, whilst the delay predicted by the over saturated model is given by the line $E2$. The Kimber and Hollis model is constructed such that the length $L1$ (the distance between the under saturated model and...
Figure A.6: Derivation of Kimber and Hollis model

\[ D = \frac{(F^2 + G)^{1/2} - F}{2} \]

\[ F = \frac{(Q - q)Qt^2 + 2(1 - L_0)Qt - 4(1 - C)(2L_0 + qt)}{2[Qt + 2(1 - C)]} \]

\[ G = \frac{2(2L_0 + qt)[Qt - (1 - C)(2L_0 + qt)]}{Qt + 2(1 - C)} \]  

(A.17)

\[ D = \text{Delay} \]

\[ C = \text{Constant depending on junction type. For a roundabout and priority junction } C = 1, \text{ for a stream at a fixed time traffic signal } C = 0.55. \]

\[ L_0 = \text{Initial length of the queue (vehicles)} \]

\[ Q = \text{Capacity (vehicle / sec)} \]

\[ q = \text{Arrival rate (vehicle / sec)} \]
A.2.5 General Queuing systems

Much of queuing theory was developed in the context of telecommunications - notably with modelling the operation of a telephone system. Many queues can be disaggregated into the following components:

- Arrival Stream - contains the inter-arrival time distribution.
- Service Pattern - number of servers, service time distribution
- Buffer arrangements - Number of buffers, how are they connected. Buffer sizes
- Queue discipline - rules for selecting the next item for service.

This is illustrated by figure A.7.

The common nomenclature used to describe the system is:

\[ A/S/K/N/QD \]

- \( A \) = Type of arrival-time distribution
- \( S \) = Type of service-time distribution
- \( K \) = No of Servers
- \( N \) = System capacity
- \( QD \) = Queue Discipline

For example, the queue \( M/M/1 \) has Poisson distributed arrivals, a Poisson distributed service time, and one server. The queue discipline is assumed to be FIFO - first in first out.
Appendix B

Platoons

This appendix provides further material concerning platoons that is referred to within the thesis. The first section describes the limitations of platoon matching to estimate link travel time. The second section explains how platooning effects can increase the amount of mid-link delay and hence increase ULTT.

B.1 Limitations of platoon matching

The platoon correlation matching techniques outlined in section 5.5 rely on the structure of the platoon being the same or at least highly similar at the upstream and downstream detector. This is explicitly mentioned by Dailey (1993) who notes that ‘the validity of the technique depends on the commonality of information in the signals, which implies rigid propagation of the traffic microstructure.’ He also states that residual terms in the cross-correlation function should be significantly less than the autocorrelation term of the cross-correlation function. Dailey found that at average occupancies greater than 15%, these residual terms increased relative to the auto-correlation term, although Dailey’s model does not suggest why this should be the case.

There seems to be little work in the transport literature about the limitations of such cross-correlation techniques and the causes of this microstructure changing. Incidents are likely to cause the microstructure to change, since incidents cause bottlenecks to occur. The bottleneck has limited capacity, and if a platoon with instantaneous flow rate exceeds the capacity of the bottleneck, then the platoon will have to be ‘flattened’ to allow all the vehicles to pass though the bottleneck. This is illustrated in figure B.1. It is thus possible to derive a constraint on the maximum occupancy before two separate platoons merge into one platoon - i.e. altering the traffic microstructure.

In figure B.1 there are 4 platoons of vehicles, A, B, C, and D, each with the same number of vehicles. The velocity of each vehicle before and after the incident is the same, \( v \). The period for each new platoon is \( T \) seconds. In platoons A and B, vehicles are close together, and this is borne out by the high flow rate, \( q_b \). Since the flow is high, the time length of the platoon is small, only \( t_b \) seconds. At a certain point in the link there
is an incident, which acts as a bottleneck, allowing only $q^a$ vehicles to flow. This forces the vehicles in the platoon to queue temporally until they can be serviced through the bottleneck. This increases the length of each platoon to $t^a$ seconds.

The number of vehicles in platoon $A$ is the same as the number of vehicles in platoon $D$. Since the flow is defined by the equation below,

$$\text{flow} = \frac{\text{no. of vehicles}}{\text{time}}$$

(B.1)

By the conservation of vehicles, the area of platoon $A$ must equal the area of platoon $D$:

$$q^b t^b = q^a t^a$$

(B.2)

$q^i = \text{Flow at point } i.$  
$t^i = \text{Length of platoon (in seconds)}$  
$a = \text{traffic characteristics after the incident.}$  
$b = \text{traffic characteristics before the incident.}$

Occupancy has been defined in the main paper in equation 3.2. The average occupancy, $o^b$, over one time period, $T$, is given by equation B.3.

$$o^b = \left( \frac{q^b}{v g} \right) \frac{t^b}{T}$$

(B.3)

$o^b = \text{occupancy before the incident}$  
$v = \text{velocity of the vehicles in the platoon.}$  
$g = \text{constant relating occupancy to density.}$  
$T = \text{Period between the start of two adjacent platoons.}$

In order for the two platoons to exist after the incident (i.e. not merged together) the
B.2 The effect of platooning on mid-link delay

B.2.1 Description of problem

On highways, conditions along the link tend to be homogenous - that is the velocity of vehicles travelling over a stretch of highway tend not to change. However in the urban context this is not true. In urban environments there are many more traffic features which can cause vehicles to slow down. For example, parked cars on the side of a road may force vehicles to slow down. Vehicles making a turn mid-link will also cause vehicles behind it to slow down. If there are several vehicles immediately behind a vehicle turning right, (i.e. when vehicles are travelling in a platoon) then all those vehicles will be delayed by the vehicle turning right. However if vehicles are spaced far apart from one another, then the following vehicles will not be affected as greatly by the vehicle turning right. This study researched how the time gaps between vehicles in a platoon affected the travel time and travel time variability.

B.2.2 Mathematical Model

This section develops a mathematical model of the above problem to predict the travel time for vehicles travelling in a platoon separated by a time-gap, \( g_i \), where there is a chance that some of the vehicles will make a right-turn mid link. This turning movement

\[
t_a < T 
\]

(B.4)

Combining equations B.2 and B.3, gives the following expression for \( t_a \).

\[
t_a = \frac{T o_b v g}{q_a} 
\]

(B.5)

Substituting this into B.4 leaves the following condition on the occupancy before the incident in order for the microstructure to remain the same.

\[
o_b < \frac{q_a}{v g} 
\]

(B.6)

This analysis assumes that all vehicles have the same performance characteristics (i.e. acceleration, length) and choose to travel at the same velocity.

This equation shows that as the effects of the incident get worse (i.e. the flow through the bottleneck, \( q_a \), is reduced), so the occupancy at which the microstructure breaks decreases. More importantly it suggests that even if there is only a minor incident on a road, the structure of the traffic may not remain constant, and the cross-correlation technique is unlikely to be able to determine the travel time over the link.
B.2. The effect of platooning on mid-link delay

It takes a certain amount of time, $m_i$, which may cause vehicles upstream to be delayed. Consider the figure B.2.

![Figure B.2: Vehicles which make a mid-link movement are delayed for $m_i$ seconds when making that movement.](image)

Vehicle 2 is delayed if the time taken for Vehicle 1 to make the turning movement, $m_1$, is greater than the time-gap between the two vehicles, $g_1$. The delay of vehicle 2 can thus be written by equation B.7.

$$d_2 = m_1 - g_1 \quad (B.7)$$

Vehicle 3 may be delayed by either vehicle 2 or vehicle 1. Thus vehicle 3 is delayed if the time taken for vehicle 2 to turn, $m_2$, plus any delay vehicle 2 experiences from vehicle 1, $d_2$, is greater than the time-gap between the vehicles 2 and 3, $g_2$. The delay of vehicle 3 can thus be written by the following equation:

$$d_3 = d_2 + m_2 - g_2 \quad (B.8)$$

Thus there is an iterative relationship between the delays of two adjacent vehicles. This is expressed below:

$$d_{i+1} = d_i + m_i - g_i \quad (B.9)$$

However, this formula suggests that there can be a negative delay if the time gap, $g_i$, is suitably large. This can never happen. The following constraint applies to the delay, $d_i$.

$$d_i \geq 0 \quad (B.10)$$

Thus equation B.9 should be modified as below:

$$d_{i+1} = max(d_i + m_i - g_i, 0) \quad (B.11)$$

Equation B.11 clearly shows that if there is a smaller gap between the vehicles, then the mid-link delay will increase. In platoons, the gap between vehicles is small, and hence mid-link delay will be greater for vehicles in platoons than vehicle dispersed equally from one another.
Appendix C

Day-to-day travel times for Russell Square 2004 dataset

Figure A3.1 shows a comparison of the ULTT measured directly by the ANPR cameras and derived from the k-NN model for the Russell Square 2004 dataset. Travel time data for each 15-minute period is shown for the 91 days between 1st April 2004 and 30th June 2004. The k-NN estimates are seen to closely match those of the actual data.
Figure C.1: Comparison of the estimates of ULTT using the actual ANPR measurements and the k-NN derived estimates. Data from the Russell Square 2004 dataset, weeks 1 - 4.
Figure C.2: Comparison of the estimates of ULTT using the actual ANPR measurements and the k-NN derived estimates. Data from the Russell Square 2004 dataset, weeks 5 - 8.
Figure C.3: Comparison of the estimates of ULTT using the actual ANPR measurements and the k-NN derived estimates. Data from the Russell Square 2004 dataset, weeks 9 - 12.
Figure C.4: Comparison of the estimates of ULTT using the actual ANPR measurements and the k-NN derived estimates. Data from the Russell Square 2004 dataset, weeks 13 - 14.
Appendix D

Characterisation of data sets

This chapter provides further illustrations and tables which characterise the three primary data sets used within this thesis. For each data set the following statistics are shown:

- General Details
- Cumulative Distribution Function of Travel Time
- Probability Distribution Function of the flows and occupancies from all ILDs
- Disaggregation of records by day-of-week
- Disaggregation of records by time-of-day

D.1 Russell Square 2003

This section describes the Russell Square 2003 data set. Data was collected for 37 days between Monday 3rd March and Sunday 13th April 2003. After filtering the data using the overtaking rule and DSA, this data set contained a total of 24,718 vehicle travel time records. Each vehicle travel time record contained the following attributes:

- Link identifier - i.e. Russell Square 2003
- Timestamp at which vehicle crossed first ANPR site.
- Vehicle type (car, bus, etc...)
- Travel time of vehicle - i.e. time taken to travel between the two ANPR sites.
- 15-minute aggregated flow across all ILDs on the link (time centred around the time the vehicle crossed first ANPR site).
- 15-minute aggregated occupancy across all ILDs on the link (time centred around the time the vehicle crossed first ANPR site).
D.1.1 CDF of Travel Time

The Cumulative Distribution Function of Travel Time for the Russell Square 2003 data-set is shown in figure D.1.

![CDF of Travel Time](image)

Figure D.1: cdf of travel times - Russell Square 2003 dataset - see section 4.7

A break down of travel time by time-of-day is shown in figure D.2. For each 15-minute time-period this shows the 5th percentile, median, and 95th percentile travel time. One interesting feature is the high 95th percentile travel time in the early hours of the morning. This is probably due to a failure of the Overtaking Rule to filter some invalid travel time records due to the smaller number of following vehicles at night.

D.1.2 pdf of flows and occupancies

On the Russell Square 2003 link the following ILDs were used:

- n02/063e1
- n02/253e1
- n02/055e1

Figure D.3 shows the probability distribution function (pdf) of the flows and occupancies of these ILDs for all records in this dataset.
D.1.3 Records by Day-of-Week

Table D.1 shows how many records came from each day-of-week in the Russell Square 2003 dataset.

<table>
<thead>
<tr>
<th>day-of-week</th>
<th>no of records</th>
<th>% records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>2096</td>
<td>8.48</td>
</tr>
<tr>
<td>Monday</td>
<td>3335</td>
<td>13.5</td>
</tr>
<tr>
<td>Tuesday</td>
<td>2826</td>
<td>11.4</td>
</tr>
<tr>
<td>Wednesday</td>
<td>2592</td>
<td>10.5</td>
</tr>
<tr>
<td>Thursday</td>
<td>4703</td>
<td>19</td>
</tr>
<tr>
<td>Friday</td>
<td>6113</td>
<td>24.7</td>
</tr>
<tr>
<td>Saturday</td>
<td>3053</td>
<td>12.4</td>
</tr>
</tbody>
</table>

Table D.1: No records for each day-of-week, Russell Square 2003 dataset

D.1.4 Records by Time-of-Day

Figure D.4 shows how many records came from each 15-minute period in the Russell Square 2003 dataset. From this figure, one important feature can be noticed. This is that the last 15-minute time-period of the day, from 23:45:00 to 23:59:59 has significantly fewer records than other periods. This feature is caused by the anonymising of the licence plate data. Each license plate is mapped onto an anonymous number in-order to meet
the requirements of data-protection. When the same license plate is read, the vehicle is given the same anonymous number. However, this mapping is refreshed at midnight every day. Thus it will not be possible to match vehicles which pass the first ANPR before midnight and the second ANPR camera after midnight. This explains why there are so few travel-time records for this 15-minute period.
Figure D.3: cdf of travel times - Russell Square 2003 dataset - see section 4.7
Figure D.4: No of records by time-of-day: - Russell Square 2003 dataset - see section 4.7
D.2 Russell Square 2004

This section describes the Russell Square 2004 data-set. Data was collected for 91 days between Thursday 1st April 2004 and Wednesday 30th June 2004. After filtering the data using the overtaking rule and DSA, this data-set contained a total of 98,973 vehicle travel time records. Each vehicle travel time record contained the same attributes as for the Russell Square 2003 dataset.

D.2.1 CDF of Travel Time

The Cumulative Distribution Function of Travel Time for the Russell Square 2004 data-set is shown in figure D.5.

![CDF of travel times for the Russell Square 2004 dataset](image)

Figure D.5: cdf of travel times - Russell Square 2004 dataset - see section 4.7

A break down of travel time by time-of-day is shown in figure D.6. For each 15-minute time-period this shows the 5th percentile, median, and 95th percentile travel time.

D.2.2 pdf of flows and occupancies

On the Russell Square 2004 link the following ILDs were used:

- n02/052a1
- n02/253e1
Figure D.6: Travel Time by time-of-day: Russell Square 2004 dataset

- n02/055e1

Figure D.7 shows the probability distribution function (pdf) of the flows and occupancies of these ILDs for all records in this dataset.

D.2.3 Records by Day-of-Week

Table D.2 shows how many records came from each day-of-week in the Russell Square 2004 dataset.

<table>
<thead>
<tr>
<th>day-of-week</th>
<th>no of records</th>
<th>% records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>8894</td>
<td>8.99</td>
</tr>
<tr>
<td>Monday</td>
<td>12370</td>
<td>12.5</td>
</tr>
<tr>
<td>Tuesday</td>
<td>16082</td>
<td>16.2</td>
</tr>
<tr>
<td>Wednesday</td>
<td>15736</td>
<td>15.9</td>
</tr>
<tr>
<td>Thursday</td>
<td>13472</td>
<td>13.6</td>
</tr>
<tr>
<td>Friday</td>
<td>13982</td>
<td>14.1</td>
</tr>
<tr>
<td>Saturday</td>
<td>18437</td>
<td>18.6</td>
</tr>
</tbody>
</table>

Table D.2: No records for each day-of-week, Russell Square 2004 dataset
D.2.4 Records by Time-of-Day

Figure D.8 shows how many records came from each 15-minute period in the Russell Square 2004 dataset.
Figure D.7: cdf of travel times - Russell Square 2004 dataset - see section 4.7
No records by time-of-day: Russell Square 2004. Total Records = 98973

Figure D.8: No of records by time-of-day: - Russell Square 2004 dataset - see section 4.7
D.3 Tottenham Court Road 2003

This section describes the Tottenham Court Road 2003 data-set. Data was collected for 37 days between Monday 3rd March and Sunday 13th April 2003. After filtering the data using the overtaking rule and DSA, this data-set contained a total of 104,810 vehicle travel time records. Each vehicle travel time record contained the same attributes as for the Russell Square 2003 dataset.

D.3.1 CDF of Travel Time

The Cumulative Distribution Function of Travel Time for the Tottenham Court Road 2003 data-set is shown in figure D.9.

![CDF of Travel Time](image)

Figure D.9: cdf of travel times - Tottenham Court Road 2003 dataset - see section 4.7

A break down of travel time by time-of-day is shown in figure D.10. For each 15-minute time-period this shows the 5th percentile, median, and 95th percentile travel time.

D.3.2 pdf of flows and occupancies

On the Tottenham Court Road 2003 link the following ILDs were used:

- n02/057c1
- n02/056g1
Figure D.10: Travel Time by time-of-day: Tottenham Court Road 2003

- n02/060g1
- n02/059c1
- n02/019v1

Figures D.11 and D.12 show the probability distribution function (pdf) of the flows and occupancies of these ILDs for all records in this dataset.

D.3.3 Records by Day-of-Week

Table D.3 shows how many records came from each day-of-week in the Tottenham Court Road 2003 dataset.

<table>
<thead>
<tr>
<th>day-of-week</th>
<th>no of records</th>
<th>% records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>7914</td>
<td>7.55</td>
</tr>
<tr>
<td>Monday</td>
<td>13863</td>
<td>13.2</td>
</tr>
<tr>
<td>Tuesday</td>
<td>14751</td>
<td>14.1</td>
</tr>
<tr>
<td>Wednesday</td>
<td>15014</td>
<td>14.3</td>
</tr>
<tr>
<td>Thursday</td>
<td>19306</td>
<td>18.4</td>
</tr>
<tr>
<td>Friday</td>
<td>20439</td>
<td>19.5</td>
</tr>
<tr>
<td>Saturday</td>
<td>13523</td>
<td>12.9</td>
</tr>
</tbody>
</table>

Table D.3: No records for each day-of-week, Tottenham Court Road 2003 dataset
Figure D.11: cdf of travel times - Tottenham Court Road 2003 dataset - part a

D.3.4 Records by Time-of-Day

Figure D.13 shows how many records came from each 15-minute period in the Tottenham Court Road 2003 dataset.
Figure D.12: cdf of travel times - Tottenham Court Road 2003 dataset - part b
Figure D.13: No of records by time-of-day: Tottenham Court Road 2003 dataset
Appendix E

List of Acronyms

Table E.1 presents a list of the acronyms used within this thesis.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full text</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADT</td>
<td>Annual Average Daily Traffic</td>
</tr>
<tr>
<td>ALOTPV</td>
<td>Average Loop Occupancy Time Per Vehicle</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>ANPR</td>
<td>Automatic Number Plate Recognition</td>
</tr>
<tr>
<td>ARIMA</td>
<td>Auto Regressive Integrated Moving Average</td>
</tr>
<tr>
<td>ASTRID</td>
<td>Automatic SCOOT Traffic Information Database</td>
</tr>
<tr>
<td>CoV</td>
<td>Coefficient of Variation</td>
</tr>
<tr>
<td>d2d</td>
<td>day to day</td>
</tr>
<tr>
<td>DB</td>
<td>Database</td>
</tr>
<tr>
<td>DM</td>
<td>Distance Metric</td>
</tr>
<tr>
<td>DOP</td>
<td>Dilution of Precision</td>
</tr>
<tr>
<td>DSA</td>
<td>Daily Statistics Algorithm</td>
</tr>
<tr>
<td>DVLA</td>
<td>Driver and Vehicle Licensing Agency, UK</td>
</tr>
<tr>
<td>FNN</td>
<td>Fuzzy Neural Network</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile Communications</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
</tr>
<tr>
<td>HEFCE</td>
<td>Higher Education Funding Council for England</td>
</tr>
<tr>
<td>HGV</td>
<td>Heavy Goods Vehicle</td>
</tr>
<tr>
<td>ILD</td>
<td>Inductive Loop Detector</td>
</tr>
<tr>
<td>INCA</td>
<td>INcident Cost benefit Analysis</td>
</tr>
<tr>
<td>k-NN</td>
<td>k Nearest Neighbor</td>
</tr>
<tr>
<td>LEM</td>
<td>Local Estimation Method</td>
</tr>
<tr>
<td>LF</td>
<td>Low Frequency</td>
</tr>
<tr>
<td>LSAD</td>
<td>London SCOOT Archive Database</td>
</tr>
<tr>
<td>LTT</td>
<td>Link Travel Time</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>MAD</td>
<td>Mean Absolute Deviation</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MAPE</td>
<td>Mean Absolute Percentage Error</td>
</tr>
<tr>
<td>MCO</td>
<td>Moving Car Observer</td>
</tr>
<tr>
<td>MIDAS</td>
<td>Motorway Incident Detection and Automatic Signalling</td>
</tr>
<tr>
<td>MNN</td>
<td>Modular Neural Network</td>
</tr>
<tr>
<td>MPE</td>
<td>Mean Percentage Error</td>
</tr>
<tr>
<td>MRE</td>
<td>Maximum Relative Error</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>MSULTT</td>
<td>Micro-Structure Urban Link Travel Time</td>
</tr>
<tr>
<td>MVO</td>
<td>Moving Vehicle Observer</td>
</tr>
<tr>
<td>OD</td>
<td>Origin-Destination</td>
</tr>
<tr>
<td>OR</td>
<td>Overtaking Rule</td>
</tr>
<tr>
<td>OTU</td>
<td>Outstation Transmission Unit</td>
</tr>
<tr>
<td>p2p</td>
<td>period to period</td>
</tr>
<tr>
<td>PATH</td>
<td>Partners for Advanced Transit and Highways</td>
</tr>
<tr>
<td>PeMS</td>
<td>Performance Measurement System</td>
</tr>
<tr>
<td>RHS</td>
<td>Right Hand Side</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>ROMANSE</td>
<td>Road Management System for Europe</td>
</tr>
<tr>
<td>RRSE</td>
<td>Root Relative Square Error</td>
</tr>
<tr>
<td>SARIMA</td>
<td>Seasonal ARIMA</td>
</tr>
<tr>
<td>SCOOT</td>
<td>Split, Cycle, and Offset Optimisation Technique</td>
</tr>
<tr>
<td>SMS</td>
<td>Space Mean Speed</td>
</tr>
<tr>
<td>SNN</td>
<td>Spectral basis Neural Network</td>
</tr>
<tr>
<td>STD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>Tfl</td>
<td>Transport for London</td>
</tr>
<tr>
<td>TMC</td>
<td>Traffic Management Centre</td>
</tr>
<tr>
<td>TMS</td>
<td>Time Mean Speed</td>
</tr>
<tr>
<td>TTV</td>
<td>Travel Time Variability</td>
</tr>
<tr>
<td>UAV</td>
<td>Unmanned Aerial Vehicles</td>
</tr>
<tr>
<td>ULTT</td>
<td>Urban Link Travel Time</td>
</tr>
<tr>
<td>UTC</td>
<td>Urban Traffic Control</td>
</tr>
<tr>
<td>v2v</td>
<td>vehicle to vehicle</td>
</tr>
<tr>
<td>VIP</td>
<td>Video Image Processing</td>
</tr>
<tr>
<td>WSDOT</td>
<td>Washington State Department of Transport</td>
</tr>
</tbody>
</table>

Table E.1: List of acronyms
References


Freeman Fox and Associates (1972), ‘Speed flow relationships on suburban main roads’.


HETA, A. N. (1999), ‘The Appraisal of Measures which aim to reduce Travel Time Variability - (TAMTTV)’, DTLR.


Martin, P. T. & Wu, P. (2003), ‘Automated data collection, analysis, and archival’, *Mountain-Plains Consortium, North Dakota State University, P.O. Box 5074 Fargo, ND 58105- USA*.


Mott MacDonald (2000a), ‘Travel Time Variability Follow on research 1.1’.


Park, D. J. & Rilett, L. R. (1998), ‘Forecasting freeway link travel times with a feedforward multilayer neural network’, *Transportation Research Record* (1617), 63–70.


