Matching GPS Data to Transport Networks

Adel Bolbol[1], Tao Cheng[1]

[1]University College London, WC1E 6BT

Tel. (+44(0) 20 7679 2000) Fax (+44 (0) 20 7679 3042)

a.bolbol@ucl.ac.uk, tao.cheng.ucl.ac.uk, www.ucl.ac.uk/spacetimelab

KEYWORDS: Network Matching, GPS Trajectories, Mode of Transport, Classification, Movement

1. Introduction into Network Matching

GPS data is widely collected and used in our everyday life through smart phones and tracking or navigation devices. The contextual travel information of GPS data, such as the type of network that a trajectory is travelling on, is very useful for many transport applications such as conducting modal split analysis, studying tourist activity (Edwards, et al., 2009) and the impact of a strike on transportation systems (Tsapakis, et al., 2012).

To find such contextual information, GPS data needs to be assigned to the network that it travels on. We coin for this process the term “Network Matching”. Hence, network matching is the process of matching positional data to their respective transportation network. This must not be confused with map matching, which is the process of assigning every point to its corresponding network link in a given network (Quddus, 2006). On the other hand, network matching is the process of selecting the network on which a trajectory is travelling. An example is detecting that a group of GPS points representing a person travelling by train is travelling on a train network. The network matching process can be used for several purposes such as a pre-map matching step (To provide prior knowledge of which network to snap to) (Quddus, 2006), analysing different network usages (or network split) (TfL, 2009), or detecting the mode of transport (Bolbol, et al., 2012).

Movement data can be broken down into several terminologies. Every two GPS fixes form a GPS segment, and every string of GPS segments of the same mode of transport form a GPS stage (TfL, 2009). Zheng et al. (2008) states that each two non-walk stages are separated by a walk stage. Moreover, Stopher (2008) states that a walk stage can easily be separated from the rest of the modes. Therefore, the remaining non-walk stages can be subject to be tested by a network matching method to identify whether they follow any transportation network.

For example, a given non-walk stage can be tested to whether all its GPS segments are within a pre-specified distance from the train network and follow a specific train route. This is illustrated in as two stages one of a car mode and the other of a train mode. The figure shows that the car fixes clearly follow the road network while the train fixes follow the train network.

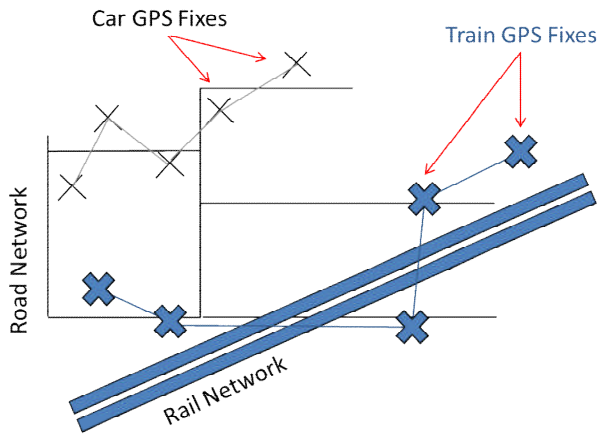

Figure 1. Conceptual Idea of Network Matching

In this work specifically, we are constrained to specific transport modes that exist within London, of which some can be network-matched and some cannot. Section 2 discusses the network types and the constraints that face the network matching process for the City of London.

2. Data

This section presents the network and GPS testing data available for testing the method developed in this paper. The section also provides the rationale for using specific networks for testing the developed method.

2.1 Network Data

We use London as our case study, as transport-rich as it is, containing a number of different transportation networks. These are illustrated in Figure 2 as the road, underground, train, footpath and bus networks.

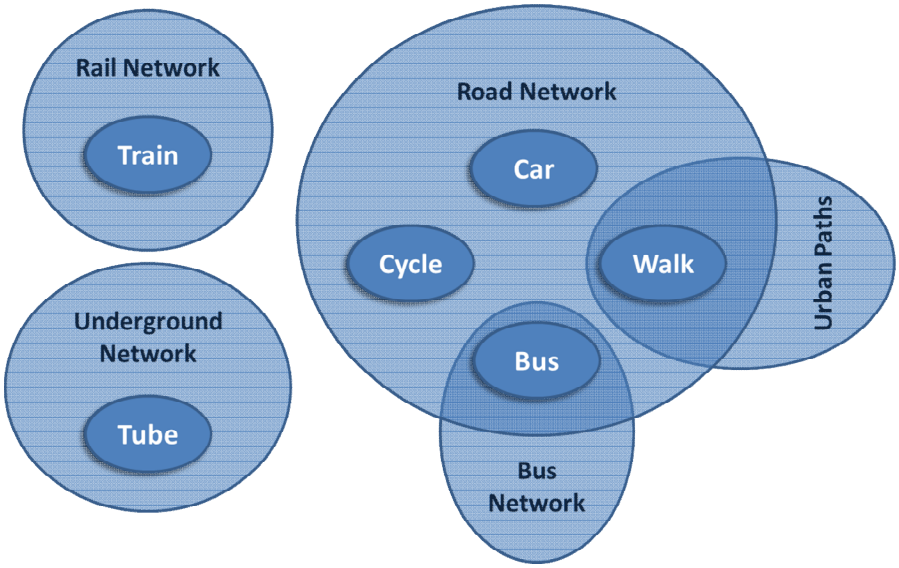

Figure 2. Different Transportation Networks in London

Each network holds one or more travel mode type and vice-versa. The ITN road network for example holds cycle, bus, walk and car modes, while the walk mode exists partially within

the road and urban path networks. So a question arises; which networks do we match to? The answer to this question depends on the constraints of the network matching algorithm. In order to ensure achieving a high accuracy, unique network routes need to be used as a constraint. This means that a trajectory is to be questioned whether it follows a certain route on the tested network. This is discussed in detail in section 3.2. Therefore, the process is to be applied to networks with specified routes with unique identifiers. This only applies to the underground, rail and bus networks. In other words, we only will consider public transit networks for such a process due to route-based structure that these networks are made of. The data of these networks are shown in Figure 3, Figure 4 and Figure 5 with a description of their sources. What could be noted is that the rail network extends into the whole UK rather than just in London.

Figure 3. London TfL Bus Network (Data Provided by TfL)

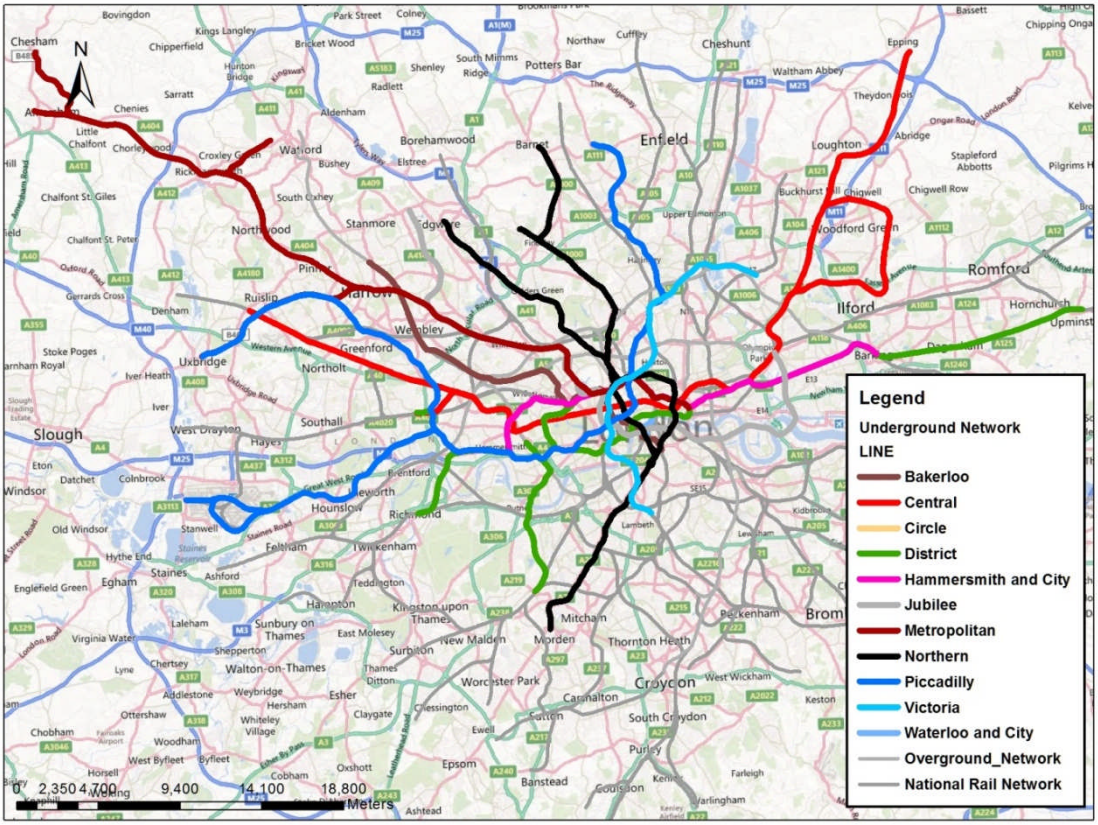

Figure 4. TfL's London Underground Network Data (Data Provided by: TfL)

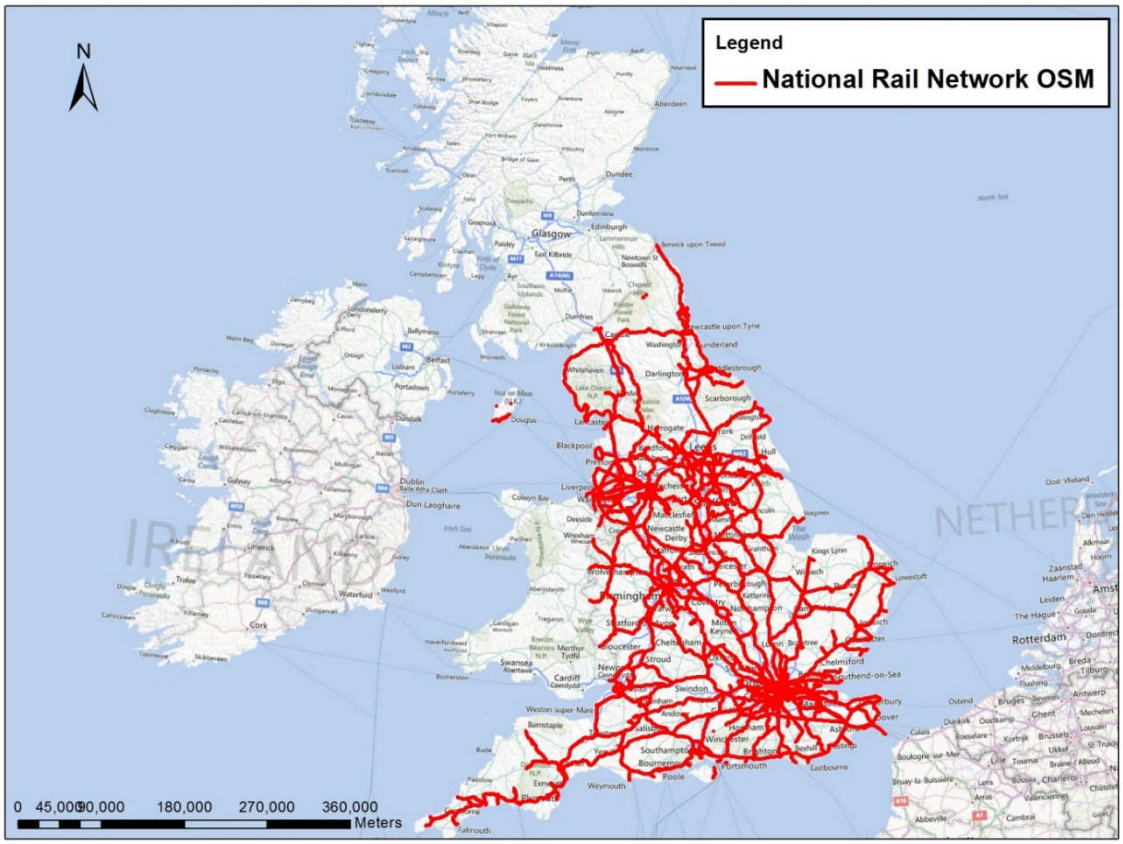

Figure 5. UK National Rail Network (Data Provided by: OpenStreetMap)

2.2 GPS Data

The data provided for testing this work is part of a project held at UCL, aiming at understanding individual travel behaviour from sparse GPS data. The sample consists of 10 participants in London who collected data over 2 weeks using u-blox GPS devices (u-blox, 2009) at a 60 seconds collection rate (Bolbol & Cheng, 2010). The mode of transport was labelled by each participant segment-by-segment, allowing the validation of the developed network matching method results.

3. Network Matching

This section discusses the idea and principle of applying the network matching process to GPS data. It also discusses the method constraints, namely; the stage time duration and GPS fixes distance from the network under investigation.

3.1 Principle

The network matching algorithm *aims at verifying whether a stage of a given non-walk mode follows the underground network in a unique route*. The method is similar for each of the bus, train or tube networks except for minor differences due to the different nature of the network, its dataset structure or the nature of movement of the trajectories that use it.

The algorithm starts by finding the nearest network routes ranked according to how close each route is from a GPS fix from closest to furthest, and fitting this information into a matrix. The GPS fixes of any non-walk stage are then tested to whether they all fall within a threshold **distance** from the same route. The **duration** of the stage is also constrained to a temporal threshold otherwise the stage is discarded from being underground-network-matched. If a given stage is matched to the underground network, it is then noted as such, as well as the name of the nearest route. Once that stage is tested for the three networks, all the results can be queried at a further phase and a final classification can be assigned to the stage.

3.2 Constraints

There are two constraints for this test, namely; (a) the **maximum distance** allowed to the underground network to allow the inclusion of an underground route, and (b) the **minimum time** duration threshold that a given stage is allowed to be, for it to be qualified as a candidate for being matched to the underground network. These two constraints are demonstrated in Figure 6 as an underground segment constrained by these two thresholds. These constraints will exist for all the public transport networks; however, their threshold values will depend on the nature of movements of the trajectories on each network and the nature network structure itself.

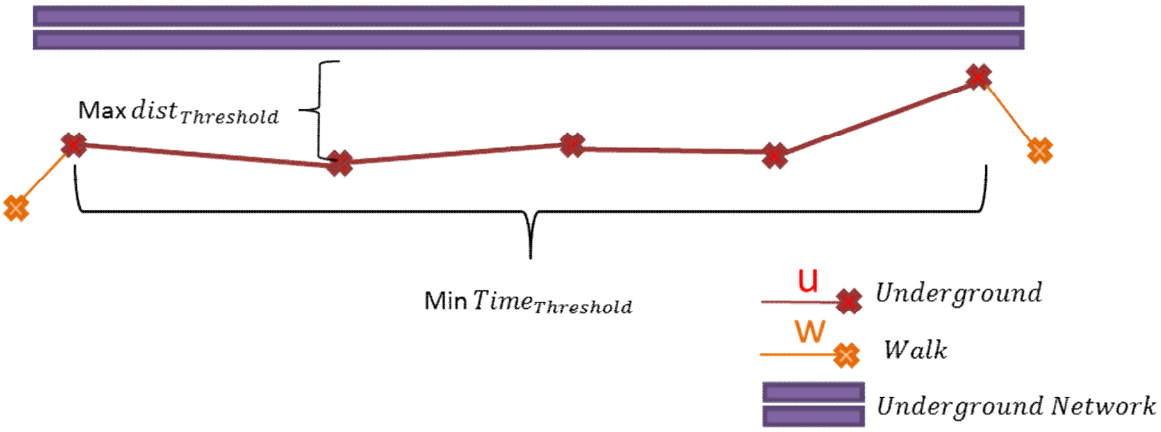

Figure 6. Underground Network Matching Constrained by a Temporal and a Distance Threshold

As for the **distance** from the network routes, it will depend on the GPS device accuracy. The stated positional accuracy of the u-blox devices is ±4.3 meters (u-blox, 2009). Nevertheless, due to the nature of the data being collected in an urban environment, many sources of error affect this accuracy such as multipath, bad GDoP, and Ionosphere and Troposphere disturbances (Hinch, 2007).Yet, for the purpose of this algorithm, a generous distance is more appraisable since the test will check anyway if all the consequent points follow a unique network route, which provides another checkpoint for the classification. Therefore, we assign a distance of **150 meters** which is around 30 times the stated positional accuracy of the u-blox devices.

On the other hand, the **temporal** threshold for the length of an underground stage will depend on the general average journey times within the one of the London's networks. The shortest distance between two underground stations is between Charing Cross and Embankment, a distance of 100m (TfL, 2012), and the time taken to do this trip is **1 minute** which is equivalent to the collection rate of the GPS data in this research. Hence, there is no need to assign a minimum threshold for the time spent on an underground stage trip.

As for the train network, according to (TfL, 2011), the average time spent travelling per day by London residents on train mode is around 8 minutes. And since this might typically be over two trips (example: home to work and back), then each leg will be of 4 minutes on average. Hence, we assign the minimum train stage duration as **4 minutes**.

For bus network, according to TfL (2011), the average time spent travelling by bus per day is around 13 minutes. This can imply the break of this average into two trips or three habitual trips between [work - (leisure or shopping) – home], since these constitute 80% of the trip purposes in London in 2008 (TfL, 2009). This means that an average bus trip would be around **4 minutes** long for each of the three daily trips assuming they are all done by bus. Hence, we use this threshold for the time duration threshold to accept bus stages in the bus NM algorithm.

4. Network Matching Accuracy

The algorithm is run through two phases, one in ***Python*** using AcGIS's library "***ArcPY***" to find the nearest neighbours of the transport networks, and one in ***R Project*** using the output tables from Python to reason about the distance, time and route following constraints. The results of the matching are discussed in this section.

4.1 Underground Network Matching

Figure 7 illustrates an example of the matching to the "District" Line of the underground network (matched GPS fixes are denoted in red). On the other hand, Table 1 shows the confusion index of the results of different GPS fixes being tested for maching to the undergound network. The table reveals a matching accuracy of **77%**, while having a confusion of 18% of bus fixes being matched too.

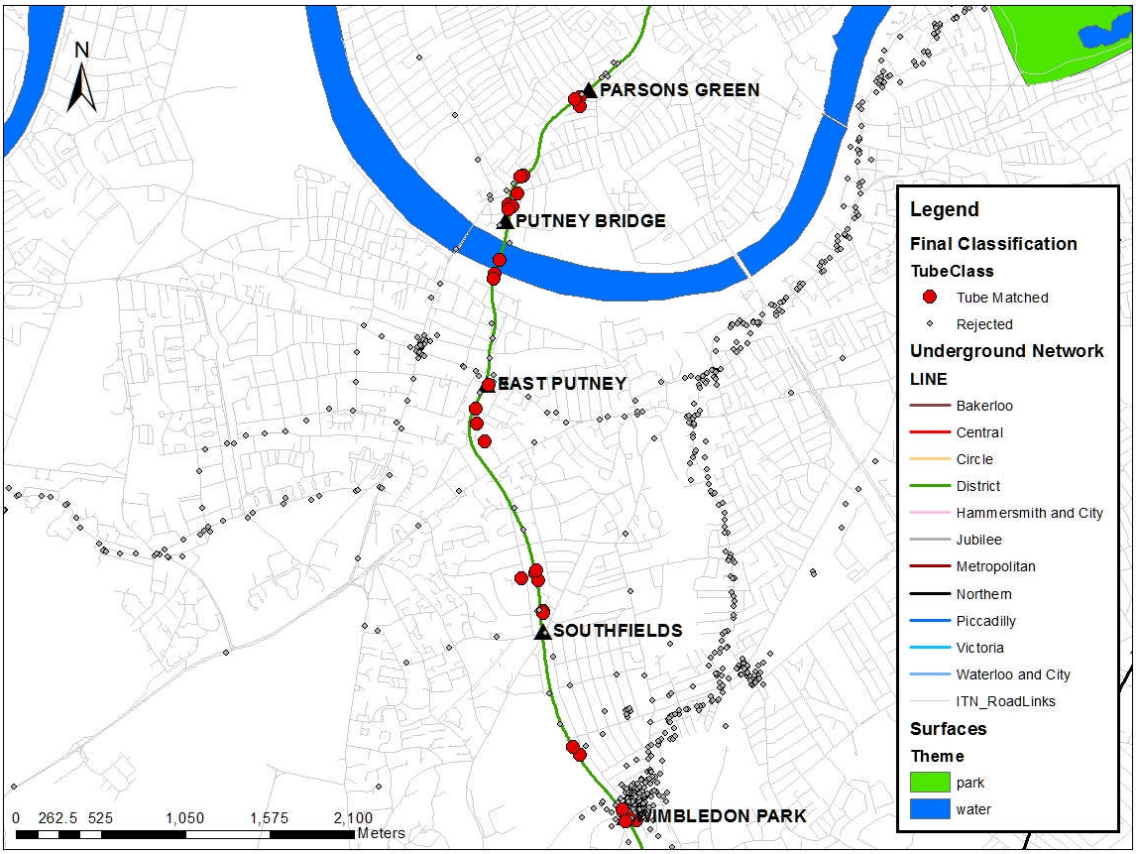

Figure 7. Example of GPS Fixes Network-Matched to the London Underground Network (District Line)

Table 1. Confusion Matrix of Results from the Underground Network Matching for GPS Fixes (Dismissing Tunnel Trips)

Truth	Inferred			
	No Entry	Not Tube	Tube	% of Tube Matched Fixes
bus	57	931	224	18.48%
car	124	2046	38	1.72%
cycle	47	1384	25	1.72%
train	17	265	0	0.00%
tube	5	51	197	77.87%
walks	10355	0	0	0.00%

4.2 Train Network Matching

Figure 8 demonstrates an example of repetitive train trips being matched to the rail network. Table 2 also shows the confusion index for GPS stages when being matched to the rail

network. Despite the confusion with car and cycle, yet the train network matching accuracy is as high as **100%**. The confusion stems from the nature of some long car/cycle trips that take place on roads parallel to rail tracks.

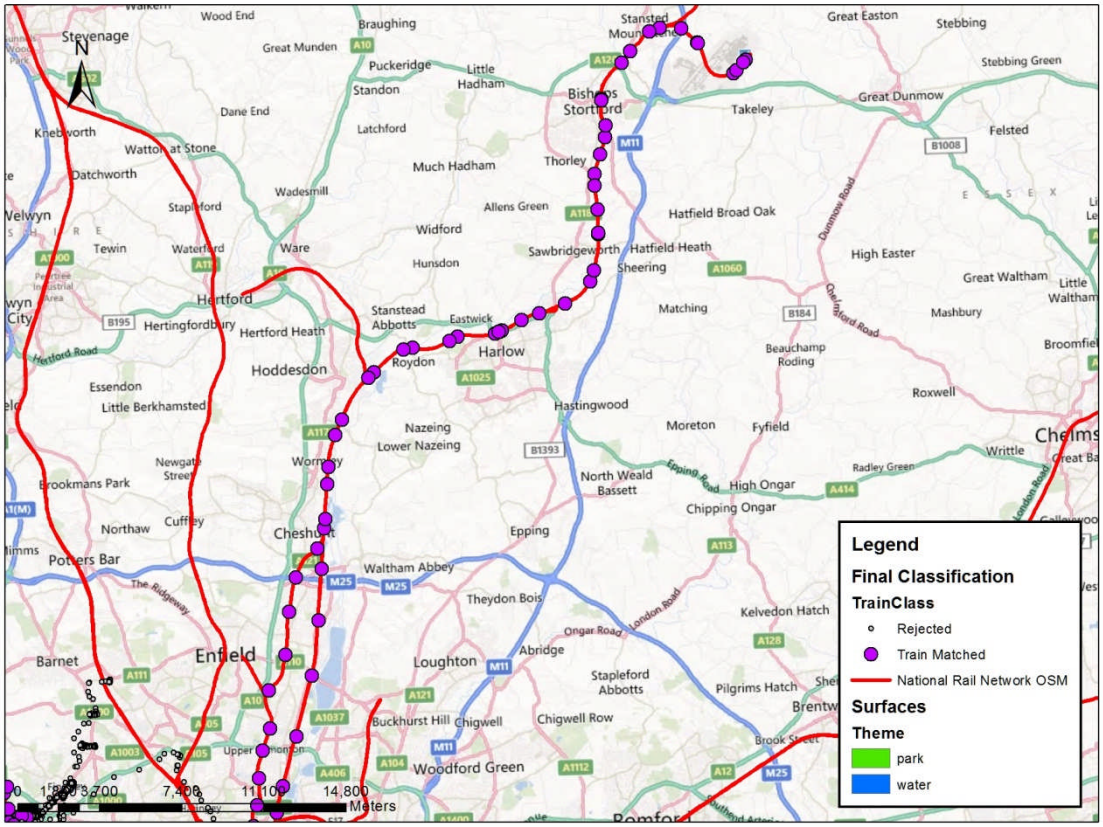

Figure 8. Example of Matching to UK National Rail Network

Table 2. Confusion Matrix for Results from the Rail Network Matching for GPS Stages Applied to SVM Classification Results

Truth	No Entry	Inferred		
		Not Train	Train	% of Train Matched Stages
bus	44	86	24	15.58%
car	87	61	62	29.52%
cycle	17	51	22	24.44%
train	0	0	25	100.00%
tube	58	17	8	9.64%
walks	389	107	83	14.34%

4.3 Bus Network Matching

Figure 9 demonstrates an example of the results attained by the matching to the bus network. Table 3 also shows the confusion index for the same process exhibiting an accuracy of **55%** of correctly matching bus stages to the bus network. Some minor confusion with the car and cycle stages also occur (15% and 12%) too due to the communal usage of the same road network.

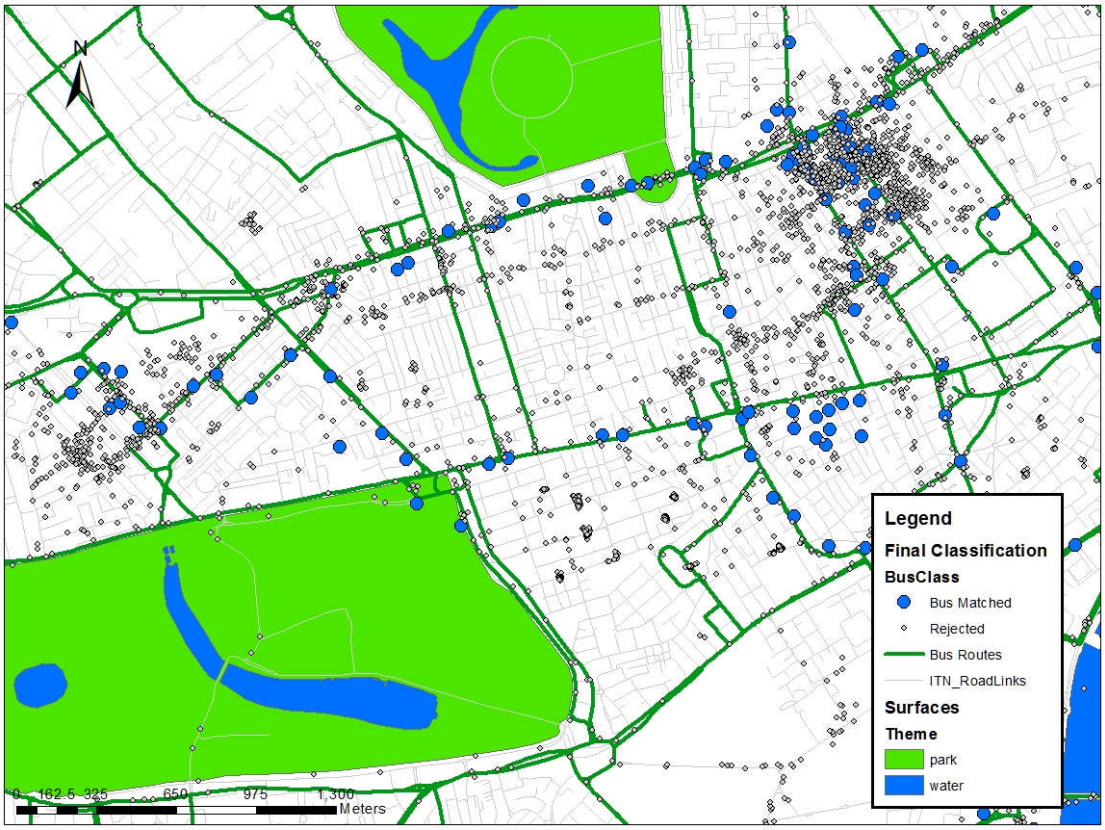

Figure 9. Another Example of GPS Fixes Network-Matched to the London Bus Network

Table 3. Confusion Matrix for Results from the Bus Network Matching for GPS Stages

Truth	Inferred			
	No Entry	Not Tube	Tube	% of Tube Matched Stages
bus	5	65	84	54.55%
car	12	165	33	15.71%
cycle	0	79	11	12.22%
train	1	24	0	0.00%
tube	38	39	6	7.23%
walks	579	0	0	0.00%

5. Conclusions

In this work we present a novel approach for testing whether GPS trajectories are travelling on specific transport networks. The purpose of the test is to serve applications such as detecting the mode of transport, acting as a pre-map matching process or calculating the network usage split. We test the method on GPS data from 10 participants in London collected over 2 weeks using u-blox GPS devices at a 60 seconds collection rate with labelled modes for every GPS segment.

The method is applied to non-walk GPS stages for all public transit networks, since walk stages are well-known to be easily identified due to their low speed values. The method achieves 77%, 100% and 55% accuracy for correct matching the underground, train and bus networks respectively. Some confusion occurs between other modes such as the car and cycle modes with the train and bus networks which is due to the vicinity (or coincidence) of the rod

network with some train and bus routes. The confusions can be resolved by a further step of reasoning between the results of the matching process to different networks, and further querying the nature of the trajectory's movement such as its direction or speed of travel.

6. Acknowledgements

We would like to express our gratitude to u-blox and EPSRC, the sponsoring bodies for this research, specifically Chris Marshall for his invaluable ideas and support. We would also like to thank TfL for providing the data for this work.

7. References

Bolbol A & Cheng T (2010). *GPS Data Collection Setting For Pedestrian Activity Modelling.* GIS Research UK 2010. London, University College London.

Bolbol A Cheng T Tsapakis I and Haworth J (2012). Inferring Hybrid Transportation Modes from Sparse GPS Data using a Moving Window SVM Classification. *Computers, Environment and Urban Systems*. In Press.

Edwards D Griffin T Hayllar B Dickson B and Schweinsberg S (2009). *Visitors to Urban Destinations: Understanding Tourist 'Experiences' and 'Behaviour' in Cities, an Australian Case Study,* Technical report. Gold Coast: CRC for Sustainable Tourism Pty Ltd.

Hinch S W (2007). *Outdoor Navigation with GPS.* 2nd ed. Berkeley: Widerness Press.

Quddus M (2006). *High Integrity Map Matching Algorithms for Advanced Transport Telematics Applications,* PhD Thesis, Department of Civil and Environmental Engineering. London: Imperial College London.

Stopher P R (2008). Collecting and Processing Data from Mobile Technologies. *The 8th International Conference on Survey Methods in Transport*, Annecy, France, ISCTSC.

TfL (2009). *Travel in London: Key Trends and Developments Report Number 1,* London: Transport for London.

TfL (2011). *Travel in London: Supplementary Report: London Travel Demand Survey (LTDS),* London: Transport for London.

TfL (2012). *Key Facts.* [Online]
Available at:
http://www.tfl.gov.uk/corporate/modesoftransport/londonunderground/1608.aspx
[Accessed 23 October 2012].

Tsapakis I Turner J Cheng T Heydecker BG Emmonds A and Bolbol A (2012). Effects of Tube Strikes on Journey Times in the Transport Network of London. *Transportation Research Record,* I(2012), p. 84–92.

u-blox (2009). *YUMA: Software and Service for Capture and Process.* [Online]
Available at: http://www.u-Blox.com/en/gps-solutions/yuma.html
[Accessed 29 August 2009].

Zheng Y Liu L Wang L and Xie X (2008). Learning transportation mode from raw GPS data for geographic applications on the Web. *17th international conference on World Wide Web*, Beijing, China, ACM Press, pp. 247-256.

8. Biography

Adel Bolbol is a PhD student at UCL, Civil, Environmental and Geomatic Engineering department. He studied for his MSc in GIS from City University - London in 2007. His current research interests span between Spatial Analysis, Movement Data Analysis, Transportation, Spatio-Temporal Human Behaviour Modelling, WebGIS, Geospatial analysis and data mining.