Matching observed public route choice data to a GIS network

Marie K. Anderson, M. Sc. in Eng., PhD Student, DTU Transport, mkl@transport.dtu.dk

Thomas K. Rasmussen, M. Sc. in Eng., PhD Student, DTU Transport, tkra@transport.dtu.dk

Abstract

This paper is motivated by a growing interest in providing insight into route choices of public transport users. Observations on route choices are collected in the Danish National Travel Survey (Transportvaneundersøgelsen, TU) where respondents describe their chosen route (e.g., modes, times, origins, destinations). In this paper methods are developed to match these data to a GIS network in order to be able to compare the observed routes with routes generated by means of a GIS application. The GIS network is a schedule-based public transport network containing addresses, train stations, bus stops, transfers, train lines, bus lines and a road network. The method developed is able to identify the train stations used, identify the bus stops used, map observations onto relevant links with knowledge of line used and points travelled through (origin, bus stops, train stations, and destination).

1 Introduction

Knowledge about actual route choices for public transport passengers is important and can be used when assessing generated choice sets for route choice modelling. An objective measure of relevant routes in actual networks does not exist and assessment of the choice sets is often based on the experience and knowledge of the analyst. The work described in this paper is motivated by the need to get insight into the route choices of passengers in public transport.

Literature has paid increasing attention to route choices of private car users, but only limited efforts have been posed toward the understanding of route choices in public transport networks. One of the reasons lies in the difficulty to collect data on actual route choices in public transport networks, since a lot of information has to be provided to describe the routes actually used by travellers. For private transport, it is possible to use GPS devices to track routes and then map the data to a physical network (see, e.g., Jan et al. 2000; Schönfelder et al. 2002; Wolf 2004; and Zabic, 2011). For public transport, the same method is of little help because relevant information about the lines used is not retrievable with these devices, signals may fall out in tunnels (metro and sections of the urban rail system), and information on the trip purpose, which is another fundamental piece of information for uncovering route choice determinants, is not retrievable.

In order to use the collected data for comparison with generated choice sets, the observations should be represented in a network equal to the network in which the choice sets are generated. This sets some requirements to the collected data and to the applicability to the network in which the choice sets are generated.

Anderson (2010) presented the development of a questionnaire to collect details about the actual route choice behaviour in public transport networks. The questionnaire was after a full scale test (Anderson, 2010) added to the Danish Travel Survey (TU) which collects daily travel diaries with questions on activities and travels of a representative sample of the population and since in February

2009 public route choice observations have been collected continuously. When the travel is by a public transport mode an additional questionnaire with the new questions collects detailed information about access modes, stations, lines, departure and arrival times, trip purposes, transfers and egress modes.

In this paper methods are developed to match these data to a GIS network in order to be able to compare the observed routes with route choice sets generated by use of the schedule-based method proposed by Nielsen (2000). The matching of the observed data to the GIS network has proved to be very important for visualisation of the actual route choices of public transport passengers in the Greater Copenhagen area and for use in the assessment of generated route choice sets.

The paper offers a short literature review in section 2. Section 3 presents the data; the TU survey, the route choice questions, the dataset from TU and the network used to map the observations in are explained. In section 4 the method used in the matching of the observations to the network are explained. Section 5 discusses the method developed and the results.

2 Literature review

The following section presents some methods of collecting public route choice data and methods to match this data to GIS networks.

2.1 Collecting public route choice data

The accuracy of collected GPS data for use in private route choice analyses has improved in recent years (see Holm, 2009 and Zabic, 2011) from a large share of missing points (Nielsen, 2004, observed missing information for 90% of the trips in Copenhagen) to more accuracy (both devices and number of satellites but still the use of this data source for route choice knowledge is questioned (see Bierlaire and Frejinger, 2009).

Stopher and Greaves (2007) described the use of GPS devices together with a traditional travel diary and predicted that the future national travel surveys will rely more on GPS data. Chen et al. (2010) and Gong et al. (2012) reported the findings from a travel survey collecting data in a multimodal transport network by the use of GPS devices, and the developing of GIS algorithms to determine the travel modes used and the trip purposes for the trips. Up to 90% of the travel modes were correctly identified but the identification was as low as 29% for rail and 53% for buses showing that these methods are still not sufficient for route choice data collection in public transport networks.

Since GPS devices could not be used to collect the desired data for this survey (no mode or trip purpose information, signals fall out, etc. as discussed above) other methods were considered. Bovy and Bradley (1985) used stated preference survey for collection of route choice data. This procedure does not provide the actual routes as desired for this survey. Ramming (2002) asked drivers in the Boston area to name the origin and destination zone of their route and the road segment used. This procedure returned a great number of incomplete route descriptions. Some of these incomplete routes were fixed by using the shortest path between two known points or by using the routes of other respondents travelling between the same points. With this method a great amount of manual work is required to map the data afterwards and the method is not applicable for public transportation since the lines used would not be revealed by listing road segments. Prato (2005) collected data on route choice in a web-based survey where respondents indicate their chosen and

other considered routes by selecting the numbered order of passing through junctions on an interactive map of the city centre of Turin. The observations are for drivers and the method is not applicable to the public transportation system since the lines used in public transportation cannot be identified by choosing junctions in a network. Vrtic et al. (2006) asked respondents to provide information on the origin and destination cities of their trip and up to three cities or locations they passed through on the way. This provides a great deal of missing information and the exact actual route cannot be reproduced from these pieces of information. This method of asking for specific points of the travel is in some way applicable to public transport if the questionnaire is created to obtain all relevant information so that the exact route can be reproduced.

For public transport route choice, only very few studies have collected data to describe the route choice of the travellers. Hoogendoorn-Lanser (2005) collected data on considered choice sets and actual choices via face-to-face interviews. The survey was carried out for train users in a specific train corridor in the Netherlands and defined as a Hub-n-spoke network. The collected route choice data was not far as detailed as required for this survey with regard to feeder modes, exact bus lines, etc.

Surveys conducted via mail, telephone, web-based etc. are all conventional ways of collecting data on route choice. Often the data collected are concerning the attributes of the traveller and the trip since the actual route is rather difficult to obtain in this way. Mahmassani et al. (1993) and Abdel-Aty et al. (1995) described different approaches to the data collection for road users by means of questionnaires. Mahmassani et al. (1993) collected data on respondent characteristics and commuting patterns by using a short paper questionnaire sent to 13,000 households (less than 3,000 answers were acceptable) and those willing to follow up answered a more detailed questionnaire about their commuting trips describing routes link-by-link. Abdel-Aty et al. (1995) combined a computer-aided telephone interview (CATI) with GIS means to register the exact road segments the driver had used.

To examine the route choices of passengers in public transportation in the Greater Copenhagen area and to be able to compare the routes with generated choice sets, a collection method using additional route choice questions in the travel dairy form has been developed in Anderson (2010). The information of the routes collected is detailed enough to enable the analyst to reproduce the actual chosen route. In order to be able to describe the route choices in the detailed network of Greater Copenhagen the data collection involves a large number of observations. Therefore methods as face-to-face interviews used in in Hoogendoorn-Lanser (2005) would be very costly.

The data collection method from Anderson (2010) takes the following important points into consideration:

- Route choices in public transportation for all trips during a day
- High level of detail
- Possibility to reproduce the route in a GIS network
- Large number of respondents
- Cover the area of Greater Copenhagen and the public transport modes within
- Limited budget constraints
- Method to continuous data collection

2.2 Matching public route choice data to a network

The literature shows scarce effort in matching observed route choice data from a questionnaire to GIS networks. The detailed level of the Danish Travel Survey with the additional questionnaire on public transportation route choice is unique and relevant research literature for methods of matching collected public route choice data to a GIS network has been difficult to find.

When route observations are collected within the modelled network, for example by respondents pointing to a map (e.g., Prato 2005), the data are more or less straightforward to use, but because of computer power the method is only applicable to smaller networks. When collecting route choice data described in words, some standard procedures to match these data to the network have to be developed. Ramming (2002) collected route choices for car drivers by asking for the origin and destination zones of their routes and the road segments used. This procedure returned a great number of incomplete route descriptions, some of which were fixed by using the shortest path between two known points or by using the routes of other respondents travelling between the same points. The method requires a great amount of manual work to map the data afterwards and this is not favourable if the data amount is huge and the collection is ongoing.

2.2.1 Automated Fare Collection Data

Several studies on matching public transport data from automated fare collection (AFC) sources to a GIS network have been carried out. These have various purposes of estimating station-to-station origin-destination trip tables (Barry et al., 2002), route choice estimation (Zhao, 2004 and Wilson et al., 2009) and statistical analyses (Trépanier et al., 2007 and Slavin et al., 2009). Some of the challenges of matching the AFC data to a network are the same challenges met when matching the questionnaire data to a network and some of the assumptions made are presented in this section.

Barry et al. (2002), Slavin et al. (2009), Barry et al. (2009) developed methods to match the automated fare collection (AFC) from the New York City MetroCard data to GIS networks. The New York City Metro card is an entrance-only system which registers information on each boarding of a public transport vehicle (bus, metro, train). For each boarding the PT line ID and time are registered and for rail modes also the boarding station. The data is truncated to six-minute intervals to save data storage.

Barry et al. (2002) processed information about stations used by metro travellers and matched these to a database of the metro stations. The systems collected entrance data only making the first stop/station and transfer (boarding) stop/station easy to identify. The destination stations were identified using a set of simplifications;

- *The destination station on the first trip was assumed to be the equal to the departure station on the second trip*
- *The destination station on the last trip of the day was assumed to be equal to the departure station of the first trip of the day*

Barry et al. (2002) looked at trips with one or two trip legs only.

Zhao (2004) and Wilson et al. (2009) used data from the entry-only automated fare collection system from Chicago operated by the Chicago Transit Authorities and extended the work of Barry et al. (2002) by including buses in the studies, looking at train-train and train-bus transfers only. The

boarding bus stop for each bus trip was identified by comparing AFC data to automated vehicle location (AVL) data for buses and GIS network attributes. Information was given on the exact time of boarding the bus and the GIS network was searched at a *given network distance from to boarding stop to identify possible alighting stations*. Zhao (2004) successfully identified destinations for 65.5% (improved to 71.2% in Wilson et al., 2009) of the trips and with the knowledge of PT line used the trips were matched to a GIS network. Finally the matched routes and one alternative route generated by TransCAD were used to estimate route choice parameters.

Slavin et al. (2009) enhanced the work of Barry et al. (2002) and included bus trip legs from the New York City MetroCard data in the study examining the full set of alternatives (also bus-bus and bus-train in addition to Wilson et al., 2009). For bus-bus transfers an intersection table was created identifying the *nearest stops* on bus routes intersecting. Similarly, *for rail and bus the transfer location for the lines was identified when the lines intersect at a single location only*. If the first trip started with a bus trip leg the boarding bus stop was located at the *bus stop on the bus line closest to the home location* (if known). Finally major simplifications were made if the stops were not yet assigned; *if the bus boarding stop was not identified a stop was randomly assigned to a stop* (with a uniform transfer time distribution) and if the trip destination was not identified a station was uniformly sampled from all trips starting at the same origin. The authors reported to have assigned origin and destination stops for most AFC transactions. Barry et al. (2009) worked with the same data and same methods as Slavin et al. (2009) but instead of sampling destination stops for those not assigned a destination stop using the algorithms they discarded the data which was 10% of the trips.

Trépanier et al. (2007) processed smart card data from the city of Outaouais, Canada. The city has buses only (regular, express and special buses) and the smart cards give access to some or all buses (regular cards are not accepted at express routes etc.). Using GPS data the boarding stop is identified and stored when the traveller boards the vehicle. The authors assumed that a traveller alight at the *bus stop closest to the boarding stop of the next trip*. Trépanier et al. (2007) made use of the continuous data and included the possibility that the *last alighting stop of the day could be identical to the first stop of the following day*. If the destination stop of a trip could not match the origin stop of the next trip (if the destination was not reachable by the train or bus line boarded) *data from previous days were searched to find a similar boarding stop*.

Even though the above methods of matching of automated fare collection separate themselves from the questionnaire route choice data collected in this study the methods can be used as inspiration for the methodology developed. The issues of defining the boarding and alighting bus stops are especially interesting for the present paper. The AFC data collection has advantages since it is easy to collect and involves large samples of travellers. It has however disadvantages especially in the missing information about the origin and destination making reconstruction of full multimodal paths impossible using only this data. In the literature presented many assumptions are made to identify the routes and only trips following these trips patterns are correctly matched to the network. In a detailed public transport network offering many route alternatives the travellers do not always follows a symmetrical pattern when choosing route for the trips during day.

3 Data

In the following, the TU survey and the GIS network to which the route choices are matched are described.

3.1 The TU Survey

The TU survey is an existing and well-established collection of travel data in the form of travel diaries and respondents' and households' socio-economic data, and is consisting of a questionnaire which is either filled out on the internet (20 %) or via telephone (80 %). In the TU survey, respondents are a representative sample of the Danish population between 10 and 84 years who are asked to describe all their trips with both private and public transportation modes on the day before the interview. Since February 2009, the public transport route choice questions have been a part of the TU survey.

Respondents provide information on all their trips during the day (e.g., selected modes, time duration, length travelled, trip purpose) and all their socio-demographic characteristics (e.g., gender, age, income, place of residence and workspace). The data are a great source of information and enables revealing many interesting details about travellers' choices in public transportation networks. For more information on TU refer to Jensen (2009) and Christiansen (2009).

The existing TU survey links information on actual travel behaviour to a list of background variables making the survey an obvious choice for the addition of the public route choice survey questions.

3.2 Public transport route choice questions

The route choice questions added to the TU survey were shortly and precisely formulated in order to keep the questionnaire simple, obtain high completion rate, and collect good and useful observations. The information should be detailed enough to enable the reproduction of the route, but also simple enough for the respondent to provide it correctly. By answering questions about specific points on the trip the route can be reproduced with knowledge of the public transportation network. The development of the data collection method is described in Anderson 2010 and the route choice question part of the TU survey is explained shortly in the following.

All modes used on the trip are asked for. The respondent chooses from a drop down list (21 modes among others car, bike, walking, bus, train). When choosing a mode additional boxes to be filled in appear according to the mode entered.

- Walking, Bike, Car, Airplane, etc.
  - Enter length and time used
- Bus
  - Enter waiting time, bus line, length and time used
- S-train
  - Enter waiting time, from-station, S-train line, to-station, length and time used
- Train, Metro
  - Enter waiting time, from-station, to-station, length and time used

The respondent also enters the start and end points of the trip as addresses which are linked to coordinates, the purposes at the start and end trip points, and the departure time. The information has to be filled in for each trip during the day. However, the start point is calculated as the end point of the previous trip.

Transport modes in correct order:

	Transport mode		Line	Length		Time		
1.	Walking			0.5	km	8	min	
				Waiting time:		3	min	
2.	Bus		701	2	km	5	min	
3.	Walking			0.2	km	2	min	
	From-station:	Nykøbing F		Waiting time:		7	min	
4.	Other train			146	km	100	min	We calculated the distance to 146 km from the stated stations
	Transfer at:	København H		Waiting time:		8	min	
5.	S-train		E	13.9	km	20	min	We calculated the distance to 13.9 km from the stated stations
	To-station:	Lyngby		Waiting time:		5	min	
6.	Bus		300s	5	km	10	min	
7.	Walking			0.5	km	5	min	
			Sum:	168.1	km	173	min	

Figure 1: Example of a route description for a public transportation trip

The list of transport modes can get relatively long, especially when using public transportation, but the boxes to fill in are relative easily understood and along the way many checks of the entered is offered to the respondent. Figure 1 shows an example of a public route description. The respondent has travelled with several different public transportation modes and walked from origin and to destination. The blue boxes are to be filled in by the respondent. This is an example of a complicated trip with many transfers, but still the questionnaire is relatively easy to fill in.

When using the above mentioned list of transport modes, the route can be reproduced. Information on the start point of the trip can be used to find the bus stop by searching within a certain buffer (the length with a margin) for bus stops where the mentioned bus line stops. When the bus alights at a train station this point is fixed and the route to the next train stations as well. When selecting *Other train* between two train stations, the route is almost definite (not many alternative routes between two train stations are served by *Other trains*). The Danish rail network almost has a tree structure generally easing the description of route choice at beginning and end of the trip. In the end of the trip, the respondent uses bus again and the alighting stop of the mentioned bus line can be found by searching from the destination point.

In section 3 the procedure of matching the data to a route in a GIS network is explained in details.

3.3 Dataset

The information in the TU survey is in six head subjects; household, person, car, journey, trip and stage characteristics (for documentation of TU see Christiansen and Haunstrup 2011). For each of these subjects a table is representative in the survey database. Below the tables relevant for this paper are described.

- **Trip.** The trip table contains information for every trip starting from 3 am, namely departure time, trip purpose, destination, travel companions, payment of fare for public transport, passenger or driver of car, and coordinates for origin and destination of each trip.
- **Stage.** Alongside the trip table, the stage table includes details about the trip. Each trip is divided in stages for each transportation mode. Information is about mode, being driver or passenger, respondent's conception of length, time, and waiting time. For stages using public transportation modes, information on bus lines (line number) and train line (S trains are described with a letter), as well as access to and egress from train stations, is also listed, thus enabling the reconstruction of the chosen route.

The information on routes is mainly in the stage table. Each use of mode during a trip is defined in this table. The number of stages with bus, S-train, metro and other train in the Greater Copenhagen area is showed in Table 1. The table also illustrates how many respondents have entered information on line use (99.8 % of bus passengers, all S-train users) and which train station was travelled via (14 % of bus users, minimum 97 % of the train users). The low share of station information for bus user is due to the fact that the respondents are not asked about the boarding and alighting bus stops, but when travelling in a bus/train combination the train station information is added to the dataset.

Table 3-1: Number of stages using the four public transport modes, number of entered lines, from- and to-stations for travellers in the Greater Copenhagen Area

Mode	No. of stages	Line	From-station	To-station
Bus	1,584	1,581	214	208
S-train	1,039	1,039	1,039	1,022
Metro	459	-	459	444
Other train	259	-	258	254

From February 2009 to May 2010, approximately 6,300 interviews were collected in the Greater Copenhagen area with more than 22,500 trips of which 2,200 use public transportation for a part of the trip.

3.4 Networks

3.4.1 Physical network

The collected route choice observations are matched to a network representing the Greater Copenhagen area. 2 million people live in the area and it is the densest area in Denmark. The public services consist of buses, metro and trains (regional, S-train and local rails).

3.4.2 GIS network

The observations are matched to a schedule-based public transport network containing addresses, train stations, bus stops, transfers, train lines, bus lines and a road network. The network is also used by the schedule-based stochastic transit assignment model based on MSA used for the generation of choice sets (Nielsen 2000), and is analogous to the one used in Orestaden Transport Model (OTM) (e.g. Jovicic and Hansen 2003).

The network contains the Greater Copenhagen area. The transportation network consists of a road and path network used for walking, bikes, cars, buses, and a rail network. Network elements include:

- **Zones:** in this survey exact start and end points are used to investigate the exact route.
- **Connectors:** connect the exact start and end points to the road /public network.
- **Road/path network:** links and nodes, used for walking, bikes, cars, buses, etc.
- **Rail network:** rail for S-trains, regional trains, metro.
- **Stops:** bus stops and train stations, where passengers board and alight buses or trains. They are defined in stop groups with one or more stops, so that two stops on the opposite side of the road are in the same stop group.
- **Changes:** transfer links connecting bus stops and train stations.
- **Lines:** definition of bus and train lines.
- **Line Variants:** different types of each bus/train line.
- **Line Variant Elements:** each *line variant* is divided in a number of elements, *SQIdx*, for each segment between two stops, with *SQIdx* as a rising number in the driving direction. In each direction the *line variant* has a new *line variant element*.
- **Runs:** different variants of the *line variant*'s stop pattern that is the stops served by the line and the order of the stops.
- **Schedule:** links *runs* to *line variants.*
- **Schedule Elements:** information on stops served by the run, whether runs allow for passengers to board and/or alight, and arrival and departure time.

Figure 2 shows the structure of the public network with the content of the tables and the connections between the tables.

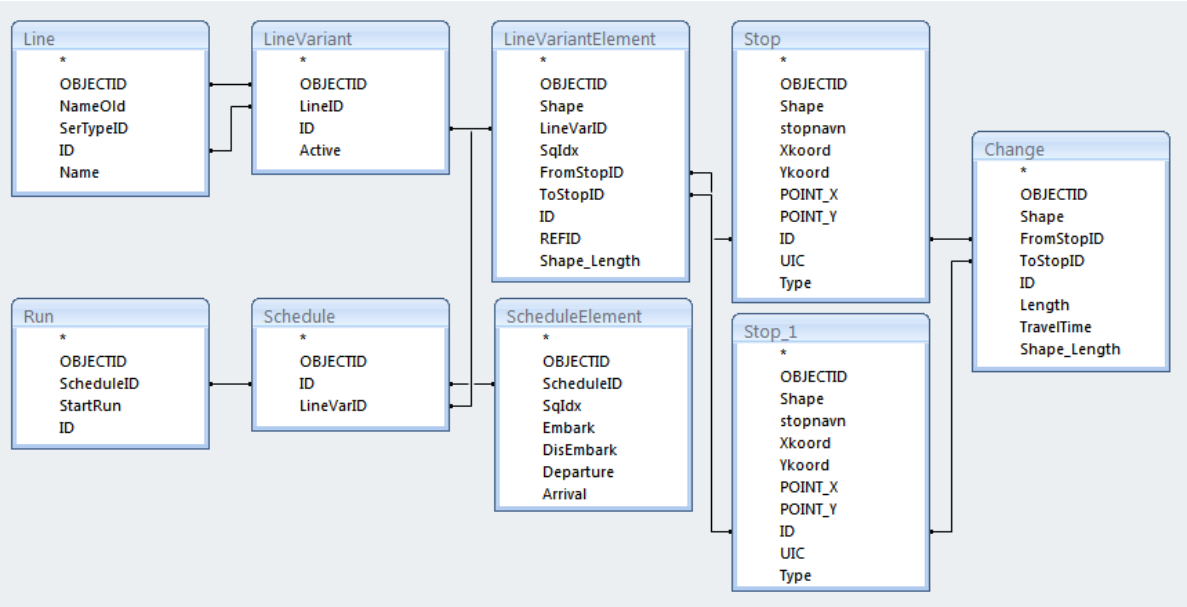

Figure 2: Database diagram of the public network structure

4 Method

1,793 observations are matched to a GIS network covering the Greater Copenhagen area. This is a schedule-based public transport network containing addresses, train stations, bus stops, transfers, train lines, bus lines and a road network. The network looks very different at different points in time since over time bus lines are rerouted, new train lines are built, and time tables are changed. It is

therefore important to have the correct network to match the observed data to in order to be able to find the actual used lines, departure and arrival stations, etc.

To obtain the correct network for the year the data was collected the GIS networks are built on the basis of the data behind "Rejseplanen.dk" (the Danish Public Transport Route Planner, see Rejseplanen, 2011). These data provide information about lines, stops, stations, time schedules, etc. at a given point in time, and accordingly allow the construction of the public transportation GIS networks. The important public network data are correct train stations, bus stops, train lines and bus lines. The observations are matched on line variant level causing the exact time schedules to be less important. The network data for a weekday in June 2010 are selected as a good representation of the observed data since stops, stations and lines correspond to the observations.

The matching of the collected data should fulfil some requirements in order to be accepted as correct. The method should be able to match a certain amount of the collected data to the correct route. The comparison of the actual route and the matched route is done manually for selected route observations. The method should be able to:

- Identify the train stations used
- Identify the bus stops used
- Map observations on relevant links with knowledge of line used and points travelled through (origin, bus stops, train stations, destination)

4.1 Identification of stations for train trips

For identification of train stations, a number of tables are available. Two methods are used and compared in order to check the correctness of the matching. The train stations serve S-train, Metro and/or other train (regional and IC).

- Train station names

  This method is based on a comparison between the station names in the TU data and in the network. The names are not defined exactly identical, but with the adding of the ending *st.* or *st. (Metro)* most of the TU stations have a match in the network data.

- Spatial placement

  For the stations in the TU data a list of coordinates are available. When plotting these coordinates and compare them to the coordinates of the network stations, all stations can be identified by using the following steps for one train service at a time:

  1. Plot coordinates for all TU train stations.
  2. Plot all network train stations.
  3. Select one of the (in 1.) plotted stations and search for network stations (2.) within 200 m. This always returns one and only one train station for the specific train service.

The results from the two methods only differ for one station which is fixed manually.

The matching between train stations in the TU data and in the network has to be done only once and the coordinates can afterwards be joined automatically. This step could be avoided if the station

names were defined exactly the same in the TU data and in the network, but this has not been done in order not to confuse the respondent and get too long drop down lists. If train stations were named according to the infrastructure serving them, the respondent had to enter a walking distance between the S-train and Metro even though they stop very close to each other.

When joining the list of station coordinates to the TU observations, the use of mode on the stage is used in order to identify the correct train station and stop type.

This steps identifies a train station (ID in the Stop table) for the 1,375 trips with at least one train leg.

4.2 Identification of bus stops for transfer to/from trains at train stations

When transferring between bus and train, a bus stop close to the train station will be used. In many cases several bus stops are placed close to a train station, and the following explains the method to identify the actual bus stop used for each stage. By means of changes from the train stations, all potential bus stops can be found. These stops are compared to the bus stops served by the bus line used and, if several bus stops are found, the one with the shortest transfer distance is used.

The method is divided in several smaller parts.

- The examined trips either arrive at or depart from a train station and consequently either the *to-stop* or the *from-stop* should be identified.
- The mode used on the next/previous stage is important to keep track of, because of the different train *types* and *changes* to train type stations.
- Most of the *changes* in the network are defined as going from a train station to a bus stop, but this is not consistent and hence both directions have to be examined.

This step identifies a bus stop (ID in the Stop table) for 439 trips.

4.3 Identification of *from-stop* for bus trips starting near origin – walking or bike as feeder modes

It is assumed that the traveller always uses the nearest bus stop served by the desired bus line. This is not always correct, since people are often willing to walk longer towards a bus stop in the travel direction rather than against the travel direction of the bus, even though the latter may be the closest to the starting point.

The identification is carried out for trips where the first part is by bus, or the second part is by bus and the first is by walking or biking. For these trips the traveller can be assumed to use the closest bus stop. When driving a car (or being a car passenger) to a bus stop, other issues may apply.

The identification is carried out by means of a network approach for one trip at a time.

1. All bus stops on the used bus line are loaded to the path and road network as *facilities.*
2. The origin coordinate set is loaded as starting point to the network (*incident*).
3. Identify routes from *incident* to *facilities* by means of the Network Analyst tool in ArcGIS.
4. The *facility* closest to the *incident* is identified and defined as the *from-stop* for the bus stage.

This step identifies a bus stop (ID in the Stop table) for 1,028 trips.

4.4 Identification of *to-stop* for bus trips ending near destination – walking or bike as feeder modes

For identification of bus stops used near destination the same approach as above is used with a few changes.

1. The examined trips use bus mode on the last stage or the second last stage and bike or walking on the last part.
2. Destination is loaded as *incident.*
3. *To-stop* is identified and updated.

This step identifies a bus stop (ID in the Stop table) for 1,013 trips.

4.5 Identification of *transfer stop* when transferring between two bus lines

In the public TU data 127 trips (7.1 percent) have transfers between two bus lines and, since the respondent does not state the exact bus stop used for transfer, a method to identify the stop is developed. Two types of trips are examined, namely trips with two successive bus stages (105 situations) and trips with two bus stages with a walking part in between (22 situations).

All possible transfers between two bus lines are identified to be stop groups served by both bus lines and bus stops connected by a *change*. If only one possible transfer is identified, this is the transfer used. When more than one possible transfer is identified, additional steps have to be run through.

Not all the above identified transfers are realistic because some can cause great detours for the traveller. Some rules are applied in order to identify the transfer stop used (arrival of the first bus line):

- If the two lines run parallel for some distance, there are often several bus stops served by both lines enabling transfer. The actual chosen transfer place depends on several factors as service level (travel time, comfort, frequency etc.) of the two lines, service level of the bus stops, transfer time, etc. This is not accounted for and the problem is simplified as follows:
    - If the arrival stop is placed so that bus no. 2 drives in the opposite direction of bus no. 1, the change is carried out at the first stop possible (left case in Figure 3)
    - If the arrival stop is placed so that bus no. 2 drives in the same direction of bus no. 1, the latest stop possible is chosen (right case in Figure 3)

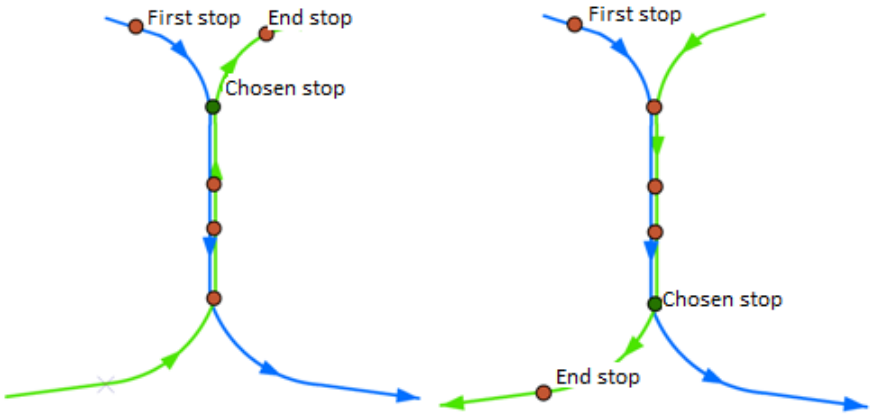

Figure 3: Transfer between two bus lines

Often the bus lines in the Greater Copenhagen area are really long and in these instances even more possibilities of transfers are at hand. Since we do not know the direction of the bus lines used several possible cases could occur, as illustrated in Figure 4.

The first case shows two lines in the same direction with a transfer on the last stop before the lines split the first time. In the following the first line takes the traveller away from the end stop, followed by a change to the second line leading him to the end stop. The last figure shows the opposite where the traveller travels longer than the end stop and changes for a bus going back to the stop. When the first shown case is possible then this is selected. The above method identifies the arrival stop for the first bus line. When the buses stop at the same bus stop group this stop is also identified as the departure stop for the second bus line. If the traveller has walked between the stops the stop with the shortest *change* is chosen.

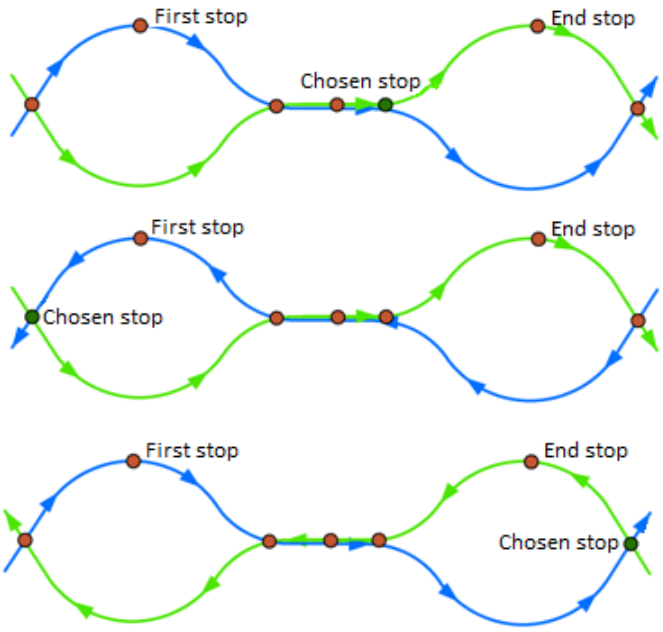

Figure 4: Long bus routes meeting and separating several times

This step identifies a bus stop (ID in the Stop table).

4.6 Identification of link pieces

When the stops used on the route are identified, the link pieces used in between are identified in order to map the complete route. The link pieces in the Line Variant Elements table are selected for each line variant and in the direction of the travel. The element *SQIdx* defines the direction of the line since the attribute is rising in number from start to end of the line.

This step identifies a number of link pieces (ID in the Line Variant Element table).

The results from the matching are shown in the next section.

5 Results

The described mapping method maps 91 % of the trips. For the remaining trips, data for the exact route are missing or incorrect or the methods are not able to identify the route chosen. In some observations the name of the train station used is missing, and in some the line number for buses is missing or is incorrect. In some cases, if the train station or bus line used is obvious, these can be corrected manually.

The trips which failed to be map matched by the algorithms can be divided according to five characteristics as presented in Table 5-1.

Table 5-1: Characteristics of trips which are not map matched, measured in percentages

Characteristic	Percentage of not matched
Bus line (missing/incorrect)	30.8
Transfer Bus-Bus	51.9
Transfer Bus-Bus-Bus	3.5
Transfer Bus-Train (or Train-Bus)	7.3
Network and other errors	6.5

The table shows that the highest number of non-matched trips is found within the bus-bus transfer trips. If the algorithm failed to identify a transfer stop or two stops with a change link between, the trip is not map matched. In 31% of the not map trips the bus line name is missing or the line name stated does not serve the area of origin/destination, etc. The transfer from bus to train (and vice versa) will fail to be map matched if the names of the lines are incorrect (so the bus line does not serve the train station entered) or the train station name is incorrect or missing.

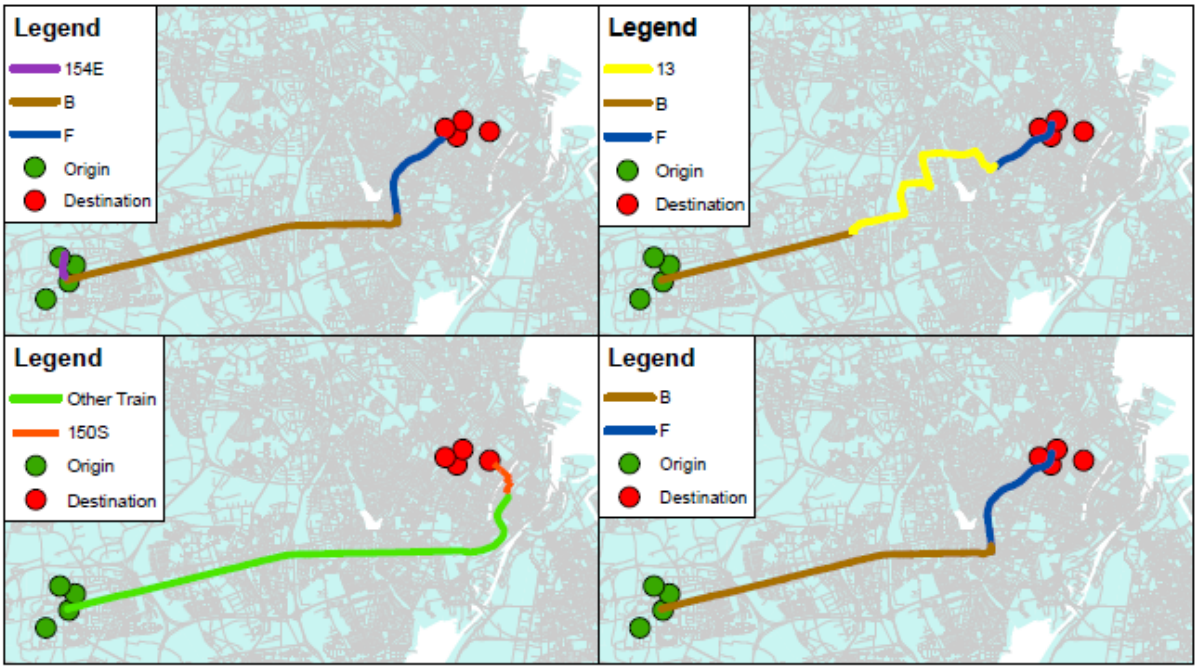

Figure 5: Maps showing routes for four trips from Høje Taastrup to Frederiksberg/Copenhagen N

In Figure 5, examples of four routes between Høje Taastrup and Frederiksberg/Copenhagen N are shown. All four routes are mapped according to the observations and the method has proven to be able to reconstruct routes correctly. The figure visualises why mapping the observations is useful. If the information was presented in a table, the routes would be difficult to compare and to assess, especially for people with less knowledge of the network. When the data are matched to the GIS network, the routes are visually comparable and easier to assess also for people with less network knowledge.

None of the four routes in Figure 5 are completely identical. Three of the four travellers used the S-train lines B and F for the greatest parts of the trip. One used a bus at the beginning of the trip (line 154E) in order to get to the train station. One of the S-train users chose to disembark the B-train at

an earlier station, take a bus (line 13) and depart the F train at a station closer to the end station of the route. The traveller has used twice the time of the other travellers in order to take this bus detour. We do not know why the traveller has chosen this detour, but the explanation can be problems on the S-trains on the day in questions or the convenience of travelling along with somebody on the bus line (following a child to school etc.). The last traveller used the regional or IC train from Høje Taastrup to Nørreport and continued by bus line 150S. The destination of this trip is somewhat different from the others, but still this route was possible for the other travellers if they were willing to walk for 1-1.5 kilometres to their destination.

Figure 6 shows the route bundle for the whole dataset of public route observations (1,793 observations are matched). The highest share of travellers is using a public transport mode in or near the inner Copenhagen area. The thickest lines leading to/from Copenhagen are the train lines and these are used by the highest number of people. Many of the bus lines in the periphery of Copenhagen are used for less than a half percent of the observations. Some routes are not used by any people in the sample, but most of these are small bus lines serving a local area and it is acceptable for the further analyses that no one have stated to have used these. When the sample gets larger, enough users of all the small routes will be included at some point.

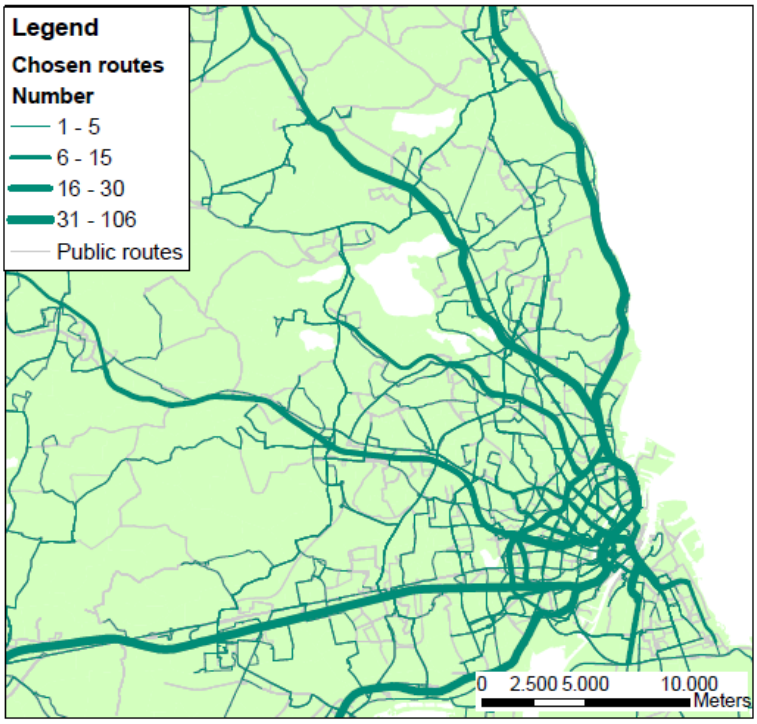

Figure 6: Route bundle for all observations

6 Discussion

For the bus stops some assumptions in the matching method could be questioned. When transferring between bus and train, departing near origin, and arriving near destination, it is assumed that the closest bus stop is used. In many situations the traveller will benefit more from using a bus stop different from the nearest. When arriving with bus to a train station the traveller will sometimes alight at the first stop enabling the traveller to run in order to catch an earlier connection. The opposite is the case when departing from a bus stop near a train station. From origin or at destination the same can be happening, for example using a stop served by the bus later than the bus stop closest by, etc.

The assumptions concerning choice of transfer stop when transferring between two buses can be questioned. The last possible stop is chosen with this method, but this can also depend on the service level of the bus (comfort, travel time, etc.) or the situation (can the traveller spot the next bus when sitting in the first).

The issues of identification of bus stops could be solved using the network attributes. When the analyst has information about the lines used it is possible to pick out only the relevant bus lines and carry out a route choice assignment between the origin and destination (or train stations) of the traveller. In this way the assignment model can assist in the identification of the most relevant transfer location between the bus lines. This of course requests the network attributes to be as precise a description of the real network as possible.

The method developed for identifying transfer bus stops is very much dependent on the underlying network attributes. If no transfer link is defined between the two bus lines the algorithm is not able to identify any possible bus stops. As presented the highest number of non-matched trips is found within bus-bus transfer trips and the network attributes could be responsible for this. Ideally the observations from respondents of actual transfer in the network could be used to define extra transfer links to improve the network.

7 Conclusion

This paper develops methods to map match the collected data in a few steps. The results from the study shows that it is possible to map match public route choice data collected via a questionnaire in a travel diary form to a GIS network.

The identification of the train stations is easy when the names of the stations used are given in the observations. The identification of bus stops are more cumbersome since these are not mentioned in the observed route choice data. Several assumptions have to be made to identify the bus stops used.

At the start and/or end of each trip including a bus:

- The traveller is assumed to board at the bus stop closest to his origin point served by the stated bus line
- The traveller is assumed to alight at the bus stop closest to his destination point served by the stated bus line

When transferring between bus and train:

- The traveller is assumed to board/alight at the bus stop closest to the train station

When transferring between two bus lines:

- If the two bus lines serve the same bus stop the traveller is assume to transfer here
- If the two bus lines serve bus stops connected by a transfer link in the GIS network the traveller is assumed to transfer here
- If multiple transfer locations possible the traveller is assumed to
    - stay as long as possible in bus one if the buses travel in the same direction
    - transfer as early as possible if the buses travel in opposite directions

The first/last bus stop and transfer at train station methods only consider distance and in several cases this will be different from the actual choice. The bus-bus transfer method assumes that the network contains all transfer links used by the travellers which is not always the case.

Identification of transfer stops using an assignment model only including the relevant bus lines and origin/destination locations could perhaps provide a higher map matching percentage or a more precise description of the actual route choices of the travellers. This will be an issue for future studies.

The map matching algorithm provided a successful map matching of 91% of the observed trips. This number is acceptable for the future use of the data since a high number of the observed trips are made useable for research purposes. The list of characteristics for the non-matched trips suggests that especially the transfer between two lines should be looked into in future research. The study emphasizes the importance of the high level of detail in the route choice observations and shows that with this level of detail it is possible to develop simple methods which reproduce more than 90% of the public route choice observations.

The matching of actual routes to the GIS network is very important for the future use of public route choice observations and the results have been used in several projects at DTU Transport. Halldórsdóttir (2010) used the data to assess and model choice of feeder mode to train stations. Rasmussen (2010) and Larsen et al. (2010) used the matched routes to assess generated route choice sets and finally an on-going project is using the routes together with the generated route choice sets to estimation parameter for route choice preferences.

8 References

Anderson M.K. (2010). Development and Assessment of a Data Collection Method for Route Choice in Public Transport, in *Selected proceedings for the Annual Transport Conference in Aalborg*.

Barry J.J., Newhouser R., Rahbee A. & Sayeda S. (2002). Origin and Destination Estimation in New York City with Automated Fare System Data. *Transportation Research Record: Journal of the Transportation Research Board, 1817*, pp.183-187.

Barry, J.J., Freimer, R. & Slavin, H. (2009). Use of Entry-Only Automatic Fare Collection Data to Estimate Linked Transit Trips in New York City. *Transportation Research Record: Journal of the Transportation Research Board, 2112*, pp. 53-61.

Christiansen H., Haunstrup B. (2011). *The Danish National Survey - Declaration of Variables, TU 2006-10, version 2,* DTU Transport.

Christiansen H. (2009). Modernisering af Transportvaneundersøgelsen, Paper presented at the *Annual Transport Conference in Aalborg*.

Halldórsdóttir, K. (2010). *Choice of Access Mode to Passenger Trains*, Master thesis, DTU Transport, 2010.

Jan O., Horowitz A., & Peng Z. (2000). Using GPS data to understand variations in path choice. *Transportation Research Record*, 1725, pp. 37-44.

Jensen C. (2009). Danskernes Transport – hvor meget, hvordan, hvor og hvornår? Paper presented at the *Annual Transport Conference in Aalborg*.

Jovicic G., Hansen C.O. (2003). A Passenger Travel Demand Model for Copenhagen, *Transportation Research Part A: Policy and Practice*, Vol. 37, Issue 4, pp. 333-349.

Larsen, M.K., Nielsen, O.A., Prato, C.G. & Rasmussen, T.K. (2010). Generation and Quality Assessment of Route Choice Sets in Public Transport Networks by means of Data Analysis. In *Proceedings of the European Transport Conference*.

Nielsen O.A. (2000). A Stochastic Transit Assignment Model Considering Differences in Passengers Utility Functions, *Transportation Research Part B: Methodological*. Vol. 34B, No. 5, pp. 337-402. Elsevier Science Ltd.

Prato C.G. (2005). *Latent Factors and Route Choice Behaviour*, PhD Thesis, Turin Polytechnic, Italy.

Ramming S. (2002). *Network Knowledge and Route Choice*, PhD Thesis, Massachusetts Institute of Technology, Cambridge, USA.

Rasmussen, T.K. (2010). *Rutevalg i kollektiv transport i Hovedstadsområdet – generering og kvalitetsanalyser af valgsæt*. Master Thesis, DTU Transport, 2010 (in Danish).

Rejseplanen (2011). *Rejseplanens API Documentation*, available at http://labs.rejseplanen.dk/files/api/rest_documentation_latest.pdf

Schönfelder S., Axhausen K., Antille N., & Bierlaire M. (2002). *Exploring the Potentials of Automatically Collected GPS Data for Travel Behaviour Analysis – A Swedish Data Source*, In: Möltgen J., Wytzisk A. (Eds.), GI-Technologien für Verkehr und Logistik, number 13 in IfGIprints. Institut für Geoinformatik, Universität Münster, Münster, pp. 155–179.

Slavin, H., Rabinowicz, A., Brandon, J., Flammia, G., & Freimer, R. (2009). Using automated fare collection data, GIS, and dynamic schedule queries to improve transit data and transit assignment model. Chapter 6 in *Schedule-Based Modeling of Transportation Networks,* Wilson, N.H.M. & Nuzzolo, A. (Eds.), Kluwer Academic Publisher, pp. 101-118.

Trépanier, M., Tranchant, N. & Chapleau, R. (2007). Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *Journal of Intelligent Transportation Systems: Technology, Planning and Operations*, Vol. 11, No. 1, pp. 1-14.

Wilson, N. H., Zhao, J., & Rahbee, A. (2009). The potential impact of automated data collection systems on urban public transport planning. Chapter 5 in *Schedule-Based Modeling of Transportation Networks*, Wilson, N.H.M. & Nuzzolo, A. (Eds.), Kluwer Academic Publisher, pp. 75-99.

Wolf, J. (2004). Applications of new technologies in travel surveys, paper presented at the *7th International Conference on Travel Survey Meth*ods, Costa Rica.

Zabic, M. (2011). *GNSS-based Road Charging Systems – Assessment of Vehicle Location Determination*. PhD Thesis, DTU Transport, the Technical University of Denmark, Kgs. Lyngby, Denmark.

Zhao, J. (2004). *The planning and analysis implications of automatic data collection systems: Rail transit OD matrix inference and path choice modelling examples*. M.S. Thesis, Massachusetts Institute of Technology, Cambridge, USA.