AECOM

Deriving Journey Patterns from Smartcard Data

Algorithm Development

9 May 2017

Quality information

Prepared by	Checked by	Approved by
Victor Sequi	Andrew Currall	Andrew Currall

Revision History

Revision	Revision date	Details	Authorised	Name	Position
0.9	14 August 2017	Completed Draft for UTG review	Yes	Andrew Currall	Principal Consultant
1.0	10 October 2017	Revised following UTG comments	Yes	Andrew Currall	Principal Consultant
2.0	22 October 2018	Updated in draft following Ticketer data work	No		
2.3	4 December 2018	Initial full release of post-Ticketer-analysis document	Yes	Andrew Currall	Principal Consultant

Prepared for:

Clare Linton
Urban Transport Group

Prepared by:

Victor Sequi
T: 01727 535612

AECOM Limited
AECOM House
63-77 Victoria Street
St Albans
Hertfordshire
AL1 3ER
UK

T: +44(0)1727 535000
aecom.com

in the document. No third party may rely upon this document without the prior and express written agreement of AECOM.

Table of Contents

1. Introduction 6
1.1 Context 6
1.2 Scope of Study 6
1.3 Report Structure 7
2. Literature Review 8
2.1 TfL Study 8
2.2 Transport Modelling Experience 8
3. Study Datasets 9
3.1 Tyne & Wear Metro Data 9
3.2 Tyne & Wear Bus Data 10
3.3 Liverpool City Region Data 11
3.4 Ticketer data (West Yorkshire) 11
4. Alighting Estimation- Concepts 13
4.1 Trip-based Alighting Estimation 13
4.2 Traveller-based Alighting Estimation 13
4.3 Full Dataset Alighting Estimation 13
5. Alighting Point Estimation- Nexus Metro Data 15
5.1 Methodology 15
5.2 Record-Level Accuracy 15
5.3 Trip Lengths 19
5.4 Conclusion 21
6. Geographic Matching- Nexus Bus Data 23
6.1 Context 23
6.2 Outline of Approach 23
6.3 Application 24
6.4 Algorithm 26
6.5 Possible Further Research 27
7. Alighting Point Estimation- Nexus Bus Data 28
7.1 ENCTS Data Full Data set 28
7.2 Trip lengths 30
7.3 Algorithm 35
8. Geographic Location for Ticketer data 36
9. Alighting Estimation- Ticketer data 37
9.1 Ticketer West Yorkshire data 37
9.2 Trip lengths 39
9.3 Algorithm 43
10. Study Conclusions 45
10.1 Tyne & Wear, Nexus Data 45
10.2 West Yorkshire – Ticketer data 45
10.3 Summary of Datasets Studied 46

Figures

Figure 3-1 Comparison between Nexus and Ticketer on the number of trips made on a day by a unique smartcard 12
Figure 7-1 Distance profile Survey, Algorithm and NTS. 0-64km (May+June 2016) 31
Figure 7-2 Distance profile Survey, Algorithm and NTS. 0-24km (May+June 2016) 31
Figure 7-3 Distance profile Survey, Algorithm and NTS, By Method (May+June 2016) 32
Figure 7-4 Distance profile Survey and Algorithm for Operator Group 1 buses (First Week of May) 33
Figure 7-5 Distance profile Survey and Algorithm for Operator Group 2 buses (First Week of May) 33

Figure 7-6 Distance profile Survey and Algorithm for Operator Group 3 buses (First Week of May) 34
Figure 7-7 Distance profile Survey and Algorithm for Operator Group 4 buses (First Week of May) 34
Figure 8-1 : Number of bus stops variability to threshold chosen. 36
Figure 9-1: Box and whisker plot of walked distance for methods 1 and 2 when the threshold was set to 4km 38
Figure 9-2: Distance profile for the algorithm by method. 0-24km 40
Figure 9-3: Distance profile Algorithm and NTS. 0-24km 41
Figure 9-4: Distance profile Algorithm and NTS. 0-24 miles 42
Figure 9-5: Distance profile Ticketer Algorithm, Nexus Algorithm and NTS. 0-24km 43

Tables

Table 3-1: Metro data set 1, Trips by total number of trips the user made on that day 9
Table 3-2: Metro data set 2, Trips by total number of trips the user made on that day 9
Table 3-3: Tyne & Wear Bus Data #71, Trips by total number of trips the user made on that day 10
Table 3-4: Tyne & Wear Bus Data #010215, Trips by total number of trips the user made on that day 10
Table 3-5: Liverpool City Region Data 11
Table 3-6: Boarding_data_set, Number of instances in which a unique Smartcard made X swipes on a day. 12
Table 5-1: Metro data, Methods 2, 3 and 4, last trip of the day 15
Table 5-2: Metro data, Trip-based and global methods, comparison of results 16
Table 5-3: Metro data, Trip-based, Traveller-based and global methods, comparison of results 16
Table 5-4 Applicability, accuracy (exact match and near match) by method. Data set 2. 18
Table 5-5: Trip Lengths, Methods 1 to , Kilometres 19
Table 5-6: Trip Lengths, Methods 1, 2, 3, 5, 6, 7, Kilometres 19
Table 5-7: Trip Length methods 1, 2, 4, 5, 6, 7, Kilometres 19
Table 5-8: Trip Length applying methods 1, 2 and 7, Kilometres 19
Table 5-9: Trip Length by time period 20
Table 5-10: Trip Length for AM only broken down by method Dataset 2 20
Table 5-11: Trip Length distribution by method (1, 2, 7) 21
Table 5-12: Trip Length distribution by method (1, 2, 8) 21
Table 6-1: ENCTS to TNDS matching numbers, First week of May Data 24
Table 6-2: Summary of the geographical match process (first week of May records) 26
Table 7-1: Method 3 Manual Review 29
Table 7-2: Alighting estimation algorithm applicability using Methods 1, 2, 5, 6, 7 29
Table 7-3: Average crow-fly distance in km from different sources (May+June 2016) 30
Table 7-4: Average trip distance in km for different operators (First Week of May Only) 32
Table 9-1: Alighting estimation applicability over the Ticketer data using methods 1, 2, 5, 6, 7 and 9. 39
Table 9-2: Average crow-fly distance trip length in km from different sources 39
Table 10-1: Alighting estimation algorithm applicability for the different bus datasets 47

1. Introduction

1.1 Context

The Urban Transport Group (UTG), which represents the seven largest city region strategic transport bodies in England, has, as part of its remit to promote collaborative working between its members, long worked on the development of smart ticketing systems, and the impact of this work is now beginning to be felt.

The growing take-up of smartcard ticketing in UK city regions is creating a valuable new source of information on travel behaviour, which could come to complement or replace more traditional surveys. However, bus smartcard systems collect information only on passengers' boarding point, which is of limited value on its own.

In order to generate origin-destination matrices, it is necessary to devise a method for inferring the alighting point, given longitudinal data on an individuals' history of boarding points. A number of transport authorities have developed such methods, often described as reverse journey matching. However, observed matching rates can vary considerably and are thought by some to be lower than is acceptable, for example, for revenue allocation purposes. There is also a sense that current methods can be subject to some forms of statistical bias.

Accordingly, UTG has commissioned AECOM to undertake some research into appropriate algorithms for estimating distributions of alighting points and perhaps to begin to develop a suitable user tool for calculating alighting points for a large dataset of smartcard data.

One particularly important application, and the key focus of this project, although not its only aim, is the calculation of appropriate operator reimbursement for agreeing to carry passengers eligible for free bus travel under the English National Concessionary Travel Scheme. In order to perform this calculation, it is necessary, amongst other things, to estimate total travel volumes and trip distance profiles relating to ENCTS passengers on the services of individual operators, as well as the market share of different ticket types amongst the wider travelling population.

The original work was carried out by AECOM in mid 2017 to research the use of smartcard bus transaction data to develop a representation of bus traveller demand. In mid 2018, UTG appointed AECOM to carry out an extension of the work to include some analysis using a new data set from the Ticketer system.

1.2 Scope of Study

The study aims to develop, validate, and prove a suitable algorithm for estimating a distribution of alighting points for smartcard data. The overall distribution (e.g. trip lengths, key movements) is what is of interest, rather than the precise alighting points of individuals, as the data will not be used at an individual level (for example, it would not be suitable for charging the bankcards of individuals).

While concessionary reimbursement is the main focus of the study, other potential applications for the data are also envisioned, such as:

- Revenue apportionment for multi-operator ticketing schemes.
- Development of travel demand patterns for transport models, transport scheme appraisal, and general transport planning.

The study aims to develop a distribution; expanding this to represent all trips (e.g. including trips that failed to swipe onto the bus correctly), is not part of the scope.

Originally it was assumed that the boarding point information provided in the smartcard data was suitable for analysis (i.e. mappable fairly simply to some geographic data). However it became clear over the course of the work that this was not the case, and the scope of the project was extended to spend some time considering the problem of mapping the supplied boarding points to geographic locations.

The original 2017 work considered several datasets of passenger boarding or ticket sales information from a number of sources. In 2018, a new data set, the Ticketer database, which contains bus passenger boardings linked to GPS coordinates identifying the position of the bus vehicle, was considered.

1.3 Report Structure

The remainder of this report is laid out in sections as follows, largely following the chronological progression of work undertaken:

- Chapter 2 briefly reviews previous work that we are aware of on similar alighting point estimation processes.
- Chapter 3 describes the test datasets that were made available for this study.
- Chapter 4 sets out principles coming out of previous work that were used as a starting point for this study.
- Chapter 5 discusses application of alighting point estimation algorithms to data from the Tyne & Wear Metro system.
- Chapter 6 discusses mapping Tyne & Wear bus boarding data to geographic locations, which was required for the following phase of work.
- Chapter 7 discusses application of alighting point estimation algorithms to Tyne & Wear bus data.
- Chapter 8 discusses how the bus stops in Ticketer data were derived from the boarding locations
- Chapter 9 discusses application of alighting point estimation algorithms to the West Yorkshire Ticketer data
- Chapter 10 summarises the study and conclusions.

2. Literature Review

2.1 TfL Study

Some research has been carried out at Massachusetts Institute of Technology (MIT) on estimating passenger flows in London using Oyster data. This is reported in *Intermodal Passenger Flows on London's Transport Network (Gordon, 2012).* Part of this thesis discusses both mapping of bus boarding observations to the geographic locations, and estimation of alighting points for bus passengers based on Oyster boarding points.

The mapping of bus boarding observations to geographic locations made use of GPS data on actual locations of individual bus vehicles at any given time, which could be referenced to the ticket boarding data. We do not, in this study, have access to such GPS data, so this approach is unlikely to be very useful.

The primary technique described for estimating destinations is the "reverse journey matching" method, whereby:

- A journey's destination is initially assumed to be the origin of the next journey on the same day.
- If a journey is the last journey of the day, the destination/alighting point is assumed to be that of the traveller's first boarding of the day.

The process also validated these initial assumptions by considering whether these points could meaningfully be mapped to any point on the bus route boarded. The processed checked:

- That the bus boarded was travelling in the correct direction for the estimated alighting point.
- That the "interchange distance" between the two bus services (if indeed they were different) was below a given maximum (750 metres was used).
- That there was time between timetabled alighting time and the next boarding event for the passenger to have made the interchange at a reasonable maximum speed.

Using this approach, the authors of the report were able to match about 75% of alighting points to plausible locations.

2.2 Transport Modelling Experience

The AECOM team undertaking this study have considerable experience in using bus ticket machine data to develop "demand matrices" of passenger flows between origins and destinations for use in transport modelling and transport planning.

Although this is not exactly the same problem, it is closely related. The ticket data generally do not contain any unique identifiers for individual, making the TfL approach impossible to apply. However, they very often do contain alighting points for *some* ticket types (singles and returns), for which the ticket records the approximate alighting point. They also contain text descriptions of fare stages, which are lacking in the smartcard data received for this project.

An approach which was used fairly heavily in processing these data was to assume that the distribution of alighting points for any given boarding point on any given bus was adequately described by the tickets for which alighting data were available, and that this distribution could be applied proportionally to any boarding for which the alighting point is unknown. This has worked well for transport modelling.

3. Study Datasets

3.1 Tyne & Wear Metro Data

Two sets of data were received from the Tyne & Wear Metro representing users of Gold Cards (this card is available to elderly passengers only and entitles them to, having paid for the card, free travel outside the morning peak). These differ from the bus data in that they *do* contain alighting points, allowing us to test algorithms on datasets in which we can check their accuracy.

Unfortunately, the data contain many records where the traveller failed to swipe at one or both ends of the trip. There are ticket gates at city centre locations, requiring users to swipe to enter or exit the network; however this is not true at many smaller stations and users sometimes fail to swipe.

The first dataset contains around 37,000 records for May, June and July 2016. The data shows the date of the trip, the alighting time of the trip, the boarding and alighting stations, the ISRN number of the card that made the trip and the proportional of trips made by that card which could be matched to an “alighting” swipe.

Table 3-1: Metro data set 1, Trips by total number of trips the user made on that day

Number of Trips made in one day	Number of trips	Percentage of total trips
1	13,284	37.0%
2	19,128	53.3%
3	2,361	6.6%
4	960	2.7%
5	150	0.4%
6	12	0.0%

The second set of data contains 120,000 records for May, June, July and August 2016. This dataset, in contrast to the first, includes “unmatched” single swipes with no other end of the trip (whether the single swipe is a boarding or alighting is not possible to determine).

Table 3-2: Metro data set 2, Trips by total number of trips the user made on that day

Number of Trips made in one day	Number of trips	Percentage of total trips
1	32,193	26.2%
2	71,460	58.1%
3	12,444	10.1%
4	4,728	3.9%
5	1,280	1.0%
6	384	0.3%

20-30 smartcards were selected randomly from the data and the journeys made on them were inspected in detail.

The simplest observed situation is simple return trips where user used the same boarding and alighting stations at both trip ends, with the order reversed. A simple match would be made taking the next boarding station for the next trip of the day as the alighting station for the journey studying.

A second situation similar to the first one where a return trip used a different station from the outbound journey the stations are within a walkable distance. Hence, a match would give as an answer geographically close to the right one.

Around 60% of the trips belong to one of these categories.

When only one trip was captured on a day it was observed that usually this was due to swiping problems where there wasn't an alighting station indicating or where the boarding and alighting station coincide, indicating that the user only swiped on one of the end of the trip on each way, possibly due to the lack of gates at the non-swiped station.

Around 25% of trips are the only trip the user made on that day.

Around 10% of trips fall into neither of the above categories.

It can be observed that for some users, usually corresponding with those with a low proportion of fully swiped trips, only one station is recorded as boarding station for several consecutive trips. This effect is accompanied in most cases by a failure to find an alighting station. Again this is most likely due to the fact that the user only swipes at those stations with gates where they are forced to swipe in order to get in or out of the Metro system.

3.2 Tyne & Wear Bus Data

Three sets of data were received from Nexus concerning Tyne & Wear bus travel.

The first dataset contained 36,000 records covering all services on a single day #01/02/2015, while the second contained 11,000 records in the whole month of February 2015 for the single service #71.

The third (full) dataset contained over 6 million records covering all services for May and June 2016.

All sets of data contain the boarding authority, the IRSN (card ID) the operator name, the service number, the date and time (at boarding) the sequence number (boarding point) and number of journeys. The full dataset (only) also included a "staffID" column linked to the identity of the bus driver (but anonymised, so that no one involved in data processing could identify a particular person). This was useful in tracking the movement of bus vehicles through the network.

Some problems with the data format were encountered on the date column in the full dataset. A very small proportion of the records appeared to contain non-printing characters that prevented the data from being read correctly by Excel. These were fixed manually.

The key problem in interpreting these data is that the "sequence numbers", which are intended to represent the location at which the traveller boarded the bus (not at the individual bus stop level, but by something close to a service "fare stage"), do not come with any lookup against real locations or text descriptions.

Nevertheless we can assume that any given service is numbered consistently, so that "service 11, sequence number 4" always represents broadly the same place.

Table 3-3: Tyne & Wear Bus Data #71, Trips by total number of trips the user made on that day

Number of Trips made in one day	Number of trips	Percentage of total trips
1	8836	81%
2	1963	18%
3	116	1%
4	20	0%

Table 3-4: Tyne & Wear Bus Data #010215, Trips by total number of trips the user made on that day

Number of Trips made in one day	Number of trips	Percentage of total trips

1	5468	15%
2	16598	46%
3	4920	14%
4	5248	15%
5	1715	5%
6	1122	3%
7	490	1%
8	280	1%
9	63	0%
10	70	0%

In order to facilitate easier analysis, some of the processing of the third (full) dataset was performed on only the first week of May, rather than the full two months. This is discussed further in chapter 7.

It may be noted that there are fewer “only one trip on the day” records in the bus data that for the Metro (only around 15% in the single-day data, and around 20% in the full dataset). This is because failures to swipe properly are much rarer.

3.3 Liverpool City Region Data

Similarly-formatted bus data were received from Merseytravel for the Liverpool City Region. This consisted of only 210 records that included date and time, service-operator ID, origin stage, the card ID, and the direction of travel (not available in the Nexus data).

Table 3-5: Liverpool City Region Data

Number of Trips made in one day	Number of trips	Percentage of total trips
1	34	16%
2	62	30%
3	54	26%
4	44	21%
5	0	0%
6	6	3%
7	0	0%
8	0	0%
9	9	4%

It was initially thought that LCR data would be the primary test dataset, but technical problems prevented this and no further data were used.

3.4 Ticketer data (West Yorkshire)

AECOM was provided with spreadsheets containing a single weeks worth of data from November 2017, for various minor bus operators in West Yorkshire. None of the operators were dominant in their area.

In addition to the passenger boarding data, an auxiliary dataset about bus journeys was available. This were not found to contain any additional information useful to the work.

The passenger boarding data covered only users of concessionary smartcards and consisted of 45,838 records. They included 23 variables, 8 of which were empty for all records.

The data contain date, time, service number, bus journey code (equivalent to bus departure time), smartcard ID (anonymised, but persistent within the dataset) bus vehicle registration, and latitude and longitude of bus position. They also include a column with bus the NaPTAN code of the bus stop; unfortunately this is missing for the majority of records (only about 10% of records include this information).

A very small number of records (about 0.5%) contained no latitude/longitude information. These were ignored.

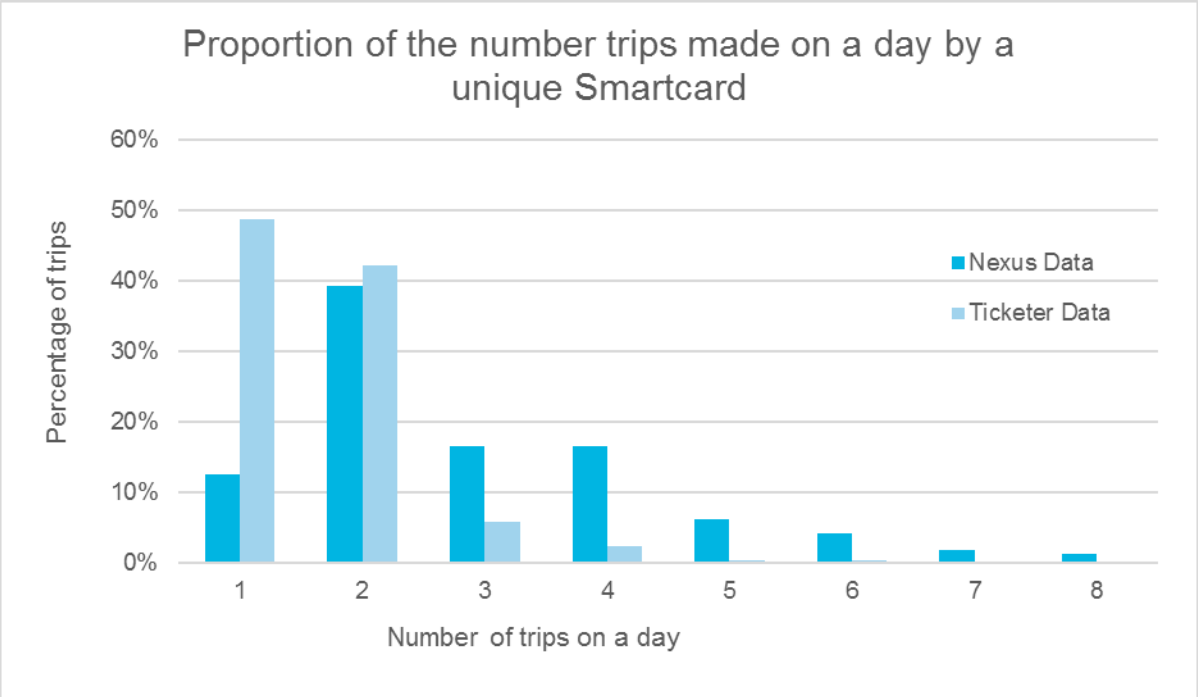
Table 3-6 describes the numbers of trips in the Ticketer dataset when grouped by the number of swipes on a day by a unique Smartcard.

Table 3-6: Boarding_data_set, Number of instances in which a unique Smartcard made X swipes on a day.

Trips made in one day	Number instances	Percentage of total trips
1	22246	49%
2	19206	42%
3	2658	6%
4	1108	2%
5	190	0%
6	120	0%

Figure 3-1 compares the number of trips that fall in each cluster when the trips are clustered by the number of trips made on a single day.

Figure 3-1 Comparison between Nexus and Ticketer on the number of trips made on a day by a unique smartcard



4. Alighting Estimation- Concepts

4.1 Trip-based Alighting Estimation

Trip-based alighting estimation encompasses methods that look at a particular trip made by one card and use information specific to that trip to estimate the alighting point. The “reverse journey matching” described in section 2.1 can be applied to a high proportion of trips (any trip with at least two trips recorded on the day). It will “fail” in general when the traveller mixes modes or starts and finishes a day in different areas.

In the following chapters, by “Method 1”, we mean matching a boarding with the next one made by the traveller on the same day, using the second boarding as the alighting station for the previous trip. This method can obviously only be applied if there is more than one trip made by one card on one day, and the journey is not the last one of the day.

Method 2 is used to estimate the alighting point for the last trip of the day. The approach taken identifies the alighting point for the last trip of the day by noting where the user started the day and assuming that they want to go back there at the end of it.

Method 4 is an extension to Method 1, tested in some of the analysis, which assumes the alighting point of a trip is the user’s next boarding point regardless of when this occurs (i.e. not necessarily on the same day).

4.2 Traveller-based Alighting Estimation

There are then methods that can be applied that look at other travel of the same user, without specifically taking account of the position of the trip in question in the user’s travel.

Method 3 selects the most common starting point of a day for the user and, taking this as the home of the user, assumes that their final trip will return to that point. In this respect it is similar to Method 2, but considers the user’s overall most common starting point rather than the starting point of the specific day of the trip.

Methods 5 and 6 require previous analysis to have already assigned alighting points for some trips. They use previously estimated alighting data for a particular individual and use this to estimate an alighting for a trip that could not be infilled using previous methods.

Method 5 selects the most common alighting point for the user, given that they board at the boarding point of the current trip. This can be applied only if at least one boarding made by the traveller at the boarding point in question was previously infilled using methods 1 to 4.

Method 6 simply takes the most common alighting point overall for the specific user (without reference to the boarding), and uses it as the alighting point for the trip case of study.

4.3 Full Dataset Alighting Estimation

Where none of the previous methods are applicable, either because the user never made more than one trip per day or other methods returned implausible (same as boarding point, or nowhere near any point on the service route) alighting points, we need a global infilling method.

a) The simplest way (Method 7) is to assign the alighting point as the most common alighting point for a given boarding point in the dataset overall. This approach is the statistically most likely of the three to get the “right” answer, if the exercise were to guess the point in which that particular user got off, but it may introduce a bias in the overall distribution.

b) A second approach (Method 8) would be to proportionally split trips according to the distribution of number of passengers using each possible alighting point for a particular boarding point. This method is obviously weaker when it comes to guessing where a particular individual alighted and more complex to apply, but likely to be better at estimating the overall distribution.

c) A stochastic approach could be used to randomly assign the alighting according to probabilities in the distribution. This is similar to method b; it avoids duplicating records and generating

fractional trips, at the cost of potentially making it hard to reproduce results exactly (random seeds could be used to ensure some measure of reproducibility). This method was not tested in any analysis, but is noted here for completeness.

d) For bus services or any other data where the concept of “service” or “line” exists , we could take the most. common alighting point for a given line or service in the dataset overall, irresepective of boarding point. This is only likely to be helpful if there are no suitable records for the specific boarding point. Methods summary

Method 1: Next boarding point on the same day.

Method 2: First boarding point of the current day (applied to the last trip of the day only).

Method 3: Most common first boarding point of the day for this smartcard.

Method 4: Next boarding point on a subsequent day (applied to the last trip of the day only).

Method 5: Most commonly chosen alighting for the current boarding point by this user

Method 6: Most commonly chosen alighting by this user.

Method 7: Most commonly chosen alighting for the current boarding across all users.

Method 8: Distribution of alighting points chosen for the current boarding, applied proportionally.

Method 9: Most commonly chosen alighting for all users of the current service.

5. Alighting Point Estimation- Nexus Metro Data

5.1 Methodology

The two Metro data sets provided were studied to understand how the matching algorithms might work. The algorithms were applied to the data using only the boarding points as evidence, to estimate the alighting points. It was then possible to see how accurate this estimation was, as the true alighting points are also in the data.

We did not attempt to calculate alighting points for records where:

- Only one match was recorded (which could have been either a boarding or an alighting).
- The boarding and alighting events matched were for the same station. This means that either the user changed their mind and did not make a journey (so the record does not represent a trip at all), or they failed to swipe in either direction at the other end of the trip and made a return journey within the 90 minute matching threshold (so the record represents two trips, but with no indicator of the other end).

Both of these indicate swiping

s, which should have no analogue in bus data, as the one thing we are sure of in bus smartcard data is that a boarding really is a boarding; there is no confusion between boardings and alightings.

However, these discarded records were used as part of the algorithm to match the valid records. For example, if one journey was actually made between Sunderland and Gateshead, and the user records a later swipe that same day for Gateshead only, that later record will be used to estimate the alighting point for the first record (correctly in this case).

Dataset 1 did not contain any unmatched single-swipe records. For comparison, some analysis has also been done with “Dataset 2 -“, which is dataset 2 with the unmatched records removed.

In assessing the accuracy of the algorithms, both exact matches (the correct station) and near matches (a station with 1.5 km of the correct station) have been calculated. This allows us to see whether we have identified broadly the right location, with the traveller perhaps having walked between stations.

5.2 Record-Level Accuracy

One important question to be answered initially is which of the three proposed methods for estimating the alighting point of the last trip of the day performs best. This analysis is shown below.

Table 5-1: Metro data, Methods 2, 3 and 4, last trip of the day

	Method 2 Start of current day			**Method 3 Most common day start for this traveller**			**Method 4 Start of next available day for this traveller**		
Data set	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match
Dataset 1	28.94%	**74.04%**	85.49%	28.39%	**73.69%**	84.18%	24.28%	**63.64%**	74.60%
Dataset 2 -	28.74%	**75.55%**	85.46%	28.35%	**75.17%**	85.04%	24.66%	**64.71%**	75.93%
Dataset 2	29.14%	**76.93%**	86.96%	29.16%	**75.67%**	85.52%	25.77%	**64.93%**	76.37%

Observing the Table 5-1 it can be seen that using the next available boarding, on a subsequent day (method 4), gives substantially poorer results; as well as being applicable to fewer trips, because a

subsequent day of data is required, which will not always be available. Methods 2 and 3 give very similar results, but 2 is consistently slightly more accurate.

Having discarded 3 and 4 as first attempts at infilling final trips of the day, we can apply methods 1, 2 and 7 in order to the datasets, producing the following results.

Table 5-2: Metro data, Trip-based and global methods, comparison of results

	Method 1 Next boarding today			Method 2 Start of current day for this traveller			Method 7 Overall most common alighting for this boarding		
Data set	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match
Dataset 1	33%	**74%**	90%	29%	**74%**	85%	38%	**33%**	52%
Dataset 2 -	34%	**70%**	92%	29%	**76%**	85%	38%	**25%**	47%
Dataset 2	43%	**70%**	93%	29%	**77%**	87%	28%	**21%**	41%

Methods 1 and 2 are quite accurate. If near matches are permitted, they have nearly 90% accuracy overall. The last trip of the day approach (method 2) is slightly less accurate than the next boarding (method 1). Method 7 is much poorer, as might be expected. Although it can be applied to all remaining trips (this is mostly trips that are the only trips that user made on the day, although it also includes a few trips for which methods 1 and 2 give the alighting station as the same as the boarding), it returns a near match less than 50% of the time.

Using datasets 1 and 2- (without the unmatched records), methods 1 and 2 can be applied to 60% of trips. Including the unmatched records allows the algorithms to be applied to 70% of trips without any loss of accuracy, although the accuracy does not improve much either.

It is quite possible that most of the residual ~10% inaccuracy in methods 1 and 2 can be attributed to users swiping incorrectly. There will be entirely missing journeys, and apparently matched journeys that do not properly represent trips in the dataset. This particular problem should not occur with bus data; it is always possible to identify an event confidently as a boarding (some boardings may occasionally be missed from the data, but that is only a problem for expansion of the results).

Some improvement in the record-level accuracy can be achieved by using methods 3, 4, 5 or 6 where possible instead of method 7. These involve looking at other travel of the same user where it is available.

When methods 1, 2, 5, 6 and 7 are applied the results are as follows:

Table 5-3: Metro data, Trip-based, Traveller-based and global methods, comparison of results

	Method 1 Next boarding today			Method 2 Start of current day for this traveller			Method 5 Most common alighting for this traveller and this boarding			Method 6 Most common alighting for this traveller			Method 7 Overall most common alighting for this boarding		
Data set	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match
DS	33%	**73%**	89%	28%	**73%**	85%	23%	**48%**	63%	14%	**26%**	58%	1%	**6%**	6%

1														
DS	43%	**70%**	93%	29%	**77%**	87%	20%	**43%**	60%	7%	**15%**	41%	1%	**9%**	11%
2														

Methods 5 and 6 are significantly more accurate than method 7 alone, and can be applied to almost all remaining trips.

Method 7 is still required for ~1% of trips where the traveller has only one boarding in the entire dataset or all other methods return alighting stations that are the same as the boarding station. It has very poor accuracy with this residual 1%. Splitting the record and using a full distribution instead of merely the most common alighting would certainly not improve the accuracy at a record level, although it might improve the quality of the overall matrix distribution.

Methods 3 and 4, discarded above in favour of method 2 where method 2 is applicable, may still be of value where methods 1 and 2 fail, either because there is only one trip made on the day, or because applying method 1 or 2 returns the same boarding as alighting.

Several combinations of methods applied to dataset 2 are shown in the table on the following page.

Four of the five combinations analysed all return very similar overall accuracy, with the combination of all seven methods being the best, but only by a very small margin. The simplest method, involving only methods 1, 2 and 7, is somewhat poorer in this respect.

In most combinations, method 6 is applied only to a relatively small proportion of trips, and method 7 to an even smaller one.

Curiously, method 3 appears to perform better than method 5. This is perhaps unintuitive; one would assume that a process based on taking account of where the trip was boarding would return better results than one that infills an alighting without reference to the boarding point. This is not simply an artefact of the order in which the methods are applied; all testing around these two methods implied that using method 3 where possible returns more accurate results.

This may be slightly misleading. Where method 3 gets the answer substantially wrong, it will often do so because it assigns the most common start point of the day as the alighting point to a trip that actually *begins* at that point. When this happens, it is possible to reject the match because the boarding and alighting points are the same. So it may be that it is simply easy to spot when method 3 gets it wrong, rather than that it is accurate per se. Caution will need to be exercised in applying the method to bus data where boarding and alighting at the same fare stage is in fact possible (unlike in the Metro system, where the “points” are individual stations).

It is important to note here, however, that while a comparison against actual alighting points is informative, the object of the process is *not* to accurately identify alighting points individually by record. Rather we are concerned about producing an accurate distribution of travel as a whole. So the method that is individually most accurate may not necessarily be the best.

In the next section, we discuss average trip lengths.

Table 5-4 Applicability, accuracy (exact match and near match) by method. Data set 2.

Data set 2	Method 1			Method 2			Method 3			Method 4			Method 5			Method 6			Method 7			Total	
	Next boarding today			Start of current day for this traveller- last trip only			Most common day start for this traveller			Start of the next available day for this traveller			Most common alighting for this traveller and this boarding			Most common alighting for this traveller			Overall most common alighting for this boarding				
Methods applied	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match	Applied to	**Exact match**	Near match	**Exact match**	Near match
1, 2, 3, 4, 5, 6 & 7	43%	**70%**	93%	29%	**77%**	87%	10%	**55%**	65%	4%	**37%**	51%	12%	**36%**	55%	1%	**20%**	52%	0%	**5%**	5%	**64.3 %**	81.9 %
1, 2, 3, 5, 6 & 7	43%	**70%**	93%	29%	**77%**	87%	10%	**55%**	65%	-	**-**	-	14%	**35%**	56%	3%	**17%**	48%	0%	**7%**	9%	**63.6 %**	81.4 %
1, 2, 4, 5, 6 & 7	43%	**70%**	93%	29%	**77%**	87%	-	**-**	-	12%	**44%**	57%	13%	**38%**	56%	2%	**19%**	50%	0%	**6%**	6%	**63.7 %**	81.8 %
1, 2, 5, 6 & 7	43%	**70%**	93%	29%	**77%**	87%	-	**-**	-	-	**-**	-	20%	**43%**	60%	7%	**15%**	41%	1%	**9%**	11%	**62.2 %**	80.2 %
1, 2 & 7	43%	**70%**	93%	29%	**77%**	87%	-	**-**	-	-	**-**	-	-	**-**	-	-	**-**	-	28%	**21%**	41%	**58.3 %**	76.7 %

5.3 Trip Lengths

Average trip lengths have been calculated for various combinations of methods in the analysis below. Because we have access to real alighting points, these can be compared with the actual trip lengths.

Table 5-5 shows the trip lengths broken down by method applied. Methods 1 and 2 are almost perfect. The high accuracy in finding the alighting point is translated in a trip length very close to the actual one. Method 3 and 4 understate the trip lengths. The understatement of methods 3 and 4 seems to be compensated by an overstatement for methods 5 and 6. Method 7 severally underestimates the trip length, but it is applied in most of the combinations to such a small proportion of trips that the impact on the overall trip length is negligible. Interestingly, when applied to a larger proportion of trips, as in the fourth table below, it performs better.

Table 5-5: Trip Lengths, Methods 1 to , Kilometres

Applied to	100%	43.3%	29.1%	10.2%	4.2%	11.6%	1.4%	0.1%
Metro validation data v2	Total	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7
Actual trip length	7.1	7.4	7.6	6.1	6.3	5.6	6.2	8.4
Estimated trip length	7.0	7.4	7.6	5.4	5.3	6.2	6.8	4.0
	-1%	0%	0%	-12%	-15%	10%	10%	-52%

Table 5-6: Trip Lengths, Methods 1, 2, 3, 5, 6, 7, Kilometres

Applied to	100 %	43.3%	29.1%	10.2%	13.8%	3.1%	0.4%
Metro validation data v2	Total	Method 1	Method 2	Method 3	Method 5	Method 6	Method 7
Actual trip length	7.1	7.4	7.6	6.1	5.8	6.1	6.6
Estimated trip length	7.2	7.4	7.6	5.4	6.8	7.1	4.0
	1%	0%	0%	-12%	17%	16%	-39%

Table 5-7: Trip Length methods 1, 2, 4, 5, 6, 7, Kilometres

Applied to	100 %	43.3%	29.1%	12.5%	13.1%	1.8%	0.2%
Metro validation data v2	Total	Method 1	Method 2	Method 4	Method 5	Method 6	Method 7
Actual trip length	7.1	7.4	7.6	6.2	5.6	6.5	8.2
Estimated trip length	7.1	7.4	7.6	5.5	6.1	6.8	4.0
	0%	0%	0%	-11%	9%	5%	-51%

Table 5-8: Trip Length applying methods 1, 2 and 7, Kilometres

Applied to	100%	43.3%	29.1%	27.5%
Metro validation data v2	Total	Method 1	Method 2	Method 7
Actual trip length	7.1	7.4	7.6	6.0
Estimated trip length	6.9	7.4	7.6	5.4
	-2%	0%	0%	-10%

Despite some significant inaccuracy in some of the methods individually, none of the combinations significantly misstates the overall trip length. Methods 3 4 5 and 6 appear, on balance and overall, to be about equally inaccurate, although 3 and 4 always understate and 5 and 6 always overstate, lengths.

We can also inspect the trip lengths by time period.

Table 5-9: Trip Length by time period

	Total	AM Period	Inter-Peak	PM Period	Off peak
Dataset 1 (methods 1, 2, 5, 6, 7)					
Actual Trip Lengths	6.5	5.3	6.7	6.7	5.6
Estimated Trip Lengths	6.7	5.8	6.8	6.7	5.6
Difference	0.1	0.5	0.1	0.0	0.0
% Difference	**1%**	**11%**	**1%**	**0%**	**0%**
Dataset 2 (methods 1, 2, 5, 6, 7)					
Actual Trip Lengths	7.1	6.1	7.1	7.4	6.8
Estimated Trip Lengths	7.2	6.9	7.4	7.3	6.8
Difference	0.1	0.8	0.3	-0.1	0.0
% Difference	**2%**	**13%**	**3%**	**-1%**	**1%**
Dataset 2 (methods 1, 2, 3, 4, 5, 6, 7)					
Actual Trip Lengths	7.1	6.1	7.1	7.4	6.8
Estimated Trip Lengths	7.0	6.4	7.1	7.2	6.8
Difference	-0.1	0.3	0	-0.2	0
% Difference	**-1%**	**5%**	**-1%**	**-2%**	**0%**
Dataset 2 (methods 1, 2, 7)					
Actual Trip Lengths	7.1	6.1	7.1	7.4	6.8
Estimated Trip Lengths	7.2	6.5	7.3	7.2	6.9
Difference	0.1	0.4	0.2	-0.2	0.1
% Difference	**2%**	**8%**	**2%**	**-1%**	**1%**

The AM period trip length based on the estimated alighting point is significantly above the actual trip length for the trips on that time period. Further study shows that this is driven mainly by methods 5 and 6.

Table 5-10: Trip Length for AM only broken down by method Dataset 2

Applied to	**100%**	**66%**	**1%**	**5%**	**6%**	**20%**	**1%**	**0%**
Method	All	1	2	3	4	5	6	7
Actual trip length	6.1	6.6	4.4	5.4	5.9	4.7	4.9	4.2
Estimated trip length	6.4	6.7	5.2	4.2	5.3	5.9	6.7	4.2
	5%	*2%*	*17%*	*-22%*	*-9%*	*26%*	*37%*	*0%*

5.3.1 Trip length distributions final infill

For the final infill used in those cases where all the previous methods failed to return a valid alighting point the algorithm uses the boarding estimated alighting data to infill the rest of the alightings. As discussed in the previous chapter, there are several approaches to how to make use of the existing data to infill the mention trips.

All previous analysis has used “method 7”, simple assignment of the most common alighting point. We can consider how well this compares with the use of a full distribution of alightings (“method 8”).

Table 5-11: Trip Length distribution by method (1, 2, 7)

Metro validation data v2	Total	Method 1	Method 2	Method 7
Actual trip length (km)	7.1	7.4	7.6	6.0
Estimated trip length (km)	6.9	7.4	7.6	5.4

Table 5-12: Trip Length distribution by method (1, 2, 8)

Metro validation data v2	Total	Method 1	Method 2	Method 8
Actual trip length (km)	7.1	7.4	7.6	6.0
Estimated trip length (km)	7.3	7.4	7.6	6.8

Surprisingly, method 8 actually returns a worse trip length validation than method 7.

Method 7 understates trip lengths. This is as expected. It uses the most common alighting point, which is quite likely to imply a shorter trip length than the mean (the mode is smaller than mean in many real-world data contexts, and certainly likely to be so in the case of trip lengths).

Method 8 returns a typical trip length (6.8) that is broadly similar to the actual overall trip length (7.1), as one would expect, because methods 1 and 2 are highly accurate and method 8 essentially uses the same distributions. However, this fails to actually generate an accurate answer, because the trips that methods 1 and 2 fail for are significantly shorter than average (only 6.0 km).

5.4 Conclusion

The object of analysis of the Metro data is informing the specification of an algorithm that can be applied to the bus data where we lack detailed information to validate against (i.e. actual alighting points). On the basis of the preceding analysis, our views were as follows:

- Methods 1 and 2 are highly accurate, far more so than any alternative explored, and should obviously be used where possible.
- Methods 3, 4, 5 and 6 are all broadly comparable in accuracy. Using some of them is desirable, as the alternative method 7 is clearly poorer. On the basis that applying all four results in methods 4 and 6 being rarely used and 4 and 6 seem slightly less accurate at a record level in any case, we suggest a somewhat simplified approach where methods 3 and 5 are used, and methods 4 and 6 are not.
- When methods 3 and 5 are used, method 7 is restricted to only ~3% of records. Consequently, there is little value in complicating method 7 by using a full distribution (method 8) instead, particularly since method 8 does not actually appear to improve the trip length

comparison. So we suggest method 7 is used to fill in any data that cannot be matched via other methods.

These views were revised as the analysis of the bus data progressed, as discussed in chapter 7.

6. Geographic Matching- Nexus Bus Data

6.1 Context

The bus smartcard data received from Nexus, as previously mentioned, code boarding points using numeric values, generally but not exclusively sequential, specific to each service. These broadly correspond to fare stages in the bus route, so for example, a service might call at 50 bus stops, divided into ten groups of 5, numbered 1 through to 10, which are used to calculate single and return fares. It is these numbers 1 to 10 that are identified in the smartcard data as boarding points.

It should be noted that the numeric identifiers do not always correspond exactly with actual fare stages, but they represent a similar concept. These numeric identifiers will be called “stages” in the text which follows.

Unfortunately, the data lack any identifier that would enable the numeric data to be placed geographically; there are not even any descriptions of each stage.

Consequently, in order to make any sense of the boarding data, and before any serious attempt can be made to begin estimating alighting points (which of course are not provided at all), it is necessary to estimate where the boarding stages actually are. Because interchanging between buses or using different service numbers on a return journey is relatively common, we will want to understand the actual geographic location of each stage, not just its position along the route.

Timetable data are available that record all bus services in the UK, their departure times, travel times, and routes through bus stops. Bus stop locations in British National Grid coordinates are also available. Therefore, some sort of mapping process between the timetable data and the smartcard ticket data is required.

The traveline national dataset (TNDS)[1] contains timetable information in a text format that can be downloaded and is suitable for use in an automated process. This uses National Public Transport Access Node (NaPTAN) codes to identify bus stops. Coordinates for NaPTAN bus stop codes are available[2].

This analysis was carried out only on the full dataset (May and June 2016), not the smaller sample datasets. To speed up the process, only the first week of May was used.

6.2 Outline of Approach

The staffID record in the ENCTS ticket data is useful in geographic matching (although it has no obvious use in alighting point estimation itself). This record identifies the driver of the bus. Using this, it is possible to track the movement of a single bus vehicle throughout the day, because a driver obviously cannot be driving more than one bus at a time, and thus understand the stages the bus travels through and time times it takes to travel between stages.

The problem to be solved can be broken down into three steps:

- Mapping services in the smartcard (ENCTS) data to services in the timetable data. This is not entirely straightforward. Service numbers are not necessarily unique, even with one transport authority area. Also, the ENCTS data sometimes record services using identifiers different from the advertised operating number (often to eliminate ambiguity, but the methodology is often unclear).
- Identifying the endpoints and order of stages within the ENCTS data for each service. While the numbers are usually sequential and in order, this is not always the case, and there may be gaps even when the stages are ordered numerically. A given service may not follow exactly the same route on all of its journeys, so the stage sequence may differ.

[1] https://data.gov.uk/dataset/traveline-national-dataset
[2] https://data.gov.uk/dataset/naptan

- Map the stages, using journey times, to bus stops or bus stop clusters in the timetable (TNDS) data. The primary difficulty here is working out which end of the service route is which. Assuming the route is relatively constant; once this is achieved the remaining stages can be allocated fairly accurately by comparing travel times between them in the ENCTS and TNDS data.

6.3 Application

For this research, only the data for the first week in May was used. The algorithm could relatively easily be extended to the full two months; this might improve accuracy slightly.

The first step to the geographical match is identifying which of the services recorded on the ticket data can be found in TNDS. The services have a name and an operator. Using these two entries most of the services can be identified, however some return more than one match. This happens when an operator runs two or more buses running with the same service number within the geographical area (North-East, because TNDS is region-based). By looking at the description we were sometimes able to eliminate services that do not enter the Tyne and Wear area and remove ambiguity.

Some of the ENCTS service names couldn't be found at all in the TNDS data. After some analysis it was concluded that some of these services had extra letters to distinguish them from other services running with the same name. For example the 16S, 16SS and 16N are all number 16 buses run by the same operator in Sunderland, South Shields and Newcastle respectively.

Following manual disambiguation as described above, 89% of ENCTS records were matched, for 8% no match could be found, and for 3% there were multiple matches that could not be disambiguated.

Table 6-1: ENCTS to TNDS matching numbers, First week of May Data

	Number of services	Number of services %	Number of boarding records	Number of records %
Pre-manual revision				
One match	199	**55%**	588,628	**70%**
Multiple matches	20	**39%**	70,109	**22%**
No matches	141	**6%**	183,803	**8%**
Total	360		842,540	
Post-manual revision				
One match	232	**64%**	750,759	**89%**
Multiple matches	7	**34%**	25,247	**8%**
No matches	121	**2%**	66,534	**3%**
Total	360		842,540	

ENCTS records that could not be matched to a single unambiguous service in the timetable data were not processed further.

The process was assisted by the "staffID" record in the ENCTS data. This allows us to break boarding records down into "shifts", which are continuous sets of records, covering a single day, bus driver and service, in ascending order of time. This gives us a picture of the travel pattern of a single bus vehicle over a period of time.

Two possible methods were explored to identify the end point stages on the route.

The first approach involved analysis of the number of times that a stage is the first stage recorded in a shift compared to the number of times that a stage was recorded in total. In general endpoints of

services appeared to be “first” stages around 40-50% of time, while other stages are much more rarely “first” stages.

Although this method has some merit and quite often is able to identify either the actual endpoints or at least stages very close to them (for example, for one service stages 1 and 11 were identified as the endpoints; it was thought from manual inspection that 1 and 12 was more likely and 12 was a rarely used stage), it did return clearly wrong answers not infrequently.

A simpler approach of selecting the highest and lowest numbered stages that are ever recorded as the first records of a shift was thought, on inspection, to actually return slightly more accurate (though of course not perfectly accurate) answers.

This was refined slightly because several services appeared to use the codes “99” or “98” as a “missing data” identifier, so these numbers were excluded from being identified as endpoints.

11 out of the 232 services that could be matched to timetable data had only one fare stage that ever occurred as start of a shift. This corresponds to 0.5% of records.

The sequence numbers identified as endpoints were then matched with corresponding departures in TNDS, by comparing the time of the last record on the fare stage for that shift, with TNDS departures times. If the difference between the TNDS time and the ENCTS time was within a threshold of 5 minutes behind or 3 minutes ahead the TNDS departure was recorded as plausible. If more than one TNDS departure occurred in this range, then the closest departure to the ticket data record was chosen.

A single TNDS bus stop was then identified for each selected endpoint by examining all the matches to TNDS departures in the full ENCTS dataset for this service and stage and picking the most commonly identified origin bus stop.

Applying the algorithm described above, two bus stops (with NaPTAN stop codes) were assigned to the endpoints of 124 services. For the other 97 services it was found impossible to achieve this, either because two endpoints could not be identified in the first place, or, more commonly, because it was not possible to match both of these to TNDS departure times to assign a bus stop using the approach above. This represents about 13% of all the records in the ENCTS data.

Validating the results generated by this process is difficult because there is no other information in the ENCTS records. However, there are two ways to check the results other than the “find the departure times in the timetable” method actually used by the process:

- In some cases, it will be obvious that one end of the bus route is likely to be much more heavily used than the other; any radial route linking a residential area with a business centre is likely to exhibit this. By checking the number of boardings at each end, we can in some cases make a good guess which end is which. This method would be difficult to apply in an automated process; one would need employment and population data and a zoning system.

- It is possible to examine the travel times between the endpoints and the following stage. If, for example, at one end of the journey, there is a 15 minute gap between the end point and the next stage, while at the other end there is a 2 minute gap; this can be compared with a published timetable. The published timetables often use bus stop groups (they rarely list every single bus stop) that correspond well with the stages used in the ENCTS data, so it may be possible to identify which end is which. This method is difficult to apply in an automated process as well, since the TNDS data record individual bus stops, which do not correspond even approximately to stages.

16 randomly selected services were inspected using these methods. It was concluded that in 8 cases the algorithm was probably correct, in 2 cases the algorithm was probably wrong (endpoints were matched the wrong way around), and in 6 cases no conclusion could confidently be drawn. Consequently, our best guess is that the algorithm is roughly 80% accurate.

Table 6-2: Summary of the geographical match process (first week of May records)

Number of records	**969,690**
Number of records matched to a timetable service	860,575
Number of records using services with two endpoints matched to timetabled bus stops	730,920

6.4 Algorithm

The algorithm arrived at for this research is the result of time trial-and-error and validation as described above. However, the budget and time available for this part of the work was limited. It is quite certain that the approach could be refined and made more accurate and robust; what follows is merely an initial attempt.

The geographical match process uses two datasets:

- ENCTS data, containing one record per bus passenger boarding, recording the date, time, ID of bus driver, service number and origin sequence number ("fare stage").
- Travelline National Dataset (TNDS) data in TransXChange format for the whole region (North-East) relevant to the ENCTS data.

The object is to assign each possible service number/ fare stage combination to a NaPTAN bus stop code, and thus a British National Grid coordinate. Each fare stage in general represents more than one bus stop; the intent is to select a representative one, ideally close to the middle of the stage.

Firstly, the ENCTS services are mapped to the TNDS services. As noted in table 6-1, only just over half of services in the ENCTS data are identical to TNDS service numbers with no duplication. This is a larger proportion of records, as the rarely-used services tend to be harder to match. Some manual intervention was applied to improve this (removal of services clearly not in the Tyne & Wear area from the timetable lists; assignment of service numbers such as "12SS" to the "12" bus in South Shields, and similar).

The ENCTS data are sorted by staffID, date and time, in that order, and grouped into "shifts". A shift ends when any one or more of the day, the service number, or the driver ID changes. This represents the travel pattern of a single bus vehicle while operated continuously by one driver. This could involve several back-and-forth journeys in general. Shifts may also sometimes split a timetabled journey if the driver changes mid-route.

The algorithm first identifies the sequence numbers that are recorded at least once as first records in a shift. The largest and smallest sequence numbers from those are identified as "end points". In picking the end points several specific sequence numbers were excluded from the selection of "largest" number, as the size of the number compared to other numbers in the route and the low usage of the number by passengers suggested that they represented either "missing data" or a sequence number on another connecting route. These numbers were 99, 98, 94, 97, 111, 110, and 104.

For each endpoint, an attempt is then made to identify the time of every departure from that endpoint. The last record boarding at the endpoint prior to a record boarding at a different endpoint is recorded as the "departure time". This is compared to the actual timetabled departure times in the TNDS data. A TNDS departure within a range of 3 minutes later than the ENCTS time and 5 minutes earlier than the ENCTS time is sought. The originating bus stop for this TNDS departure is recorded. If more than one departure is found, the one closest to the ENCTS time is chosen. If none are found within the 8 minute range, no match is recorded.

This process assigns bus stops to every instance of an endpoint fare stage in the ENCTS data. The most commonly chosen one for each endpoint fare stage is selected as the correct bus stop.

This assigns bus stops to endpoints. It is then necessary to assign bus stops to intermediate fare stages. This is done by estimating travel times between fare stages.

Within each shift, the time between the previous endpoint's final record and the first record boarding at a given fare stage is recorded. This generates a large number of "time from start of route" estimates for each fare stage. Any time greater than the journey time for the entire bus route (from the timetable) is rejected. The average value of the remaining estimates is chosen as the average time from start of route.

This time is compared to the timetable. The first timetabled departure after noon with the correct origin point (i.e. same as endpoint) is selected, and the algorithm calculates which bus stop this bus would have reached by the time-from-start-of-route. This is recorded as the estimate of the bus stop to be assigned to the fare stage.

With this procedure the algorithm finds bus stops for the intermediate stations twice, one for each end point. i.e. the algorithm compares the time it takes to get to a given intermediate sequence number when starting the journey from both ends. Whichever estimate is associated with a greater number of records (i.e. larger sample size) is selected.

6.5 Possible Further Research

A number of possible improvements to the algorithm or areas for further investigation are described briefly below:

- Some in-depth dialogue with the bus operators to understand the data better and possibly work out a more robust approach from scratch, probably using some additional data. Or at least a more complete method for allocating smartcard records to timetabled services in TNDS.
- More investigation of methods for mapping endpoints to bus stops, possibly using intervals between stages and population levels to validate the initial allocation.
- Using wider thresholds for mapping endpoint departures to timetabled services but taking into account quality of match when comparing against other matches.
- A more robust method of averaging travel time between stop estimates that more clearly rejects unusually long times. Possibly using the median rather than the mean would be preferable.

7. Alighting Point Estimation- Nexus Bus Data

7.1 ENCTS Data Full Data set

Having completed the process described in chapter 6, the algorithm developed in chapter 5 for the Metro data was then applied to the Tyne & Wear (Nexus) bus data. There are a few differences in application to the two datasets.

All records for which the geographic matching process described in chapter 6 failed were not processed further in estimating alighting points. This is about 20% of all records, divided roughly evenly into records boarding services for which no timetable data could be unambiguously identified, and records boarding services for which one or both endpoints could not be allocated to bus stops.

While for a minority of these “un-matched” records an estimate could have been made, such an estimate would have no value in itself because the location of the boarding point is not known.

The key difference between the Metro and bus datasets is that the Metro system is treated as a single entity, so any boarding point can be mapped to any alighting; while each bus service has its own fare stages and we know that a user must alight from the same bus that they board.

Consequently, in applying method 1 and method 2, if the two bus services compared are not the same service, we must map a subsequent boarding to an alighting on the current service using geographical distance. In applying method 6, only records boarding the same bus service were considered.

Methods 1, 2, 3, 5, 6 and 7 were studied. They are described in full in section d), and are summarised again below. Following disappointing performance on the Metro data, we did not investigate Method 4 further. Method 8 was not used either, as Method 7 was only required for a small proportion of records, so the additional complexity was not felt worthwhile.

Method 1: Next boarding point on the same day.

Method 2: First boarding point of the current day (applied to the last trip of the day only).

Method 3: Most common first boarding point of the day for this smartcard (applied to the last trip of the day only).

Method 5: Most commonly chosen alighting for the current boarding stage by this user

Method 6: Most commonly chosen alighting by this user for this bus service.

Method 7: Most commonly chosen alighting for the current boarding stage across all users.

Ties in “most common” are broken by simply selecting the first record in the list.

Method 1 is simple to apply if the two relevant journeys use the same bus services. In this case the algorithm works as it did with the Metro data.]

If the next boarding is made on a different bus route the following approach is taken.

The next boarding’s geographical coordinates are inspected. If they are missing (because the following service is unmatched), method 1 is discarded and not used. If they are present, the algorithm tries to find the closest stage to this on the current service, within 4km. 4km may appear long for a walk distance. However, it is necessary to make allowances for errors in the geographic matching process itself, as well as allowing the traveller to walk a modest distance between bus stops. If there is no stage on the current service within 4km, method 1 is discarded and not used.

The same mapping logic is applied for method 2, using the first boarding of the day rather than the following boarding. The average point-to-point distance “walked” where methods 1 and 2 are applied is about 930m.

In the TfL study, two additional checks were performed following a match, other than checking the distance was within reasonable limits. Firstly, the time that would be necessary to walk between

services was compared with the actual time between boardings. It is not possible to do this with the Nexus data; the inaccuracy in estimating geographic locations is too large. Secondly, a check was carried out that the boarded service was travelling in the correct direction to permit the selected alighting point. This could in principle have been done here; however we would have to estimate direction of travel, which is extremely difficult given the data available.

In the Metro data, any attempt to assign the alighting point to be the same as the boarding point was automatically rejected. This is less simple in the bus data. As a fare stage represents more than one bus stop, journeys wholly within a fare stage are quite possible. The approach adopted was to accept "board=alight" when applying method 1, 2, 5 or 7, but to reject it in applying method 3 or 6.

As with the Metro data, the methods are applied in order; for example Method 3 would only be considered if methods 1 and 2 fail, either because the user has made no other trips that day, or because the match is more than 4km away from the closest point on the service boarded.

Method 3 performed quite well in the Metro data, but there was uncertainty whether it would be as valuable for bus travel. Accordingly, a manual review of the algorithm was undertaken. 20 records for which method 3 was applied if enabled were manually inspected and the estimates using method 3 or a subsequent (5,6 or 7) method were compared.

The results are below. For most records it was not possible to say confidently whether either method was correct or not, but nonetheless, the results strongly suggest that method 3 does not improve the quality of the estimates. Accordingly, it was removed from further analysis.

Table 7-1: Method 3 Manual Review

Method	Probably right	Probably wrong	Uncertain
3	15%	45%	40%
Other method (5,6,7)	25%	20%	55%

An overall summary of the process for the two 2016 datasets is presented below. We did not attempt to apply the full algorithm to the smaller 2015 datasets for 1 day and 1 service.

Table 7-2: Alighting estimation algorithm applicability using Methods 1, 2, 5, 6, 7

Method	First Week of May 2016 only	% of matched records	May+June 2016	% of matched records
1 (Different Service)	201,372	26%	1,274,827	26%
2 (Different Service)	94,365	12%	591,635	12%
1 (Same Service)	131,499	17%	837,815	17%
2 (Same Service)	98,709	13%	623,632	13%
5	76,411	10%	898,311	18%
6	28,671	4%	181,952	4%
7	142,484	18%	488,281	10%
All methods failed	88	0%	5	0%
Matched records	773,511	-	4,896,453	-
Total number of records	969,690		6,202,203	

The two datasets performed very similarly, but it is notable that method 5 is significantly more applicable with the full two months, as there are more records available to compare against and find a suitable boarding at the same stage.

About two thirds of matched records are infilled using methods 1 and 2, which we expect to be quite accurate (subject to uncertainty in the geographic matching process). With the full dataset, only 10% are infilled using method 7, which we expect to have relatively poor accuracy.

7.2 Trip lengths

It is not possible, as with the Metro data, to check the accuracy of the alogorithm fully at the record level. However, we do have access to the survey data formerly used for concessionary reimbursement, collected by surveyors on-board buses around Tyne & Wear. This does not have a sufficiently large sample to validate the algorithm at a boarding stage level, or even generally service, level, but it does allow us to compare against overall trip lengths. Due to the lack of complete routeing data, we have used point-to-point ("crow fly") distances from both sources; the survey supplies boarding and alighting bus stops by NaPTAN code.

We have also compared against the National Travel Survey, which is a household survey undertaken continuously by the Department for Transport. This collects travel diary information from individuals, and would collect all bus trips made by residents of a household in the course of the survey week. Unfortunately, we do not have access to this data at a detailed geographic level, so it was necessary to use data for the North-East of England region. Tyne & Wear represents about 40% of the population of the North-East, but it may be noted that the rest of the North-East is significantly more rural, so differences in trip lengths between the two may not be negligible.

NTS data of course represent complete "in-vehicle" distance. A factor of 1.3 was applied to convert this to crow-fly distance. This factor was based on transport modelling experience; it will be roughly appropriate.

Average trip lengths are shown in the table below.

Table 7-3: Average crow-fly distance in km from different sources (May+June 2016)

Source	Average distance trip Km
Survey	4.39
NTS North-east	5.89
Algorithm (Overall)	4.54
Algorithm (Method 1)	4.42
Algorithm (Method 2)	4.31
Algorithm (Method 5)	4.43
Algorithm (Method 6)	4.54
Algorithm (Method 7)	5.84
NTS NE in-vehicle distance (before crow-fly correction)	7.66

The algorithm overall performs well against the survey, with an average trip length only 3% higher than the survey. NTS trip lengths are significantly (about 33%) longer. In general we would, if anything expect NTS to be biased slightly high (as respondents are more likely to forget shorter journeys); it is also quite possible that typical bus journeys in the North-East are a bit longer than in Tyne & Wear, as the region as a whole is more rural.

All methods return broadly similar trip lengths, except method 7 which is significantly longer. Although we do expect method 7 to be the poorest, this does not *necessarily* imply that method 7 is wrong; it is possible that the trips for which others methods cannot be applied do tend to be genuinely longer than average.

Complete trip length profiles are shown in the figures below.

Figure 7-1 Distance profile Survey, Algorithm and NTS. 0-64km (May+June 2016)

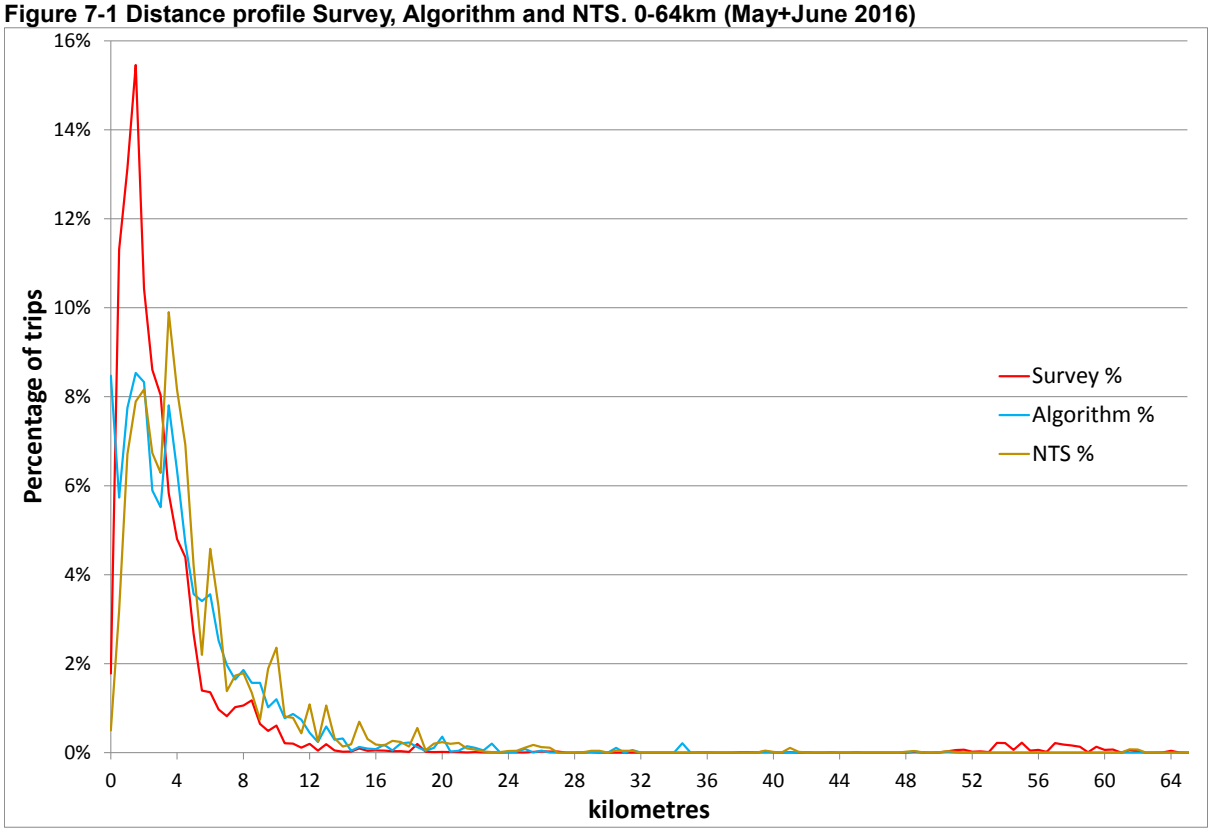
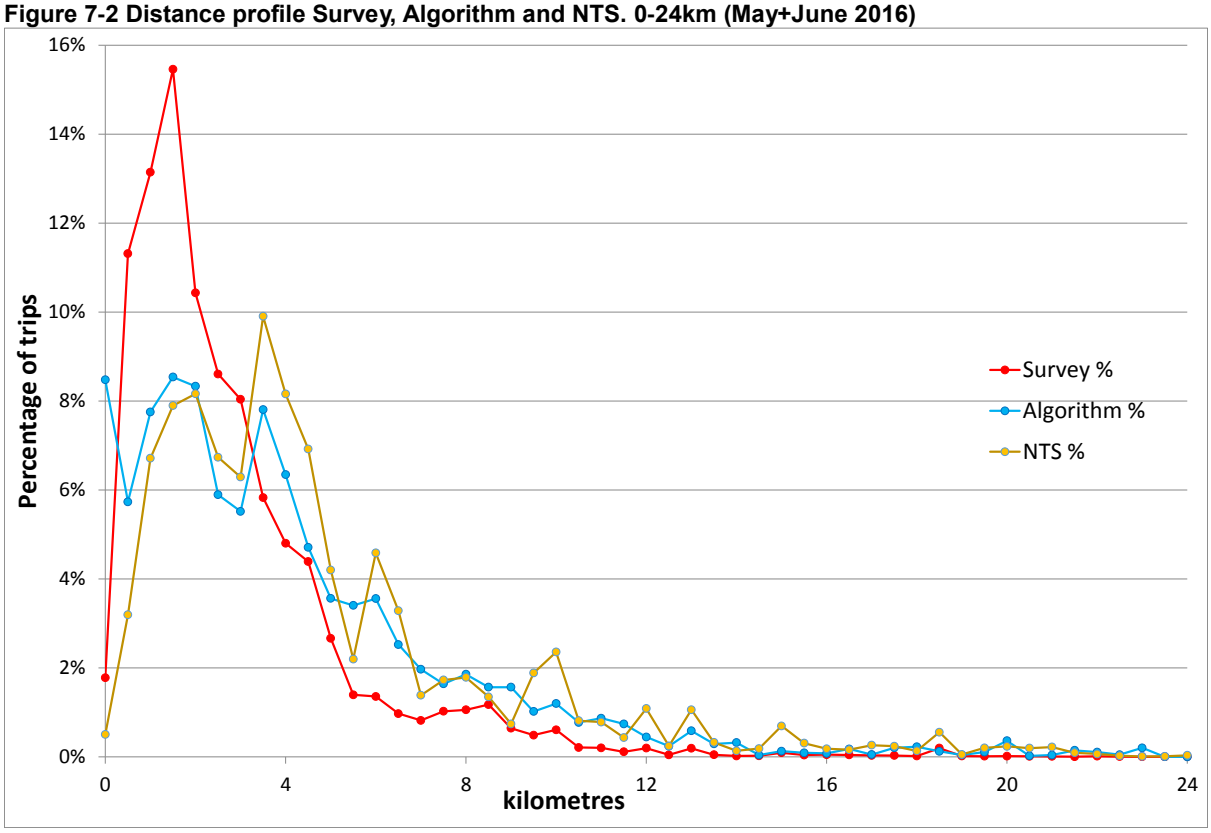


Figure 7-2 Distance profile Survey, Algorithm and NTS. 0-24km (May+June 2016)



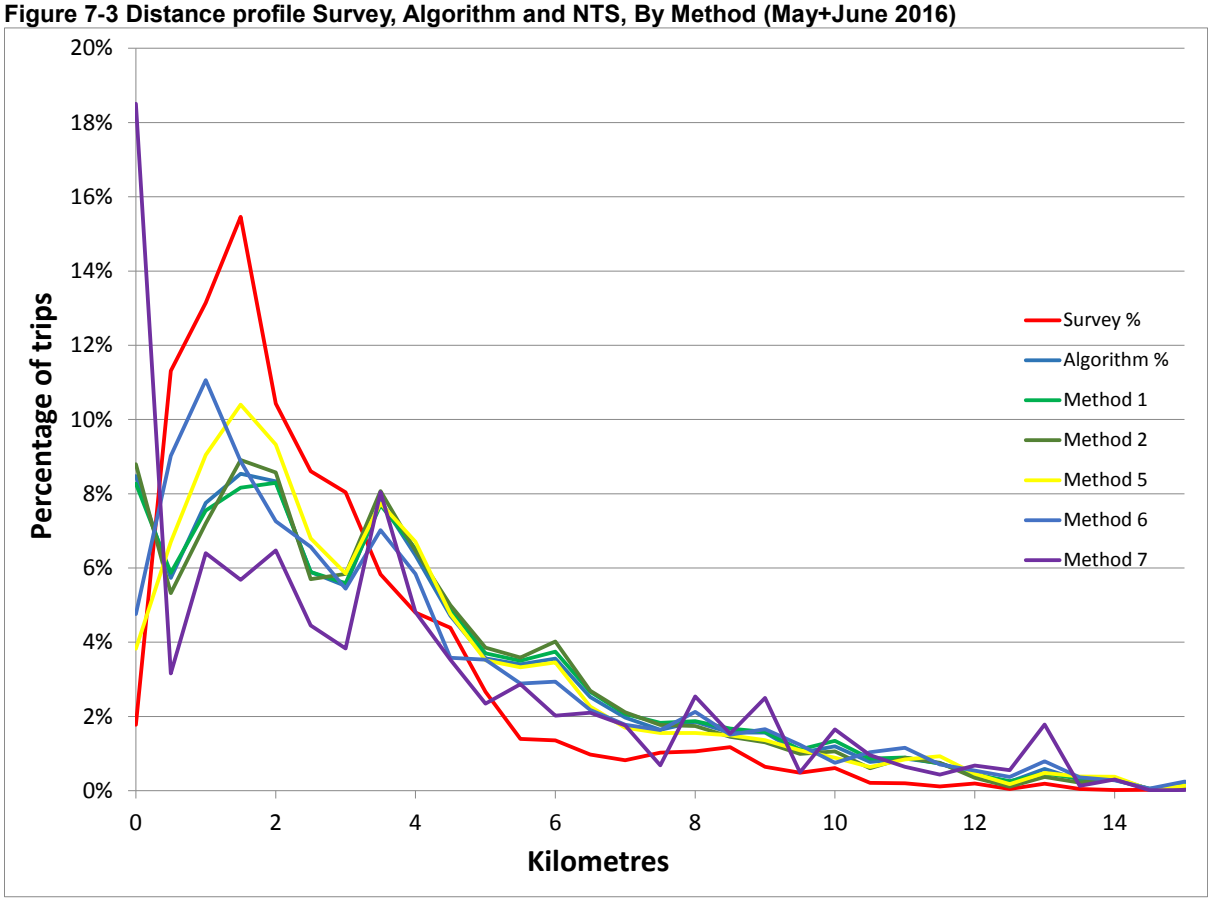
The algorithm does not match the survey data nearly as well when the complete pattern is considered. Curiously, it reproduces the NTS profile very much better, expect for the shortest trips where it lies between the two.

The algorithm clearly overstates trips in the shortest (0-0.5km) band. This is because it occasionally selects an alighting fare stage that is the same as the boarding. This could be correct in some cases, but it seems likely that it is wrong in at least a substantial proportion.

The survey data have significantly more trips in a 0.5-3.5 km band than either the algorithm or NTS. It must be noted that none of the three curves is a perfect representation of reality, and the survey data may be biased as well.

The trip lengths profiles are shown by method below.

Figure 7-3 Distance profile Survey, Algorithm and NTS, By Method (May+June 2016)



Most of the methods behave quite similarly. Methods 6 and 7 are notable for being somewhat different.

We can also compare the algorithm and the survey by operator. Unfortunately the May and June 2016 ticket data from Nexus only categorises operators into four groups; there is more detail in the survey.

Table 7-4: Average trip distance in km for different operators (First Week of May Only)

Average distance trip Km	Survey	Algorithm	Difference
Operator Group 1	6.91	6.55	-5%
Operator Group 2	5.61	5.04	-10%
Operator Group 3	3.09	3.68	+19%
Operator Group 4	2.76	1.75	-37%

It is clearly that the algorithm broadly captures the variation by operator, although it does significantly understate the trip length for the small operators. Full trip length distributions are shown below; as before these do not match very closely. We cannot extract NTS data by operator.

Figure 7-4 Distance profile Survey and Algorithm for Operator Group 1 buses (First Week of May)

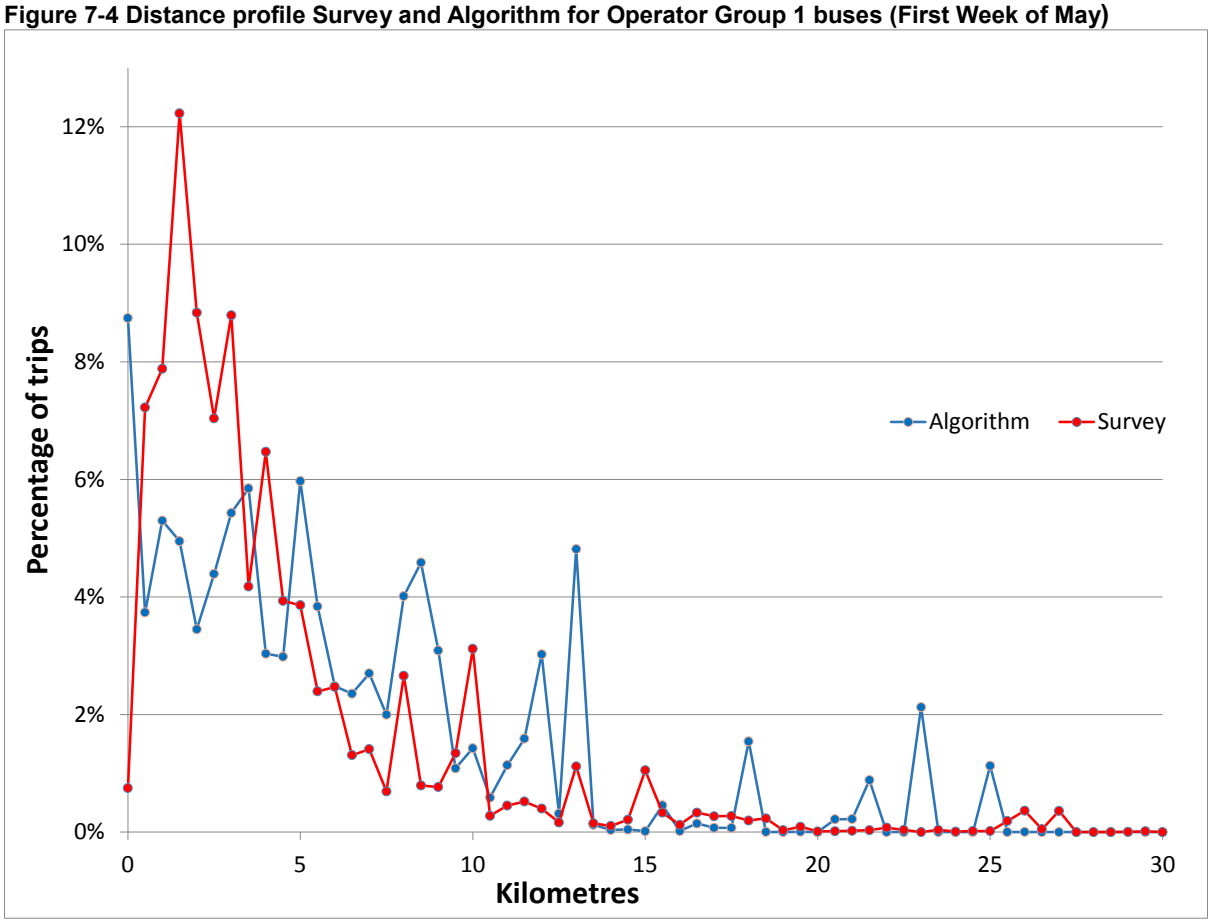


Figure 7-5 Distance profile Survey and Algorithm for Operator Group 2 buses (First Week of May)

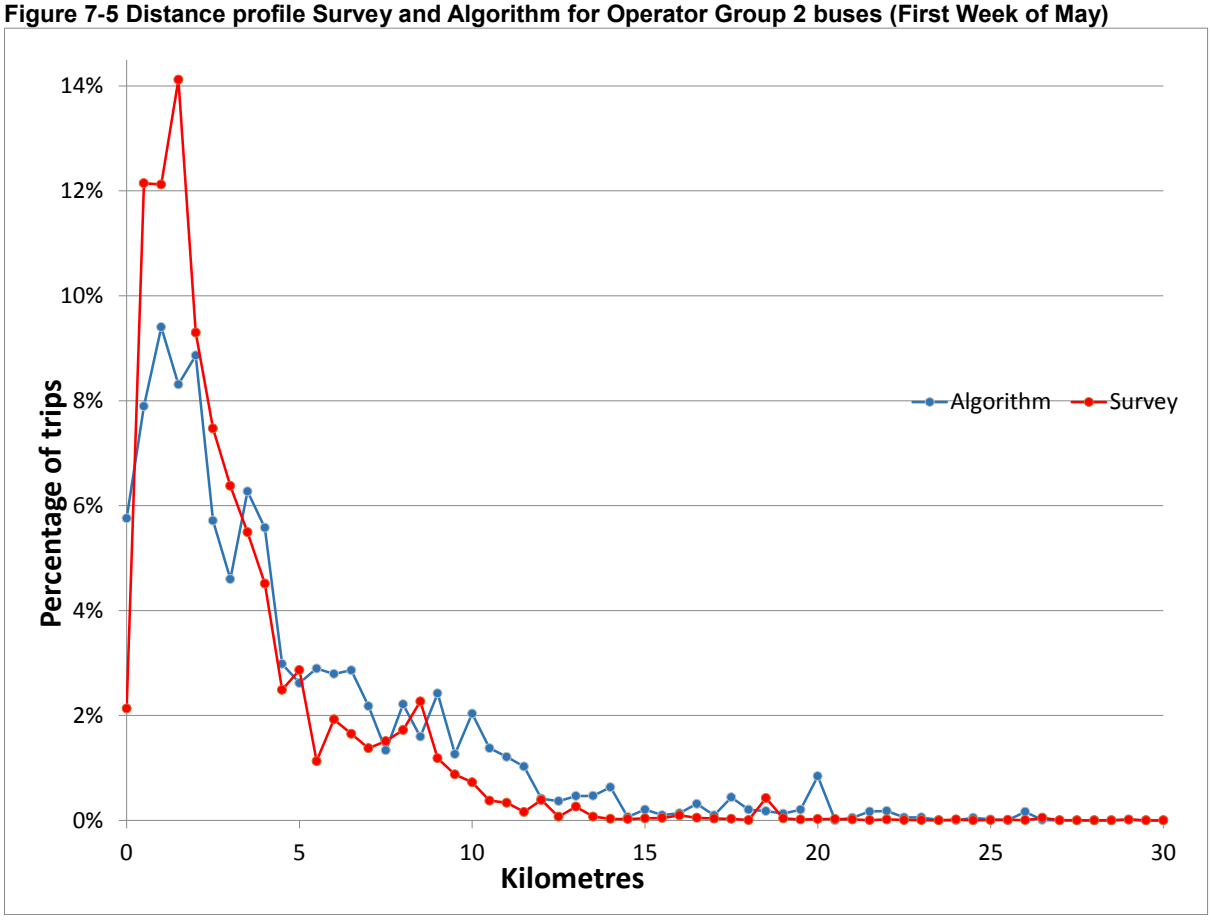


Figure 7-6 Distance profile Survey and Algorithm for Operator Group 3 buses (First Week of May)

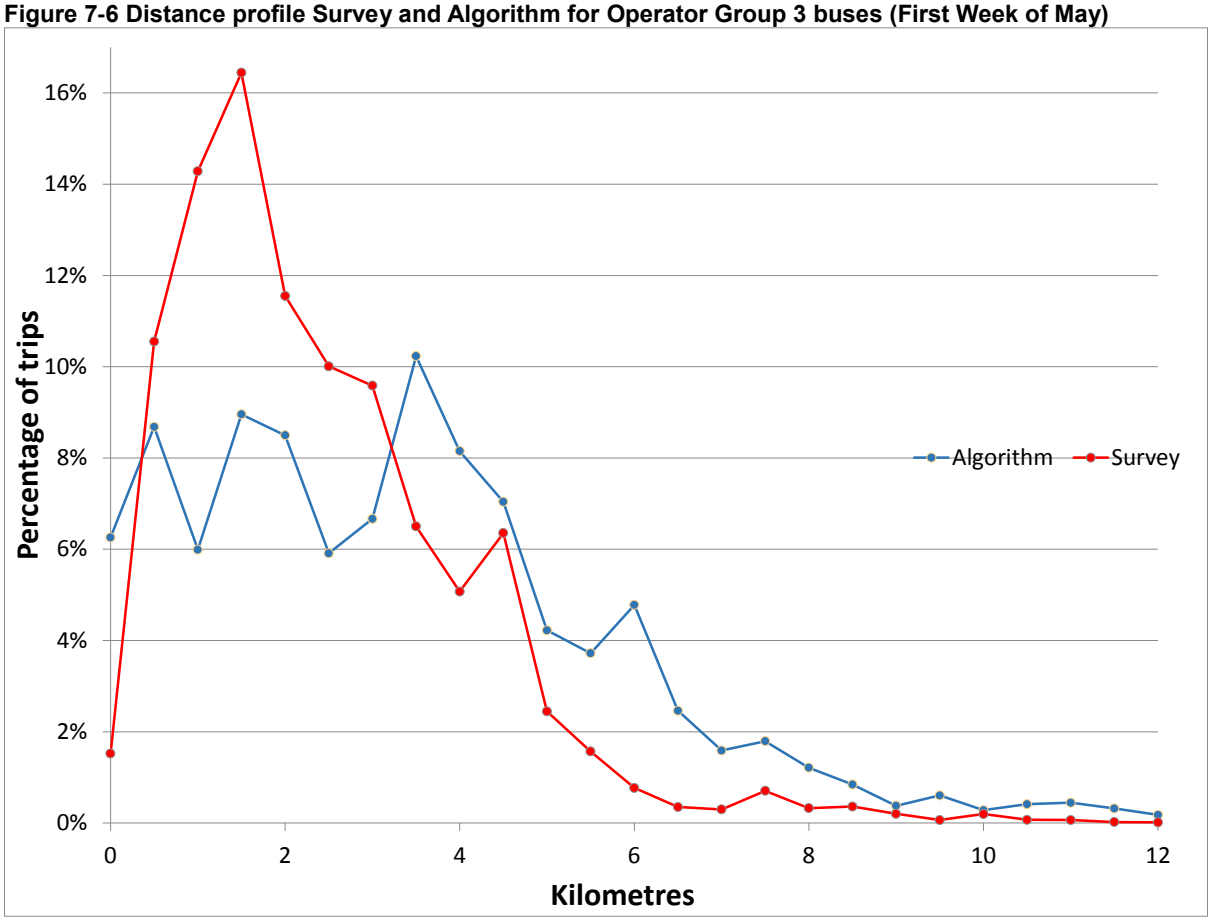
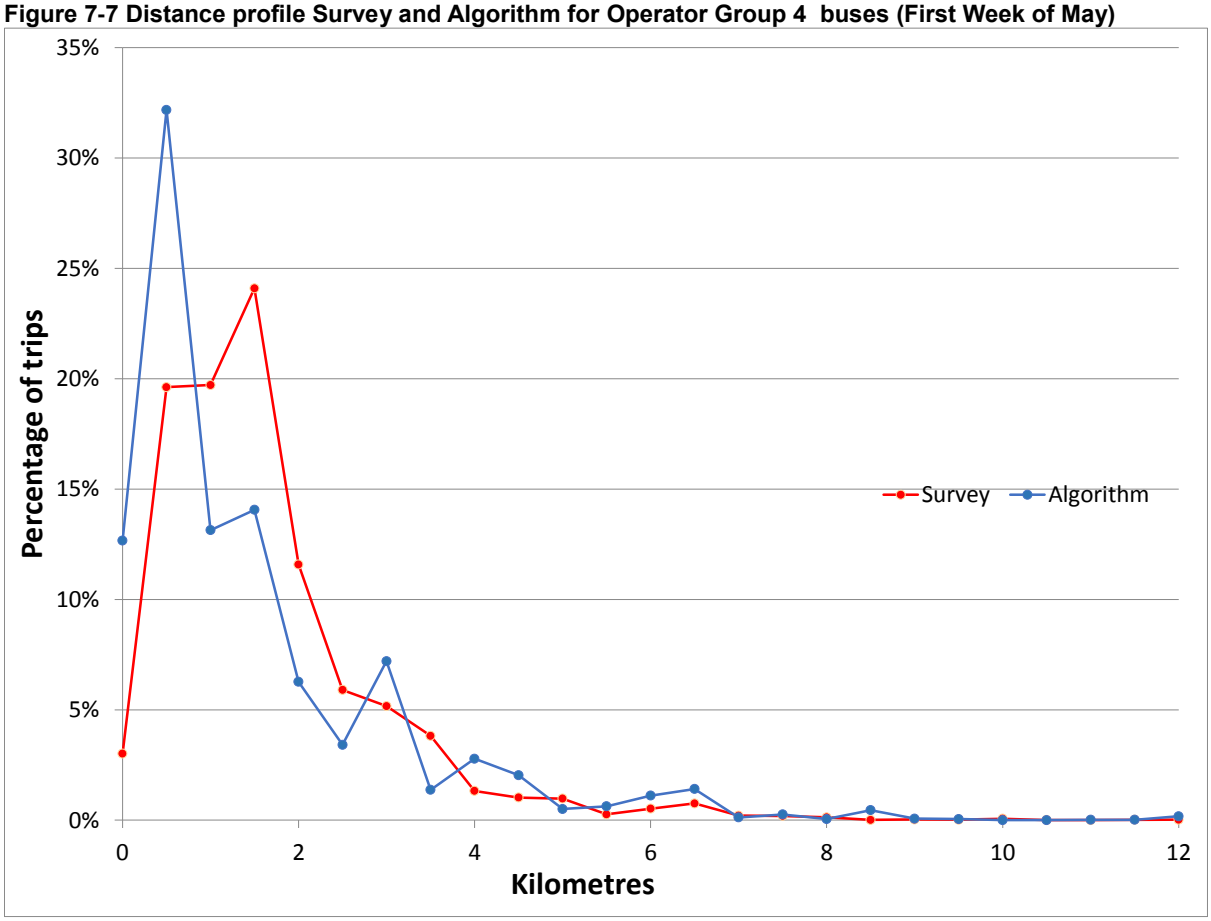


Figure 7-7 Distance profile Survey and Algorithm for Operator Group 4 buses (First Week of May)



7.3 Algorithm

As with the geographic matching process, this algorithm is the result of a certain amount of testing and validation time, but could almost certainly be improved. The approach is set out as a series of steps. Each journey is considered separately. A “traveller” is assumed to be identified by a smartcard. The process does not consider one smartcard being used by more than one person, or (more likely) one person using more than one smartcard. “Journey” and “trip” are used below interchangeably with “data record”.

An “unmatched” journey is one where the service boarded could not be fully reconciled with a timetable, so we don’t know where the stages are located. The process is not applied to unmatched journeys, but they are not deleted from the dataset for the purpose of analysing other records.

1. **(Method 1)** If the journey is the only one the traveller made on that day, proceed to step 3. Otherwise, if it is the last trip the day, proceed to step 2. Otherwise, consider the following trip made by the user on that day. If this is “unmatched”, proceed to step 3. Otherwise, find the boarding point of this following journey and find the closest stage on the service the current journey uses. If this is within 4km, choose this closest stage as the alighting point and stop. Otherwise proceed to step 3.
2. **(Method 2)** Consider the first trip made by the user on this day. If this is “unmatched”, proceed to step 3. Otherwise, find the boarding point of this first trip, and find the closest stage on the service the current (final of the day) journey uses. If this is within 4km, choose this closest stage as the alighting point and stop. Otherwise proceed to step 3.
3. **(Method 5)** Identify all other boardings this user makes at the same boarding stage on the same service throughout the dataset. If there are any such boardings for which method 1 or 2 can successfully be applied, select the most common alighting chosen in such cases (ties are broken by selecting the first record on the list) and stop. Otherwise proceed to step 4.
4. **(Method 6)** Identify all other boardings this user makes on the same service throughout the dataset. If there are any such boardings for which method 1 or 2 can successfully be applied, select the most common alighting chosen in such cases (ties are broken by selecting the first record on the list). If this is the same stage as the boarding point of the current service, or no such records exist, proceed to step 5. Otherwise select this most common alighting as the alighting for the current trip and stop.
5. **(Method 7)** Identify all other boardings made at this boarding stage on this service throughout the whole dataset, across all smartcards. If there are any such boardings for which method 1 or 2 can successfully be applied, select the most common alighting chosen in such cases (ties are broken by selecting the first record on the list). If no such records exist, stop. Otherwise select this most common alighting as the alighting for the current trip and stop.

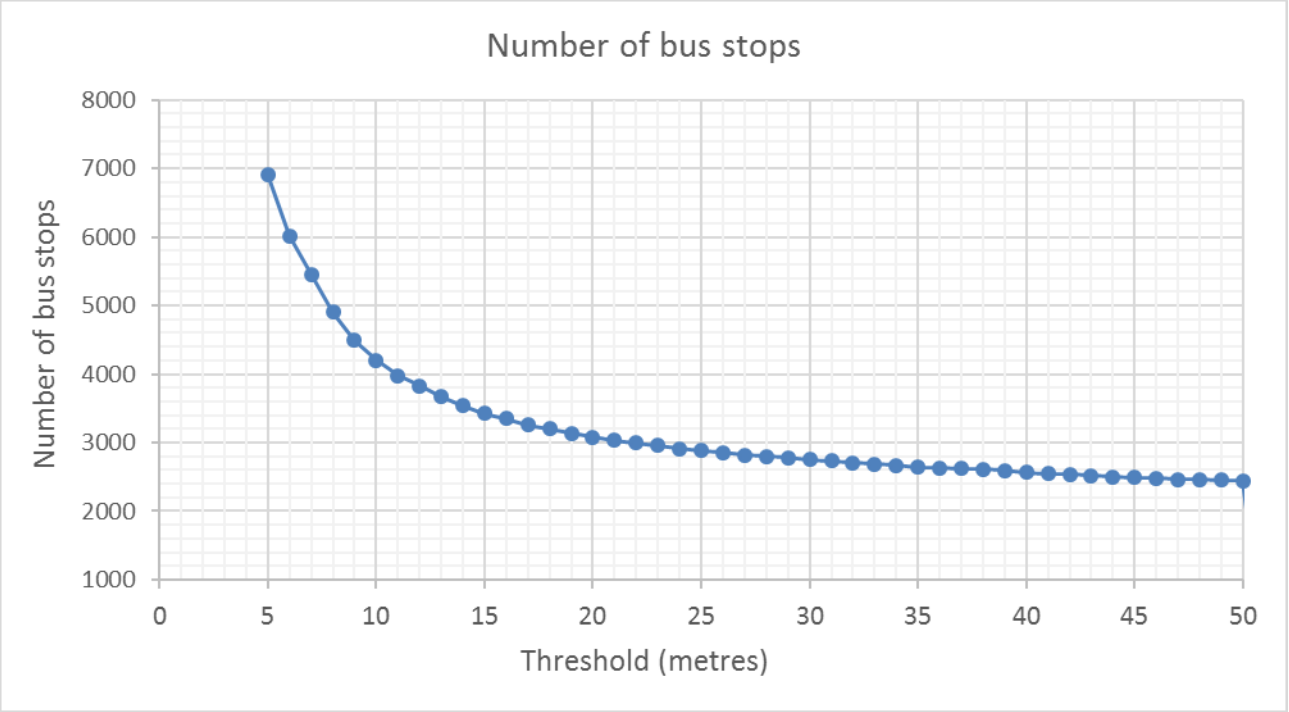
8. Geographic Location for Ticketer data

As noted in chapter 3.4, the Ticketer data contain coordinates (Latitude and Longitude) for each recorded boarding. These are subject to a small amount of error (from the GPS system) and variation (jn precisely where the bus stops when serving a given bus stop). Accordingly, we clustered GPS locations where they were very close to try to identify bus stops.

The clustering threshold was set to 30 metres, hence if two boardings were less than 30 metres apart they were taken as occurring at the same bus stop. The 30 metres threshold was chosen based on professional judgement and clustering optimization using the curve shown on

Figure **8-1** which is the curve described by the function, number of bus stops to cluster threshold used.

Figure 8-1 : Number of bus stops variability to threshold chosen.



Once the bus stops data set was built through this clustering mechanism, we then attempted to map the derived bus stops coordinates to NaPTAN bus stops. The percentage of bus stops for which a NaPTAN bus stop was found within 60 metres was 89%; this represents 96% of passenger boardings as unmatchable bus stop clusters are significantly less heavily used than average. The mean distance between the clustered bus stops location and the matched NaPTAN location is 12 metres.

9. Alighting Estimation- Ticketer data

9.1 Ticketer West Yorkshire data

All the concepts explained in chapters 4, 5 and 7 of this report apply to the Ticketer data process with minor refinements in some of the definitions as outlined below..

The particularities of the Ticketer data (specifically the very detailed stop segregation and relatively small size of dataset) meant that a new method, Method 9, had to be introduced. Method 9 is as 7 and 8 a “full dataset alighting estimation” method. Which means that it is applied were none of the previous methods returned a plausible answer or were applicable and it’s based on previous infills made using methods 1 and 2. Method 9 infills the destination with the most common destination for this service, regardless the boarding point.

The Ticketer data, like the Nexus data, represent bus boardings, which means that the user is recorded as boarding a specific service at a specific bus stop. This particularity, constrains the number of plausible alighting points (relative to the Metro dataset used for Tyne & Wear) because the infilled alighting bus stop has to be a bus stop that is served by the specific service that the user boarded.

Based on the experience gathered on the Nexus Bus data alighting estimation it was decided to implement only the methods that were eventually approved on the mentioned data set i.e. Methods 1, 2, 5, 6 and 7. As mentioned before method 9 was included to complete the infilling.

In constrast to the other bus datasets, the mapping was undertaken at a bus stop (strictly, cluster of bus stops within 30 metres) level, rather than by fare stage. This should improve the geographical accuracy of the process significantly. In constrast to the process for the Metro data, there was no need here to begin by trying to geographically reference the fare stages, as we already had coordinates for the boardings points.

Brief definition of the methods:

Method 1: Next boarding point on the same day.

Method 2: First boarding point of the current day (applied to the last trip of the day only).

Method 5: Most commonly chosen alighting for the current boarding stage by this user

Method 6: Most commonly chosen alighting by this user for this bus service.

Method 7: Most commonly chosen alighting for the current boarding stop and service across all users.

Method 9: Most commonly chosen alighting for all users for this bus service.

If two consecutive boardings on the same day (or last boarding to first of the day) are closer than 200 metres, algorithm 1 is not applied; similarly algorithm 2 is not used if the first boarding of the day and the last boarding are within 200 metres of each other..

Bus trips under 100 metres weren’t considered.

Method 1 is simple to apply if the two relevant journeys use the same bus service number. In this case the algorithm works as it did with the Metro data.

If the next boarding is made on a different bus route the following approach is taken.

The next boarding’s geographical bus stop coordinates are inspected. If they are present, the algorithm tries to find the closest bus stop to this on the current service, within 200 metres. If there is no stage on the current service within 200 metres, method 1 is discarded and not used.

The reason for using 200 metres comes from statistical analysis of the results using 4km (the maximum distance walked used for Nexus data). As it is shown on

Figure 9-1, any walked distance greater than 192 metres was identified as an outlier.

Figure 9-1: Box and whisker plot of walked distance for methods 1 and 2 when the threshold was set to 4km

The same mapping logic is applied for method 2, using the first boarding of the day rather than the following boarding. The average point-to-point distance walked, where methods 1 and 2 are applied is about 15m.

As explained in chapter 7.1, in the TfL study, two additional checks were performed following a match, other than checking that the distance was within reasonable limits (which we do here with a 200m threshold). Firstly, the time that would be necessary to walk between services was compared with the actual time between boardings; With a maximum walked distance of 200 metres between bus stops and an average of walked of 15 metres, the time to walk from one stop to another would always less than 2.5 minutes, that time is too small for the level of accuracy that the estimating method would provide time-wise.

Secondly, a check was carried out that the boarded service was travelling in the correct direction to permit the selected alighting point. This could in principle have been done here; however, we would have to estimate direction of travel, which is extremely complicated given the data available, although it is technically possible.The bus journey data include a “direction” column, but as this just lists “inbound”, outbound”, “clockwise” or “anticlockwise”, it is difficult to map to bus stop coordinates. A direction check was not applied for our analysis of any previous datasets either.

Method 5 infills the most common alighting bus stop for the boarding stop for the boarded service, for the user making the trip. This method doesn’t need any other constraints, because the fact that this trip was infilled previously using methods 1 or 2 makes the trip plausible and likely.

Method 6 infills the most common alighting bus stop for the boarded service, for the user making the trip. It doesn’t take into account what the boarding stop is. Therefore, is more error prone than method 5. The trip length distribution of this method has a different shape to the rest of them, as will be shown further down in this report which implies there is potentially room for refinements of it.

Method 7 and 9 are equivalent to 5 and 6, but the OD matrix they use to infill the alighting is made up of trips for all the users, instead of, being made up of user specific trips.

Table 9-1 shows the applicability of the methods used.

Table 9-1: Alighting estimation applicability over the Ticketer data using methods 1, 2, 5, 6, 7 and 9.

Method used	Number of estimations	% of estimations to valid records
1 (Same service)	8686	19.1%
1 (Different service)	2881	6.3%
2 (Same service)	7853	17.2%
2 (Different service)	1699	3.7%
5	3608	7.9%
6	1134	2.5%
7	16948	37.2%
9	2515	5.5%
All methods failed	230	0.5%
Matched records	**45324**	-
Valid records	**45554**	-
All records	**45838**	-

Methods 1 and 2, the most accurate, infill 46% of the trips. In the Nexus data this percentage was 68%. The poorer performance for the algorithm on Ticketer data is due to the fact that the dataset contains many instances in which there is only one trip a day for the user as shown in Figure 3-1. This is because the dataset include only minor operators and is not a complete representation of travel within the area.

This underscores the need to have complete datasets in order to apply this infilling methodology reasonably robustly.

Methods 5 and 6, based on the specific user travel patterns, infilled 10%. This leaves the remaining 43% to be infilled using methods 7 and 9. This is probably reasonable for the purpose of obtaining reasonable overall distribution of travel, for example trip lengths, but obviously less desirable.

9.2 Trip lengths

For the Ticketer data AECOM wasn't provided with any validation data that we could use to estimate the level of accuracy the algorithm achieved. Nevertheless, as was done for the Nexus data, it is possible to compare the trip lengths distribution that the different methods provide and compare this with National Travel Survey (NTS)data.

The average trip length by method is shown in Table 9-2.

Table 9-2: Average crow-fly distance trip length in km from different sources

Source	Average distance trip Km
NTS Yorkshire	5.48
Algorithm (Overall)	4.08
Algorithm (Method 1)	4.39
Algorithm (Method 2)	4.26
Algorithm (Method 5)	2.90
Algorithm (Method 6)	3.35
Algorithm (Method 7)	4.06

Algorithm (Method 9)	4.15

As can be observed in the table above, the algorithm average trip length is 4.08 km, a distance shorter than the average relevant bus trip length for in Yorkshire.

The trip length here named “NTS Yorkshire” represents the trip length for bus trips in Yorkshire made by concessionary cards users after 10am 2010-2016. This data was provided in miles. On the conversion from miles to kilometres a factor of (1/1.3) was used to account for the fact that the travelled distance in NTS is not crow-fly as the algorithm is. This factor was based on transport modelling experience; it will be roughly appropriate. As with the Nexus data, the algorithm returned average trip lengths shorter than NTS.

Figure 9-2 shows the distance profile for the different methods used in the infilling algorithm. The algorithm as a whole infills 13% of the trips in the band from 0 to 1km, peaks in the band of 1 to 2km with 24% and then declines quasi-steadily being 16% for the band 2 to 3 km, 12% for the band 3 to 4, 10% from 4 to 5 and so on. Methods 1, 2 and 7 make up most of the infilling hence their high similarity to the distribution of the algorithm as a whole. Method 5 peaks in band 1 to 2 with 31%, the highest percentage observed in the trip length distribution table. Method 6 as opposing to the other methods does not peak in the band 1 to 2km. Method 9 infills around 18% of the trips for each of the three first bands, 0 to 1, 1 to 2 and 2 to 3, hence the poorer profile it shows between 0 and 3km.

Figure 9-2: Distance profile for the algorithm by method. 0-24km

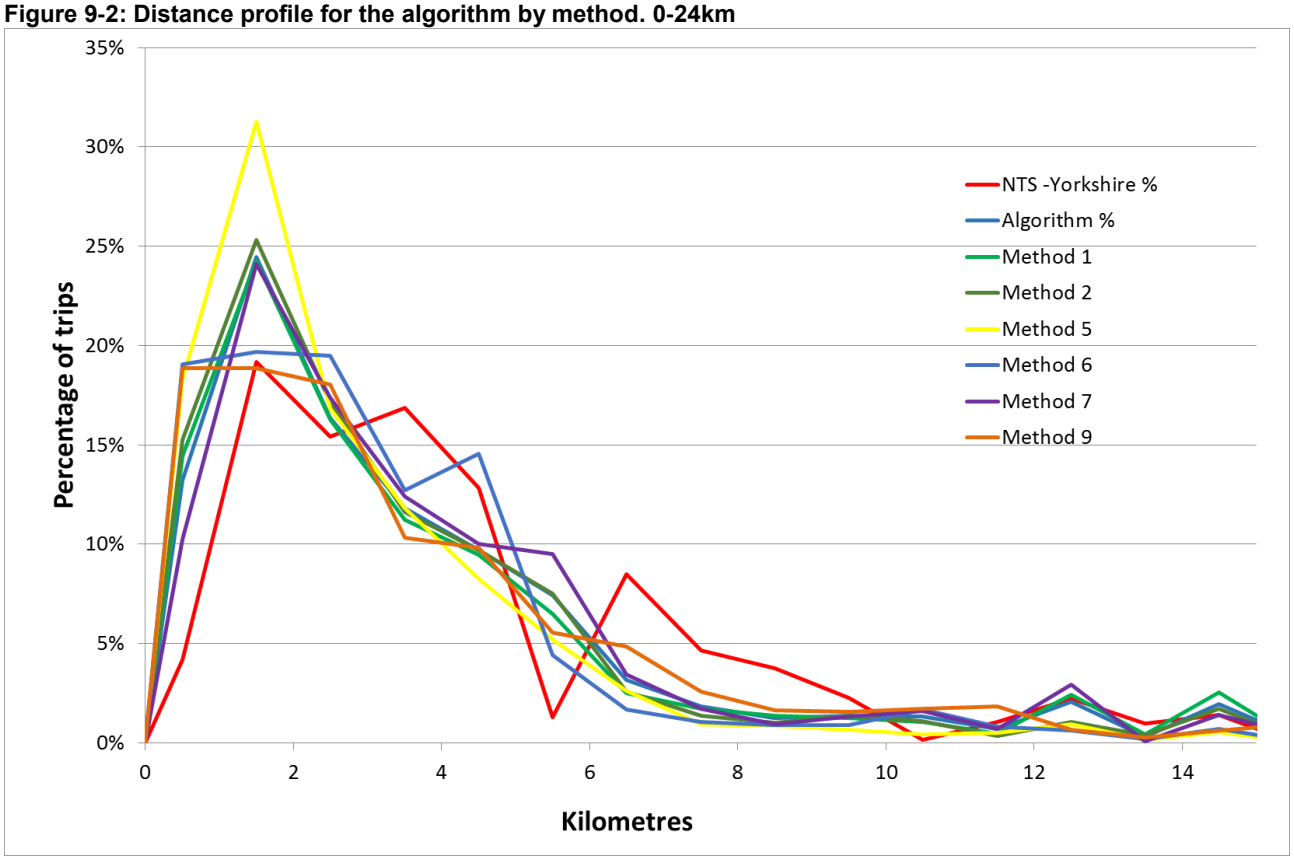
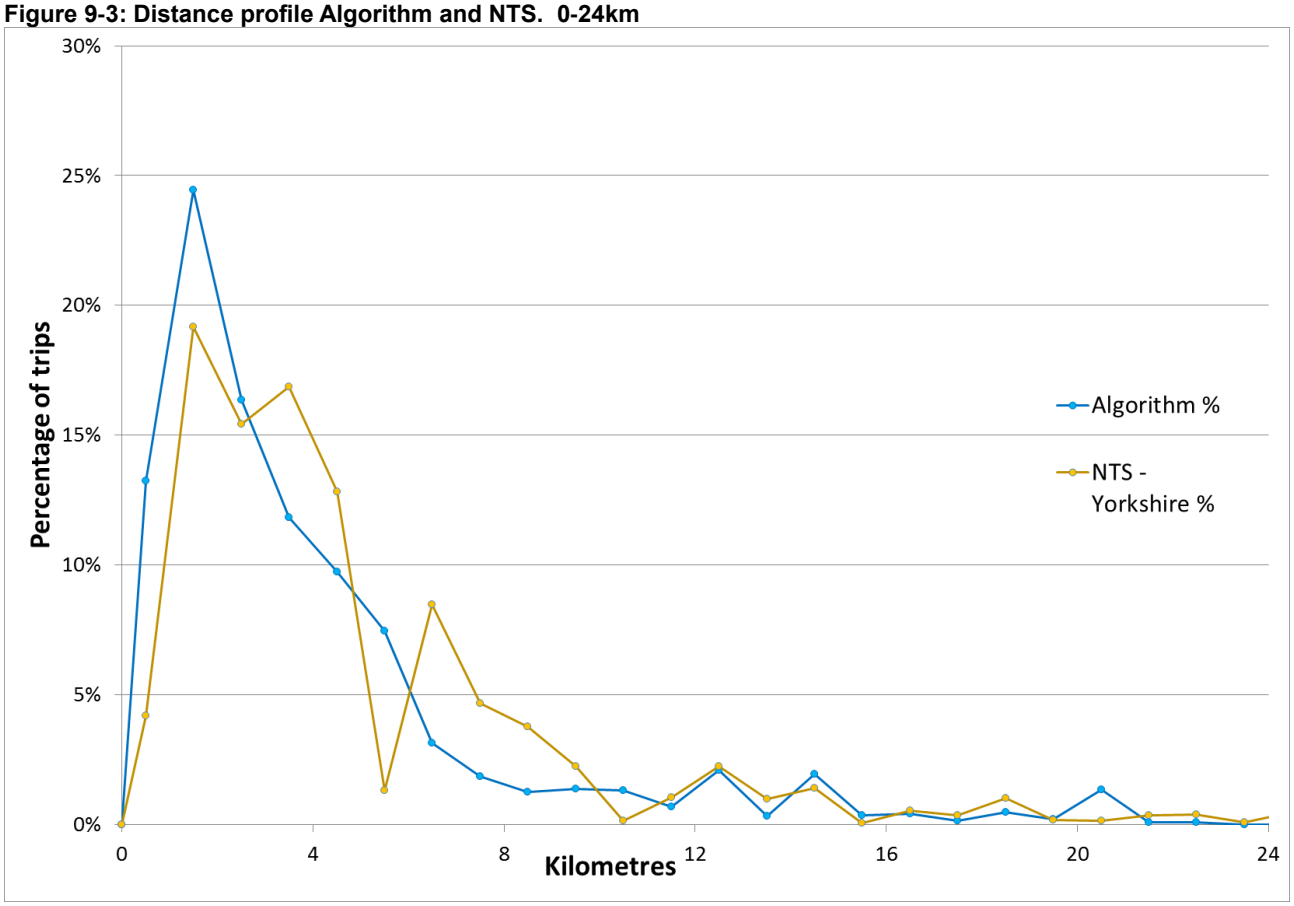


Figure **9-3** shows the trip length distribution of the algorithm and the bus trip length distribution from the National Travel Survey for concessionary card holders trips after 10am in Yorkshire. The algorithm roughly replicates what the only validation available does. The Peak for both curves is clearly in between 1 and 2 km. There is some variation on the NTS Yorkshire data between 2 and 7 km, with an abrupt drop on the number of trips between 5 and 6km followed by a raise between 6 and 7. These variation is unlikely to be genuine in the area and may be due to the distortion of introduce during the conversion of the from a mile bandwidths to a 0.2km bandwidths.

The fact that the dataset studied has its particularities, such as leaving the trips made by the larger transport providers aside, must be considered a possible reason for the deviation between the two curves.

Figure 9-3: Distance profile Algorithm and NTS. 0-24km



Because NTS data is provided in miles bandwidths there is a distortion the data is converted to 0.2km bandwidths.

Figure 9-4 shows the same concept as

Figure **9-3** but in miles.

Figure 9-4: Distance profile Algorithm and NTS. 0-24 miles

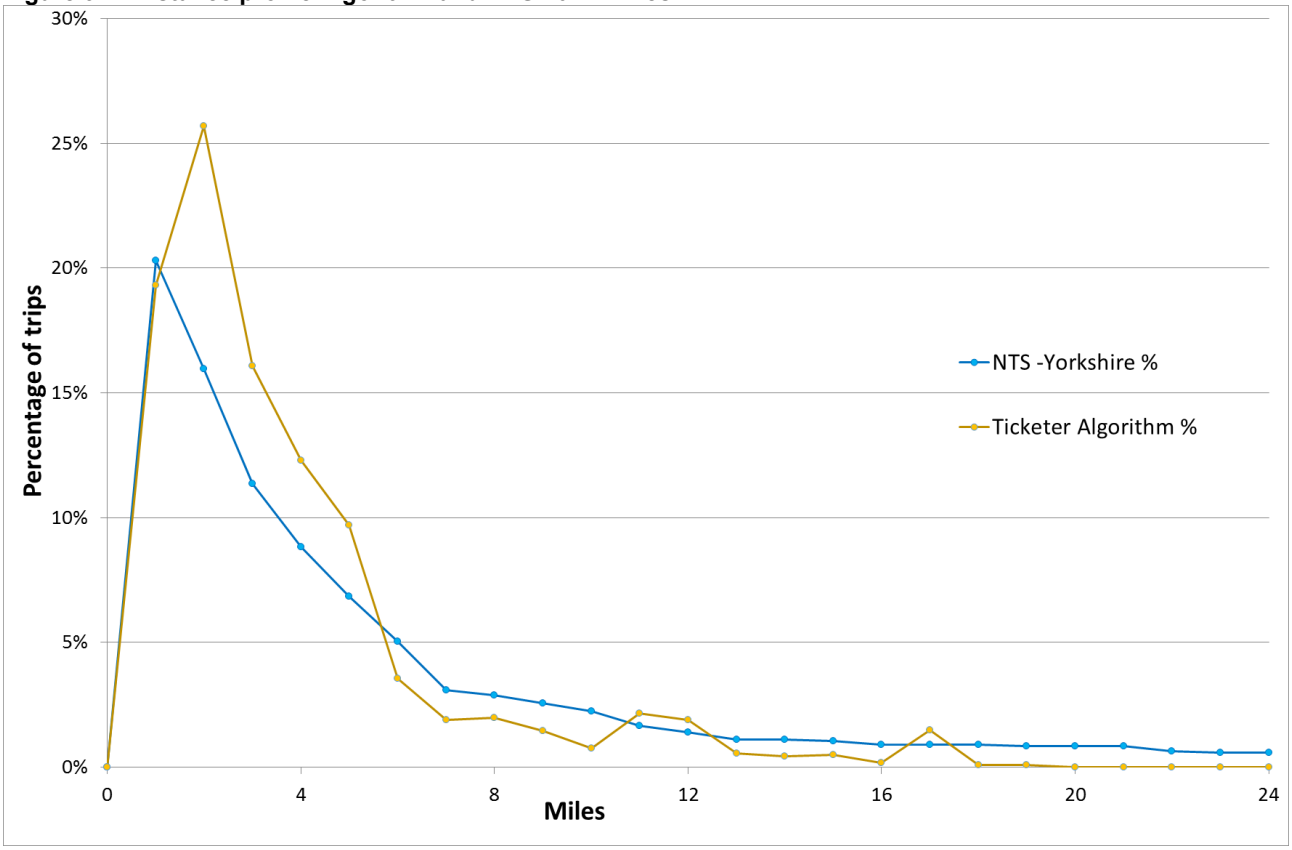
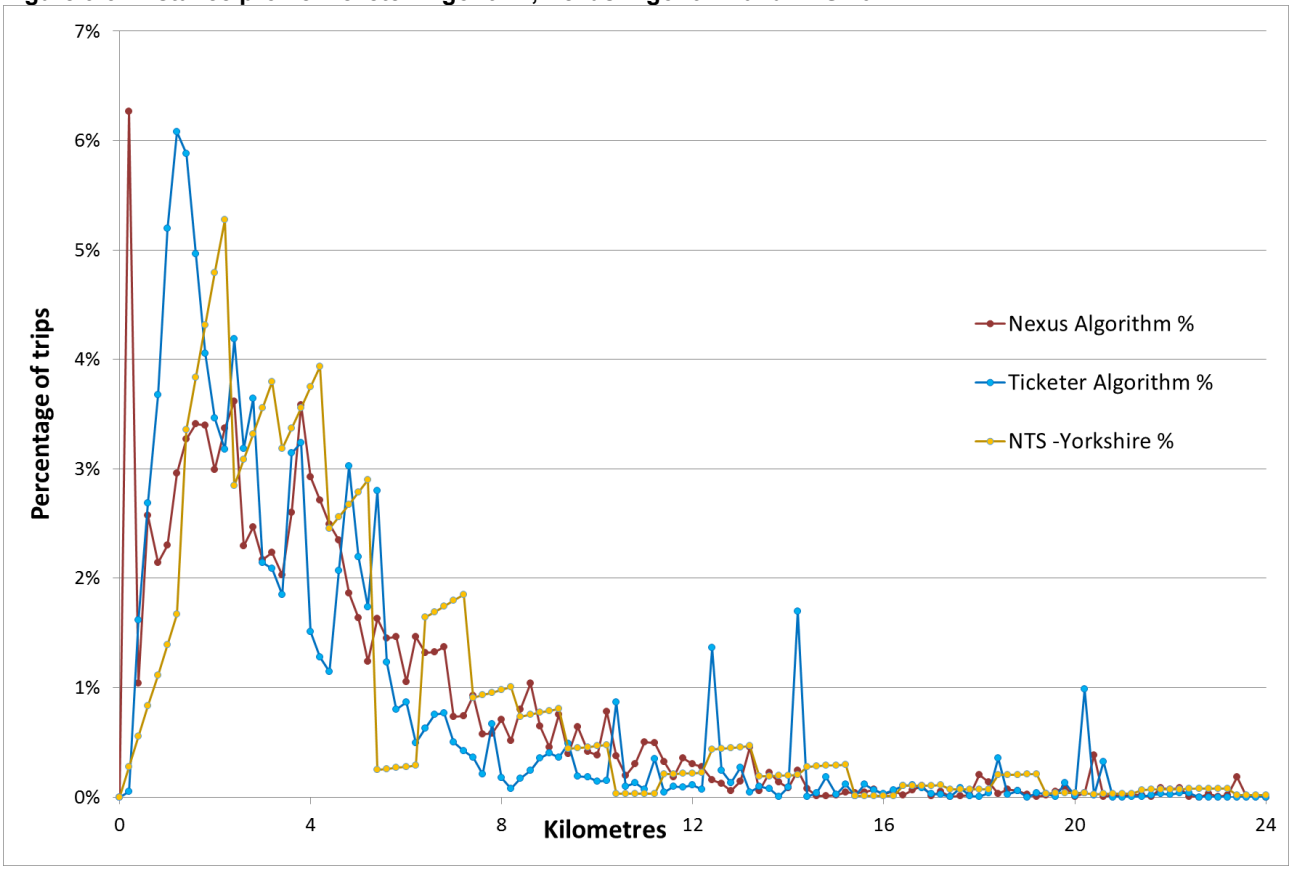


Figure **9-5** shows how the trip lengths infilled by the Ticketer algorithm compare to those infilled by the Nexus bus data algorithm. As can be observed, the Nexus algorithm infilled many short distance trips, possibly due to the fact that the algorithm worked at a stage level rather than a bus stop level. These would mean that trips within the same fare stage are taken as 0 km trip length. For both datasets, the algorithm and NTS, the average trip length is lower in the case of the Yorkshire study.

Figure 9-5: Distance profile Ticketer Algorithm, Nexus Algorithm and NTS. 0-24km



9.3 Algorithm

The approach followed for the Ticketer data is very similar to that of the Nexus data. A “traveller” is assumed to be identified by a smartcard. The process does not consider one smartcard being used by more than one person, or (more likely) one person using more than one smartcard. “Journey” and “trip” are used below interchangeably with “data record”.

If two consecutive boardings on the same day (or last boarding to first of the day) were closer than 200 metres, algorithms 1 and 2 were not used.

Bus trips under 100 metres aren’t considered.

1. **(Method 1)** If the journey is the only one the traveller made on that day, proceed to step 3. Otherwise, if it is the last trip the day, proceed to step 2. Otherwise, consider the following trip made by the user on that day. Find the boarding point of this following journey and find the closest bus stop on the service the current journey uses. If this is within 200 metres, choose this closest bus stop as the alighting point and stop. Otherwise proceed to step 3.
2. **(Method 2)** Consider the first trip made by the user on this day. Find the boarding point of this first trip, and find the closest stage on the service the current (final of the day) journey uses. If this is within 200 metres, choose this closest stage as the alighting point and stop. Otherwise proceed to step 3.
3. **(Method 5)** Identify all other boardings this user makes at the same boarding stage on the same service throughout the dataset. If there are any such boardings for which method 1 or 2 was successfully applied, select the most common alighting chosen in such cases (ties are broken by selecting the last record on the list) and stop. Otherwise proceed to step 4.
4. **(Method 6)** Identify all other boardings this user makes on the same service throughout the dataset. If there are any such boardings for which method 1 or 2 was successfully applied, select the most common alighting chosen in such cases (ties are broken by selecting record furthest away from the boarding point, so long as is no further away than 5km, if both were further than 5km then the last record on the list is taken). If this is the same bus stop as the boarding point of

the current service, use the second most common alighting for the user, for the service.If no such records exist, proceed to step 5. Otherwise select this most common alighting as the alighting for the current trip and stop.

5. **(Method 7)** Identify all other boardings made at this boarding stage on this service throughout the whole dataset, across all smartcards. If there are any such boardings for which method 1 or 2 was successfully applied, select the most common alighting chosen in such cases (ties are broken by selecting the last record on the list). If no such records exist, proceed to step 6. Otherwise select this most common alighting as the alighting for the current trip and stop.
6. **(Method 9)** Identify all other boardings made on this service throughout the whole dataset, across all smartcards. If there are any such boardings for which method 1 or 2 was successfully applied, select the most common alighting chosen in such (ties are broken by selecting record furthest away from the boarding point, so long as is no further away than 5km, if both were further than 5km then the last record on the list is taken). If this is the same bus stop as the boarding point of the current service, use the second most common alighting for the user, for the service. If no such records exist, stop. Otherwise select this most common alighting as the alighting for the current trip and stop.

10. Study Conclusions

10.1 Tyne & Wear, Nexus Data

Two main datasets of public transport passenger boardings have been analysed in detail. Both come from Nexus, the public transport authority for Tyne & Wear; one contains Metro travel and one bus travel. Both include all travel in Tyne & Wear over a long period. The Metro data contain alighting points. This renders the study algorithm unnecessary, but allows the Metro data to be used to validate the algorithm.

The bus data unfortunately specified boarding location using un-keyed “stage” numbers that we cannot relate to any other data. It was thus necessary to develop a process to estimate stage number locations (in terms of British National Grid coordinates). This generates an estimate for 80% of records, and is estimated to be roughly 80% accurate (within a few kilometres) for those records for which it generates an estimate.

An algorithm has been developed, with the help of the Metro data, to estimate alighting points given only boarding points. This is described in full in section 7.3. This algorithm was not applied to the 20% of total bus records that couldn’t be matched to geographic locations.

For around two-thirds of matched trips (in all data sources) a method called “reverse journey matching” is used. This attempts to use the boarding point of a subsequent (or first) trip on the same day to identify the alighting point. This is believed to be generally very accurate (around 90%).

For most remaining trips, methods involving other travel of the same user, not necessarily on the same day, are used. These are not precise at a record level, but are believed to be reasonable estimate of general distribution.

For about 10% of trips, a general infill method is used, using travel of all passengers. This is not believed to perform very well in general.

The algorithm generates average trip lengths that compare well (within 5%) with on-bus survey data from 2015. It also reproduces variation in trip lengths by bus operator fairly well. However, the full trip length distribution does not compare as well with the on-bus survey, although it matches the National Travel Survey (a DfT household survey) much better. All three data sources have advantages and disadvantages.

The algorithm is known to overstate very short (under 500m) bus trips. This could probably be improved.

In general, further work could be done to improve the algorithm, although as always there would be diminishing returns; the better the algorithm becomes, the more effort would be required to make a given improvement.

10.2 West Yorkshire – Ticketer data

The dataset provided contained a week of boarding data in November 2017 for smartcard concessionary users that boarded services operated by small operators.

The data provided had coordinates attached to them so there was no need to proceed to a geographical location of the boarding, although the boardings needed clustering into bus stops.

The algorithm developed for the Tyne & Wear study was adapted to work for the new data set. The particularities of the data, especially the scarcity of smartcards with more than one trip a day, meant that methods 1 and 2 were not as widely applicable as they were for the Tyne & Wear case and a new method of alighting estimation had to be introduced. It is therefore important to note that complete datasets of bus trips are very helpful in robust application of the algorithms and smaller datasets with only one or a few smaller operators will not return as robust results.

For this dataset no validation data was provided. The only way to evaluate the algorithm’s performance is based understanding how often the most reliable methods are used and looking how the trip length distribution and the average trip length compare with those of the NTS. As with the

Nexus data, the average trip length was somewhat shorter than NTS; both NTS and the algorithm implied slightly shorter trips in West Yorkshire than in Tyne & Wear. There are reasons to expect both NTS to slightly overstate trip length and the algorithm to slightly understate it.

46% of the trips were infilled using “reverse journey matching”, the most reliable way of infilliing alightings that was found. 10% of the trips were infilled using the smartcard user travel patterns.The remainder 43% of the trips had to be infilled using general travel patterns. Using general travel paterns is the best and only simple way to infill the remainder trips, but the likelyhood of achieving the right answer is expected to be in the low range.

The trip length profile compares reasonably well with NTS as before.

10.3 Summary of Datasets Studied

Table 10-1 shows the percentage of applicability of the different methods on the different bus datasets reported. As discussed previously, it’s noticeable the low applicability of methods 1 and 2 (different service) for the Ticketer data.

The dataset with only one service (#71) obviously did not return any matches across different services. The other two small datasets (010215 and Mersey travel) returned higher applicability of methods 1 and 2 for different services, but this is because we did not check geographic matching distances for these datasets. In other respects, the algorithms perform very consistently across datasets.

Table 10-1: Alighting estimation algorithm applicability for the different bus datasets

Data set	Method 1	Method 1	Method 2	Method 2	Method 5	Method 6	Method 7	Method 9	Remainder
	Next trip of the day principle. Same service	Next trip of the day principle. Different service	Last trip of the day principle. Same service	Last trip of the day principle. Different service	Most common alighting for this boarding stage for this user	Most common alighting for this service for this user	Most common alighting for this boarding stage for all users	Most common alighting for this service for all users	
#71	17%	0%	17%	0%				-	66%
#010215	17%	35%	14%	19%				-	16%
Mersey travel	17%	37%	12%	17%				-	16%
Week 1, May	17%	26%	13%	12%	10%	4%	18%	-	-
May+June 2016	17%	26%	13%	12%	18%	4%	10%	-	-
Ticketer data	19%	6%	17%	4%	8%	2%	37%	6%	-