Computers, Environment and Urban Systems 36 (2012) 526–537

Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification

Adel Bolbol *, Tao Cheng [1], Ioannis Tsapakis [1], James Haworth [2]

Civil, Environmental and Geomatic Engineering, University College London, United Kingdom

ARTICLE INFO

Article history:
Available online 12 July 2012

Keywords:
Transportation mode
Classification
Variable selection
SVM
Travel behaviour

ABSTRACT

Understanding travel behaviour and travel demand is of constant importance to transportation communities and agencies in every country. Nowadays, attempts have been made to automatically infer transportation modes from positional data, such as the data collected by using GPS devices so that the cost in time and budget of conventional travel diary survey could be significantly reduced. Some limitations, however, exist in the literature, in aspects of *data collection* (sample size selected, duration of study, granularity of data), *selection of variables* (or combination of variables), and *method of inference* (the number of transportation modes to be used in the learning). This paper therefore, attempts to fully understand these aspects in the process of inference. We aim to solve a classification problem of GPS data into different transportation modes (*car*, *walk*, *cycle*, *underground*, *train* and *bus*). We first study the variables that could contribute positively to this classification, and statistically quantify their discriminatory power. We then introduce a novel approach to carry out this inference using a framework based on Support Vector Machines (SVMs) classification. The framework was tested using coarse-grained GPS data, which has been avoided in previous studies, achieving a promising accuracy of 88% with a Kappa statistic reflecting almost perfect agreement.

1. Introduction

Understanding travel behaviour is important for many applications such as studying tourist activity (Edwards, Griffin, Hayllar, Dickson, & Schweinsberg, 2009) or the impact of a strike on transportation systems (Tsapakis et al., in press). To understand travel behaviour, some standard data collection practices have been in place in order to collect travel data. Among these practices are GPS-based travel surveys, where participants carry a GPS device for a certain duration of time and following this up by a prompt recall survey to report trip information, such as the transportation modes they used in every trip (e.g. *cycle*, *walk*, *bus* and so forth) (Stopher, 2008). Yet, many studies have reported high underreporting rates due to the participants' burden to fill in daily details of their activities (Bricka & Bhat, 2006).

As a result, research has emerged in the previous decade attempting to infer the transportation mode from GPS data. This inference could largely replace or complete a lot of the feedback required by users when labelling and tagging travel diaries. Studies aiming at inferring the transportation mode could be divided into procedural and Machine Learning (ML) approaches. Procedural approaches attempt mainly to make inferences based on logical assumptions, such as how a typical person would travel (Stopher, Clifford, Zhang, & FitzGerald, 2008a). Other assumptions include the surrounding environment, such as the nearest transportation networks (Chung & Shalaby, 2005), or the temporal logic assumptions of activities, such as people are more likely to have no activity after mid-night (Liao, Fox, & Kautz, 2007). On the other hand, ML approaches attempt to do the inference based on learning from existing data, possibly combined with similar logical assumptions. Examples of these studies use Decision Trees (Manzoni, Maniloff, Kloeckl, & Ratti, 2011; Reddy et al., 2010; Zheng, Chen, Li, Xie, & Ma, 2010), Bayesian Networks (Stenneth, Wolfson, Yu, & Xu, 2011), Fuzzy Logic (Schüssler & Axhausen, 2009), Hierarchical Conditional Random Fields (Liao et al., 2007), and Support Vector Machines (SVMs) (Zheng, Liu, Wang, & Xie, 2008). These ML approaches could be broken down into three aspects (or phases): *data collection* (sample size selected, duration of study and granularity of data), *selection of variables* for inference (or combination of variables), and *method of inference* (the details of the learning algorithm used). Most of the previous ML attempts possess several limitations in each of these three previously mentioned aspects. Accordingly, these limitations could be arranged into three categories: *data collection*, *variable selection*, and *method*-related issues.

* Corresponding author. Tel.: +44 788 983 1988.
E-mail addresses: a.bolbol@ucl.ac.uk (A. Bolbol), tao.cheng@ucl.ac.uk (T. Cheng), i.tsapakis@ucl.ac.uk (I. Tsapakis), j.haworth@ucl.ac.uk (J. Haworth).
[1] Tel.: +44 781 560 1230.
[2] Tel.: +44 781 607 6958.

0198-9715/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compenvurbsys.2012.06.001

First, some *data collection-related* processes and properties limitations exist in literature, such as the selected sample size, duration of study, granularity of the data, the data collection methods, and the validation techniques used. Second, there are several issues regarding the *variable selection* to be used for the inference. Different studies use different variables (or combination of variables), such as speed, acceleration, maximum or medians speed, and acceleration and length between GPS fixes. However, none of the studies, to the best knowledge of the authors, have based their variable selection process on statistical evidence. Third, the *method*-related issues include the usage of a limited number of transportation modes in the learning, the high dependence on segmentation into transportation modes, and high reliance on temporal information. Moreover, some *method*-related assumptions are often made in previous work such as that certain modes cannot follow each other in a GPS sequence and that every two GPS consecutive fixes are analysed individually ignoring the track as a whole.

Therefore, in this work we attempt to fully understand and account for these three aspects in the process of inference. We aim to solve a classification problem of GPS data into different transportation modes (*car*, *walk*, *cycle*, *underground*, *train* and *bus*). First, we overcome the data-related issues by collecting GPS data for a recommended sample size for this study (81 participants) for a period of two weeks to account for the weekly seasonal variation (Bolbol, Cheng, Tsapakis, & Chow, 2012). We set the devices to a collection rate of 60 s to conform to the recommended collection rates suitable for such studies. The data is collected in a continuous manner over the two weeks to ensure the natural flow of the travel patterns of every participant. The data is simplified for the participants to label themselves by segmenting the track into individual trips on an online platform. Section 2 highlights these data-limitations in detail and describes the sample size calculation and characteristics used to determine the number of participants and duration of study needed for an efficient validation of our proposed framework. We also describe the GPS data collected for this research with a brief account on its descriptive statistics.

Before introducing the inference framework, it is essential to select the best classifier(s) or independent variable(s) IVs to be used to classify GPS points into transportation modes (Mitchell, 1997). Therefore, as a second phase, we run an analysis of variance (ANOVA) test to select the IV(s) that best discriminate between the different transportation modes. In turn, this should improve the performance of whichever classification algorithm that would be used in the following phase. We statistically compare the candidate variables using different statistical measures, such as Wilks' Lambda and between-groups F to assess each variable's discriminatory power. The results from the classification, based on the selected variables are then analysed and compared illustrating the power of each over different modes (categories). This analysis is presented in Section 3.

Finally as a third phase, we attempt to identify transportation modes from the collected sparse GPS data, without information or assumptions about the participant's temporal or location contexts, which some of the previous approaches were based on. We use Support Vector Machines (SVMs) to perform the inference from speed and acceleration values calculated from GPS data. Due to its high quality of out-of-sample generalization and ease of training, SVMs provide far beyond the capacities of traditional ML methods used in previous research. Furthermore, using SVMs, the selected kernel could be applied directly to the data without the need for a feature extraction process. This is advantageous in the context of learning from the structure of the data, since a lot of this structure is lost by the feature extraction process. This enables us to study a sequence of movements of a participant rather than each movement individually, and hence, achieving a better classification. We achieve this by using a moving window that classifies instances of data consequent blocks. We complement this by using logical filters that apply a transition matrix between different phases of the trip. This is presented and described in detail in Section 4. The results of this inference are presented in Section 5 along with some discussions and conclusions in Section 6.

2. Data collection

This section illustrates the most striking data collection-related limitations. First, we describe the basis of the sample size calculation method which is used to determine the number of participants and duration of study needed. Then, we also define the sample characteristics that would lead to an efficient validation of our proposed inference framework, overcoming the data limitations that exist in past research. Lastly, we describe the GPS data collected for this research with a brief account on its descriptive statistics.

2.1. Travel survey definitions

In order to de-construct a GPS track, some definitions have been standardised to be used for the description of different fragments of the *trip*. For example, the route between any two consecutive GPS points is called a *segment*. Trips also consist of a number of *stages* (a group of segments). A new stage is defined when there is a change from one mode of transport to another, or where there is a change in vehicle of the same mode (Anderson, Abeywardana, Wolf, & Lee, 2009).

2.2. Limitations of data collection for inferring the transportation mode

Over the last decade, a plethora of studies have attempted to infer the transportation mode from GPS data collected by travel surveys. Most of these studies have been carried out in complex urban study areas using either: mobile applications on smart phones (Manzoni et al., 2011; Stenneth et al., 2011); strictly GPS devices alone (Chung & Shalaby, 2005; Liao et al., 2007; Schüssler & Axhausen, 2009; Stopher, Clifford et al., 2008a; Zheng et al., 2010) or integrated with other devices, such as accelerometers (Reddy et al., 2010), or; others through mobile phone call detail records (CDR) (Wang, Calabrese, Di Lorenzo, & Ratti, 2010).

Such studies resulted in collecting a large amount of diverse data to test different approaches. However, the majority of the studies did not have any specifications for their sampling techniques. For example, they did not base their sample size calculations on any statistical framework. Some of them used as low as 4 participants (Liao et al., 2007), 60 trips (Chung & Shalaby, 2005) or as many as 4882 participants (Schüssler & Axhausen, 2009) without providing statistical justification for using such numbers. Another issue is the study's duration, where several studies use less than even one day's worth of data (Manzoni et al., 2011; Reddy et al., 2010), whilst others use data of less than a week duration (Liao et al., 2007; Schüssler & Axhausen, 2009). Note that both cases do not account for the weekly seasonal variation. We discuss this in another work of ours by running statistical analysis of traffic data from similar complex urban cities, resulting in a recommendation of no less than 81 participants for at least 2 weeks, as a minimum guideline for sample sizes to be used in such studies (Bolbol et al., 2012).

Another issue is the temporal granularity of the GPS data (also called the epoch rate of collection), where most studies use a 1 s collection rate. The question that arises is: Do we need such detail? Not only would that create a load on memory and on battery restrictions on current GPS devices or smart phones, but it will also add to the computation cost of any of the used algorithms. This will

Table 1
A summary of previous studies' accuracies and sample sizes and durations.

Study	Accuracy (%)	Sample size	Duration	No. of modes
Chung and Shalaby (2005)	92	60 Trips		4
Liao et al. (2007)	90	4 Participants	6 Days	3
Zheng et al. (2010)	76	65 participants	10 Months	4
Reddy et al. (2010)	94	16 Participants	7.5 h	5
Manzoni et al. (2010)	82	5 Participants	Several hours	7
Stenneth et al. (2011)	93	6 Participants	3 h	6

also impose a daily (if not hourly) burden on the participants to charge their devices and act as a constant reminder that they have a tracking device. This gives rise to typical participant-related reported problems, such as feeling vulnerable when carrying the device, or influencing their normal behaviour (Anderson et al., 2009). We discuss this issue in Bolbol and Cheng (2010a) demonstrating that a collection rate of 30–60 s is sufficient enough for a city like London for this type of study.

Some of the current studies have other data limitations, such as using one-purpose trips (Manzoni et al., 2011; Stopher, Clifford et al., 2008a) that influence the results of any inference of any method used, restricting the outcome to one or two modes. Also, some of the validation methods did not seem to be based on actual labelled data by the participants. For example, Schüssler and Axhausen (2009) use consensus data from previous years of the same city to evaluate the classification results, while Wang et al. (2010) use Google Maps to verify the results by comparing them to the proposed modes by Google for the corresponding trip travel times.

Although other methods using labelled data have achieved accuracies of 90% or more, yet sample sizes and durations were often not adequate to give full accreditation to the results. Table 1 lists some of these studies along with the accuracies achieved, sample sizes, survey durations and the number of modes considered while validating each method's performance.

We overcome these data-related issues by collecting GPS data for this study for a period of two weeks to account for the weekly seasonal variation. We set the devices to a collection rate of 60 s to conform to the recommended collection rates suitable for such studies. The data is collected in a continuous manner over the two weeks to ensure the natural flow of the travel patterns of every participant. The data is simplified for the participants to label themselves by segmenting the track into individual trips on an online platform. The rest of this section describes the sample size calculation method used to determine the number of participants and duration of study needed for an efficient validation of our proposed framework, along with the description of the data properties.

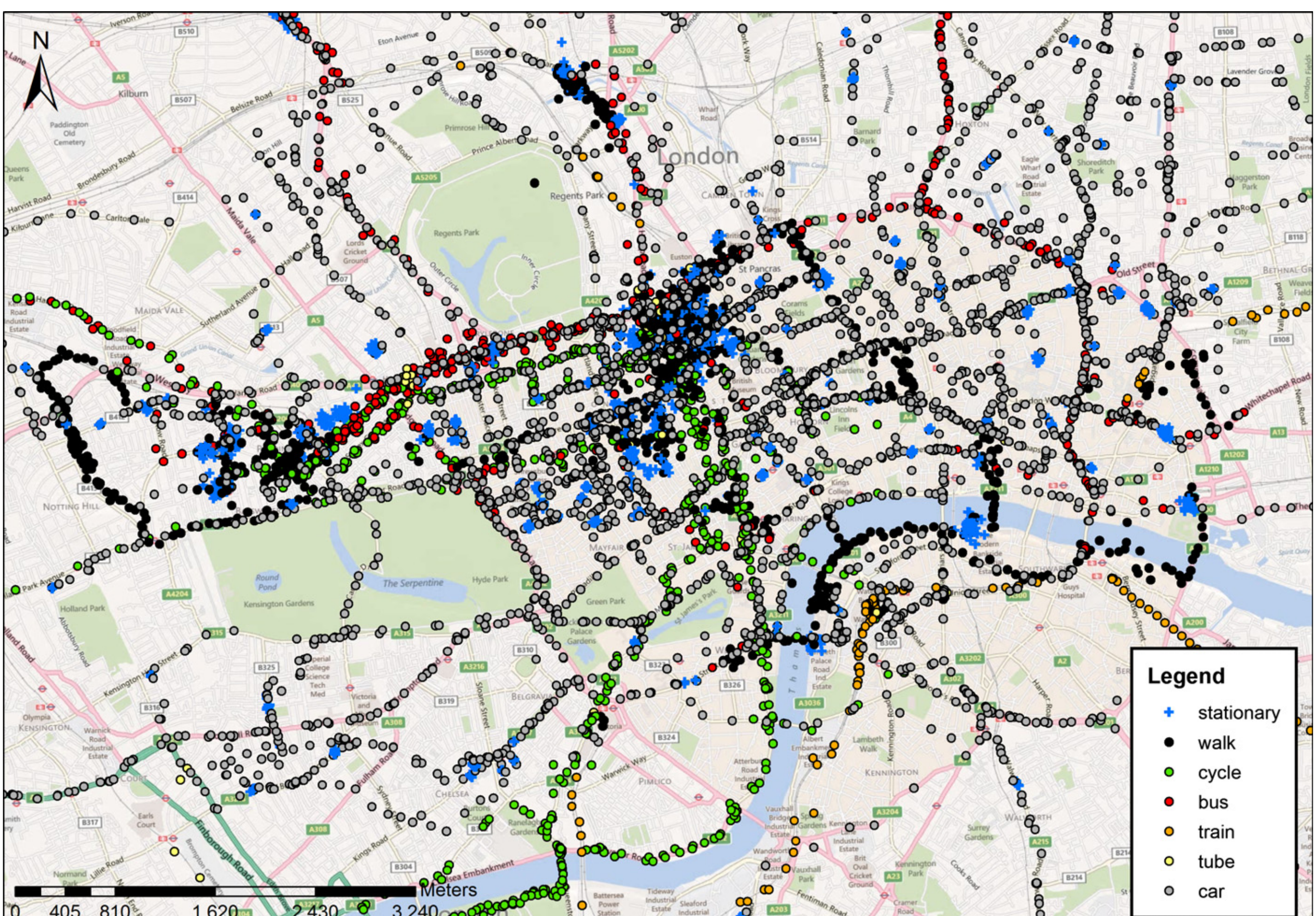


Fig. 1. The study area and dataset in Greater London.

2.3. Sample size calculation

The calculation of the minimum sample size is an important consideration in this study and for travel behaviour studies in general. For conventional one-day or two-day travel surveys, sample size procedures are well known and widely applied; for example, the Travel Survey Manual by Cambridge Systematics (1996). The corresponding sample size procedures for GPS-based panel surveys however, are less well developed (Xu, 2010). Therefore, the estimation of a statistically adequate sample size, for whichever survey type, requires good knowledge of the variables under investigation, their coefficient of variation and the desired accuracy of measurement together with the level of significance associated with it (Smith, 1979). The variables to be investigated in this study, speed and acceleration, are calculated from the collected GPS data.

The Coefficient of Variation (CV) is also an important element for the estimation process, where the sample size largely depends on how much the variable deviates from its mean. The CV is a normalized measure of dispersion of a probability distribution, or a statistical measure of the dispersion of data points in a data series around the mean. It is calculated by dividing the standard deviation by the mean of the population. The third element is the accuracy desired (and significance level), where the accuracy level is the percentage error acceptable to the analyst. Both the accuracy and the significance level are context-dependant elements to be decided by the analyst according to the analyst's experience (Ortúzar & Willumsen, 2011). Once these three factors are defined, the sample size (n) could be computed from Eq. (1).

$$N = CV^2 Z_\alpha^2 / E^2 \tag{1}$$

where E is the level of accuracy and Z_α is the standard normal value for the confidence level (α) required. Since the acceleration is a derivative of speed, the CV of speed could also represent the acceleration's variability. According to different studies that aim to measure and analyse the variability of speed for different modes, a minimum of **81 participants** was calculated for an adequate sample size (Stopher, Kockelman, Greaves, & Clifford, 2008b; TFL, 2011; Thompson, Rebolledo, & Thomson, 1997; Weidmann, 1993), which is the number we are using in this research. The details of this calculation are out of the scope of this paper and therefore are not mentioned here explicitly, yet further details on this calculation is available in Bolbol et al. (2012).

2.4. Sample data

As mentioned in Section 2.3, the optimum sample characteristics that would be ideal for such a study in an urban area should consist of a minimum of 81 participants for a minimum period of 2 weeks. On the other hand, the specifications of the data to be collected for this type of survey could be determined and made clear from the discussion in Section 2.2 for the context of this study. These specifications include that the data should contain all transportation modes that would exist in a typical complex urban setting including *walk, cycle, bus, car, train* and *underground*. The data also has to be of a moderate temporal granularity to avoid battery and memory constraints, which lessens the participants' burden. Nevertheless, it should also provide the minimum epoch rate duration sufficient for this type of study, as well as cutting down the computational cost (Bolbol & Cheng, 2010a). The data also has to cover at least a period of two weeks to account for the weekly seasonal variation. Also, the nature of the data has to be of a mode-continuous nature (i.e. avoiding recording only single mode trips). Furthermore, in order to ensure the quality of the validation process, the data has to be assigned transportation mode labels by the participants for each stage of their trips.

Table 2
A summary of transportation modes' attributes.

Transportation mode	Average speed (m/s)	Average acceleration (m/s^2)	Average distance (m)	Average time difference (s)
Bus	3.70	0.08	244.17	68.20
Car	7.13	0.23	239.06	61.17
Cycle	4.87	0.09	326.79	461.36
Train	18.65	1.34	818.11	59.16
Tube	6.27	0.08	2021.16	529.34
Walk	0.58	0.01	53.53	533.54
Grand total	3.74	0.14	199.07	353.97

Therefore, the training dataset used for testing consists of 2-weeks long multi-modal tracks (2 waves – to account for seasonal variation) of 81 users within 2010–2011 (Fig. 1). The tracks are collected within London at 1 min frequency. London is selected due to its complexity and the diversity of its transportation networks. The transportation mode of each segment in the dataset was labelled by the users themselves using an online platform enabling them to edit their own tracks. More details on this online process could be found in Bolbol, Cheng and Paracha (2010b). The dataset was then filtered for the *car*, *cycle*, *bus*, *walk*, *tube* and *train* modes, so as to use SVM classification to infer these modes.

The number of fixes of the *walk* mode in the dataset was the highest amongst other modes and almost as double as the second highest mode (*car*). This demonstrates the high occurrence of walks within an individual's daily journey. This is due to the fact that walking often occurs as an intermediate link between different modes. The underground (*tube*) mode recorded the least number of GPS fixes since nearly half of the tube network in London is actually underground, which causes the loss of GPS coverage. In this case, a typical underground tube segment would consist of only two points; an entrance and an exit fix.

Table 2 shows the respective averages of speed, distance and time difference between every two fixes of each of the 6 modes in this dataset. Outliers due to GPS errors are accounted for and removed. The table demonstrates a clear confusion and overlap between the speed averages of *bus* and *cycle* modes. This is due to London's high congestion during the peak hours of the day. This emphasizes the nature of different forms of commute in London's network. It also appears that *train* and *tube* modes have long distances between their consecutive fixes due to their high speeds. However, the average time difference among the *tube* mode fixes seems to exceed that of the *train* by a tenfold due to loss of signal.

3. Independent variable selection

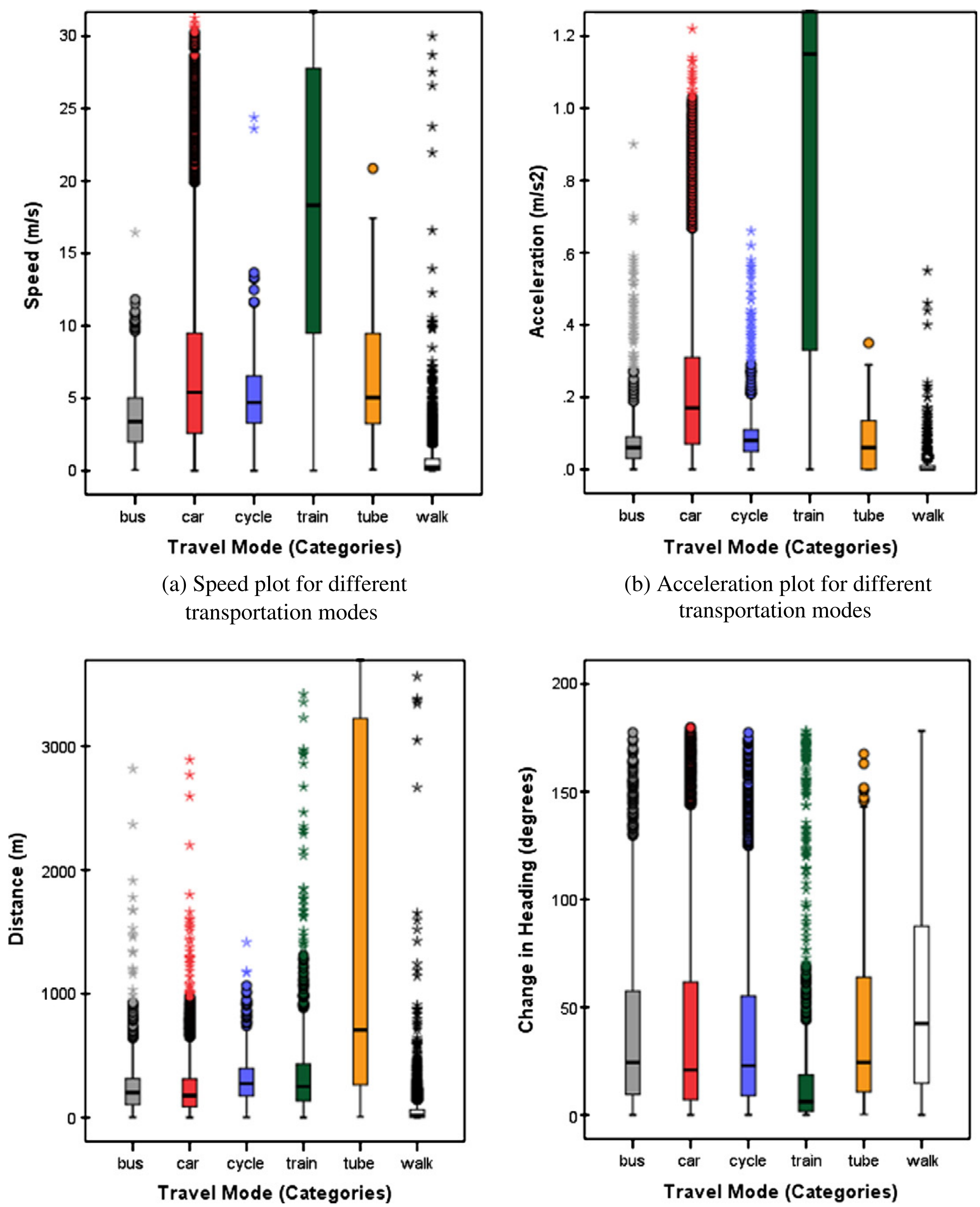
Generally in a classification problem, the variable that is to be predicted is known as the dependent variable (transportation mode in our case) because its value depends upon, or is decided by, the values of all the other attributes. The other attributes that help to predict the value of the dependent variable, are known as the independent variables (IVs) in the dataset. The less correlated (or statistically dependent) the IVs are the more the outcome of the classification is inclined to be biased.

A major limitation in methods attempting to infer the transportation mode is the choice of IVs to be used for classification. For most studies the variables chosen were not based on any statistical evaluation justifying the variable choice being made. Most studies use variables such as length, speed, acceleration, maximum or median of speed or acceleration through a stage (Schüssler & Axhausen, 2009; Zheng et al., 2010), either together or alone for classification without providing a statistical basis for the choice. The correlation of the chosen IVs in these studies was neither accounted for. Therefore, in this section we conduct a statistical evaluation of different IVs that could discriminate between different classes (modes) in this classification problem. The outcome of the evaluation identifies the best IVs to be used for the classification. Sampling for this kind of GPS-based study could also be a pushing problem in the context of transport studies. Therefore, we discuss the sampling method used to identify the sample size and the period of such studies in the remaining part of this section.

3.1. ANOVA test for variable selection

Four potential variables were taken into consideration for the analysis; three of which are distance, speed and acceleration, which are highly inter-correlated where they all stem from one another. We also consider the change rate in heading (direction) as was suggested by a previous study (Stopher, Clifford et al., 2008a).

The testing sequence starts with group statistics to examine the differences between the categories on each of the independent variables using category means and ANOVA test. The mean differences between distance, acceleration and speed suggest that these may be good discriminators as the separations are large. This separation is clear in Fig. 2 representing the distribution across different modes as box plots for each variable in a separate plot. These 3 variables effectively discriminate the *walk* and *train* modes from the rest, as illustrated in the figures. However, acceleration appears to discriminate the *car* mode from the rest quite well. On the other



(a) Speed plot for different transportation modes

(b) Acceleration plot for different transportation modes

(c) Distance plot for different transportation modes

(d) Change in heading plot for different transportation modes

Fig. 2. Box plots for the values of different independent variables.

Table 3
Tests of equality of group means.

	Wilks' lambda	F	df1	df2	Significance
Distance (m)	0.896	441.820	5	18934	0.000
Speed (m/s)	0.486	4004.532	5	18934	0.000
Acceleration (m/s^2)	0.459	4462.582	5	18934	0.000
Difference in heading (°)	0.965	135.945	5	18934	0.000

Table 4
Tests of equality of group means results using different independent variables between *car*, *train* and all other transportation modes as a third category.

	Wilks' lambda	*F*	df1	df2	Significance
Distance (m)	0.970	293.326	2	18934	0.000
Speed (m/s)	0.540	8079.834	2	18934	0.000
Acceleration (m/s^2)	0.464	10951.800	2	18934	0.000
Difference in heading (°)	0.975	238.416	2	18934	0.000

hand, the rate of change in direction does not seem to significantly discriminate between any of the modes, except for the *train*, which could be caused by the fact that *train* trajectories follow fixed tracks for long distances (Fig. 2d).

Equality of group means results are presented in Table 3. In order to assess the discriminability of the different IVs two statistical measures are introduced: the Wilks' Lambda Λ and the Between-Groups *F*. The former is used in multivariate analysis of variance (MANOVA) to test whether there are differences between the means of identified groups of subjects on a combination of dependent variables (Everitt & Dunn, 1991). Wilk's Lambda is a statistic that takes into consideration both the differences between groups and the cohesiveness or homogeneity within groups (Klecka, 1980). However, a variable which increases cohesiveness without changing the separation between centroids may be selected over a variable that increases separation without changing the cohesiveness. When the IVs are considered individually, Λ is given from Eq. (2).

$$\Lambda = \frac{\text{Within Groups Sums of Squares}}{\text{Total Sums of Squares}} = \frac{w_{il}}{t_{il}} \tag{2}$$

$$w_{il} = \sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk}X_{ijk}X_{ljk} - \frac{\sum_{j=1}^{g}\left(\sum_{k=1}^{m_j} f_{jk}X_{ijk}\right)\left(\sum_{k=1}^{m_j} f_{jk}X_{ljk}\right)}{n_j} \tag{3}$$

$$t_{il} = \sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk}X_{ijk}X_{ljk} - \frac{\left(\sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk}X_{ijk}\right)\left(\sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk}X_{ljk}\right)}{n} \tag{4}$$

where g is the number of groups; p is number of variables; i, l is $1, \ldots, p$; X_{ijk} is value of variable i for case k in group j; X_{ljk} is value of variable l for case k in group j; f_{jk} is case weights for case k in group j; n_j is sum of case weights in group j; n is total sum of weights; m_j is the number of cases in group j.

In Table 3, large values of lambda indicate that group means are close, while small values are indicators of different means. Acceleration and speed seem to be the best discriminators in this case, with a small difference between their performances.

The second statistical measure used, the Between-Groups *F*, takes into consideration the sample size of the groups. This differs from a test that is solely based on squared distance (Klecka, 1980). Comparisons between small groups will be given less weight than comparisons between large groups. The advantage here is that this criterion will maximize differences between pairs containing larger groups. Acceleration and speed are still the best discriminators in this case; however, the difference between them is higher. This finding could be attributed to the sample size of the *car* mode having a high significance in manipulating the value of this statistical measure. Eq. (5) is used to calculate the *F* statistic based on another statistic called Mahalanobis Distance D^2, which is the distance between two groups (a and b) (Klecka, 1980) and is calculated from Eq. (6).

$$F = \frac{(n-1-p)n_1 n_2}{p(n-2)(n_1+n_2)} D^2_{AB} \tag{5}$$

$$D^2_{ab} = (n-g)\sum_{i=1}^{p}\sum_{j=1}^{p} w^*_{ij}(\overline{X}_{ia} - \overline{X}_{ib}) \times (\overline{X}_{ja} - \overline{X}_{jb}) \tag{6}$$

where n_z is the sample size of the group z; $\overline{X}_{ia}$ is mean of ith variable in group a; $\overline{X}_{ja}$ is mean of jth variable in group a; $\overline{X}_{ib}$ is mean of ith variable in group b, and $\overline{X}_{jb}$ is the mean of jth variable in group b.

It could be noted that speed appears to be a better discriminator for only some categories when calculating the Wilk's Lambda and the Between-Groups *F*. On the other hand, acceleration is better for most of the categories and/or for the categories of the highest sample sizes. From Fig. 2c we could also note that the *car* mode is better discriminated using acceleration rather than speed in Fig. 2a. We therefore ran the same analysis again but this time using only 3 categories namely: *car*, *train* and the rest of modes aggregated into one category. This categorisation was due to the natural division of the acceleration data illustrated in Fig. 2c. Table 4 shows the results of this second run, proving that acceleration produces a better discrimination of the 2 categories. It also performs much better than speed, while it yields a bigger difference than that shown in Table 3.

We could comfortably conclude from this statistical evaluation that speed and acceleration are the best IVs for discriminating between different transportation modes, given the specifications of the data collected in this research. We can also conclude that each variable is better at discriminating certain categories. On the other hand, using two variables that are highly correlated will bias the inference results. Section 5 discusses the results of the inference model using each of these IVs by quantifying the difference in the classification accuracy for each mode.

4. Inferring the transportation mode

This section highlights the method-related limitations in previous attempts to infer the transportation mode. This section also describes the framework used to classify the GPS segments into transportation modes. The framework is based on a SVM classification problem based on the speed and acceleration of the trajectory, as proven to be the best IVs due to the statistical evidence discussed in Section 3. The framework uses an innovative sliding window approach to learn and classify the data instances separately for each variable. A transition matrix is later applied to amend the sequence of consecutive trip stages. A segmentation process is applied afterwards, based on the idea that a *walk* stage mostly exists as a transition between every two other stages in any trip. This enables a further reasoning on the final classification of non-*walk* stages. A final stage of identifying underground travel is carried out, followed by an integration stage of all the results from the previous stages.

4.1. Method-related limitations

The range of the methods used to infer the transportation mode from GPS data has extended from logical procedural to Machine

Learning (ML) approaches in order to resolve a classification problem. Stopher, Clifford et al. (2008a) uses a process of elimination of different modes at different phases of the algorithm. Schüssler and Axhausen (2009) developed an open source fuzzy logic engine using the median of speed, the ninety-fifth percentile of the speed and the acceleration distributions as fuzzy variables. Several studies employ decision trees to perform this classification, either alone or integrated with other techniques, such as Hidden Markov Models (HMM) (Manzoni et al., 2011; Reddy et al., 2010; Stenneth et al., 2011; Zheng et al., 2010).

A slight limitation is that the majority of these studies only considers a limited number of transportation modes. Some use as few as 3 modes (Liao et al., 2007; Wang et al., 2010), while most studies exclude the train and underground modes. Others generalise the motorised modes together (Reddy et al., 2010), grouping bus and car modes.

A common practice is to start the process by segmenting the GPS track into trips, based on either a "dwell time" period (Stopher, Clifford et al., 2008a), a threshold of time without fix. Other studies go a step further by segmenting each trip into stages, identifying the change points of mode switches. However, some of these studies start by performing a stage-level segmentation and then perform the classification based on the identified stages (Schüssler & Axhausen, 2009; Zheng et al., 2010). This exerts a shortcoming in that the classification accuracy is highly reliant on the segmentation's efficiency. On the contrary, if a *car* stage was identified as two segments, based on the fact that it moved from a speedy main road to a highly busy street, the latter might be misclassified accordingly.

Other studies that are not dependant on segmentation classify each GPS segment individually into a transportation mode and not classifying the consequent segments as a block, i.e. the change in a trajectory's motion across several segments. Even studies that perform segmentation beforehand tend to ignore this consequence across the mode switch points.

Most studies also assume that any two stages are always separated by a *walk* stage. This, while true for most cases, might fail in cases of *cycling* or driving the *car* out of a *train* station's car park for example. A useful way to account for this is to use a transition matrix to verify the mode switch between consecutive stages according to a probability matrix of such switches (Zheng et al., 2010).

A couple of studies also use temporal information for mode inference. Liao et al. (2007) use the time of day to use in a probability model building assumptions about the participant's context. While this might be a useful technique to identify different activities, it might not be applicable to participants that have abnormal working hours for example. Stenneth et al. (2011), on the other hand, depends on live bus and train times information to make some inferences too, which would require a continuous input of such information for any period of time.

The framework we propose in this work is based on SVMs to classify GPS segments into respective transportation modes. An advantage of using SVMs over other ML methods is that they can be easily trained and are applied directly to the data without the need for a feature extraction process. This allows us to learn from the structure of the data. The proposed method uses a moving window across every group of consecutive segments in order to capture the nature of participants' movements though different transportation modes. We consider all the possible transportation modes, while testing the algorithm to avoid any mode aggregations or exclusions. A segmentation process is applied to the classified data after the initial SVM inference is performed to avoid the reliance on the segmentation accuracy if we have had applied the segmentation before the classification. We also avoid using any temporal assumptions to ensure the robustness of our algorithm over different samples. A transition matrix is also applied to assign modes in the case of potential transitions between any two non-walk stages. We finally enhance the detection of underground stages by using the underground station locations with any loss-of-signal incidents. The rest of this section provides a detailed account of our proposed framework and a summary of the chosen SVM model, while results and some discussions are presented in Sections 5 and 6.

4.2. Support Vector Machines (SVMs) classification and model selection

A Support Vector Machine (SVM) is a non-probabilistic binary linear classifier. A SVM constructs a hyper-plane or a set of hyper-planes in a high- or infinite-dimensional space to achieve the largest separation between different classes (Steinwart & Christmann, 2008). SVMs use an implicit mapping of the input data into a high-dimensional feature space, defined by a kernel function (a function returning the inner product $(\Phi(x), \Phi(x'))$ between the images of two data points x, x' in the feature space). The learning then takes place in the feature space, and the data points only appear inside dot products with other points.

The kernel functions return the inner product between two points in a suitable feature space, hence defining a notion of similarity. Kernel functions do this with little computational cost even in very high-dimensional spaces, since it does not involve any actual computations in that high-dimensional space, which is a major advantage of using SVMs. In this research, we use a Gaussian Radial Basis Function (RBF) kernel (Eq. (7)). The Gaussian and Laplace RBF kernel is a general-purpose kernel used when there is no prior knowledge about the data.

$$k(x, x') = \exp(-\sigma||x - x'||^2) \tag{7}$$

When classifying, Support Vector Machines separate the different classes of data by a hyper-plane contained by the decision function in Eq. (8).

$$f(x) = sign(\langle w, \Phi(x)\rangle + b) \tag{8}$$

And the SVM solution w has an expansion presented in Eq. (9), in terms of a subset of training patterns that lie on the margin. These training patterns, called support vectors, carry all relevant information about the classification problem.

$$w = \Sigma_i \propto_i \Phi(x_i) \tag{9}$$

The optimal hyper-plane (Vapnik, 1998) will be the one with the maximal margin of separation between two classes. In order to extend this binary SVM into the multi-class problem, there have been reformulations of the support vector quadratic problem that deal with more than two classes. One of these reformulations, introduced by Crammer and Singer (2000) and referred to as "spoc-svc", works by solving a single optimization problem including the data from all classes. The algorithm is presented in Eq. (10).

$$\text{Minimise} \quad t(\{w_n\}, \varepsilon) = \frac{1}{2}\sum_{n=1}^{k}||w_n||^2 + \frac{c}{m}\sum_{i=1}^{m}\varepsilon_i \tag{10}$$

$$\text{subject to}: \quad \langle\Phi(x_i), w_{yi}\rangle - \langle\Phi(x_i), w_n\rangle \geq b_i^n - \varepsilon_i \quad (i = 1, \ldots, m) \tag{11}$$

where the decision function is:

$$argmax_{n=1,\ldots,k}\langle\Phi(x_i), w_n\rangle \tag{12}$$

where m is the number of training patterns; C is the cost parameter.

The cost parameter C of the SVM formulation in Eq. (10) controls the penalty paid by the SVM for misclassifying a training point and thus, the complexity of the prediction function. A high cost value C will force the SVM to create a complex enough predic-

Data Instance	Label
X1	*Walk*
X2	*Walk*
...	...
Xi	*Bus*
...	*Bus*
...	...
Xn	...

(a) 1 segment instances

Data Instance			Label
X1	X2	X3	*Walk*
X4	X5	X6	*Walk*
X7	...	...	...
...	...	Xi	*Bus*
...	...	...	*Bus*
...	...	...	...
...	...	Xn	...

(b) Multi-segment instances

Data Instance			Label
X1	X2	X3	*Walk*
X2	X3	X4	*Walk*
X3	...	...	...
...	...	Xi	*Bus*
...	...	...	*Bus*
Xn-3	Xn-2	Xn-1	...
Xn-2	Xn-1	Xn	...

(c) Moving window multi-segment instances

Fig. 3. Division of data into equal-sized instances (three in this case).

tion function to misclassify as few training points as possible, while a lower cost parameter will lead to a simpler prediction function. The best C selected was found to be of value 3, where they generated the best results. This value is not too small where it allows less error in training (due to GPS errors), and since the data is very inseparable, yet it also is not too large that the model is over fit. A *k*-fold cross validation on the training data of value 3 is performed to assess the quality of the model (the accuracy rate for classification).

Another advantage of SVMs and kernel functions is that the selected kernel could be applied directly to the data without the need for a feature extraction process. This is particularly important in problems where a lot of structure of the data is lost by the feature extraction process (e.g. the sequence of a GPS trajectory's movements: such as the way a *car* can move fast, stop for traffic and then move again).

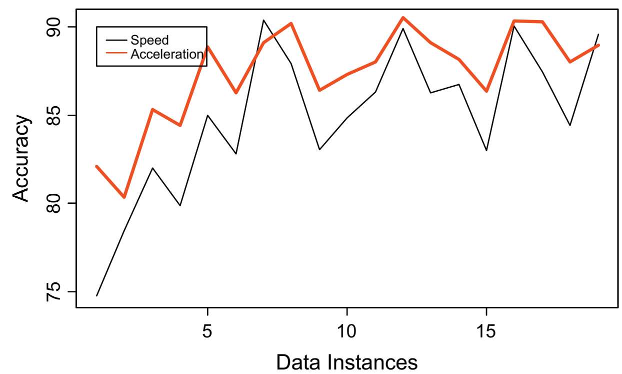


Fig. 4. SVM classification accuracies using different lengths of data instances.

4.3. Window-based SVM classification

The loss of GPS coverage due to indoor activity causes the track to be filled with long gaps with no movement till the first point that follows that gap. Therefore, the first step is to segment the track due to these gaps, as an initial segmentation process. The data is then ready and prepared for the SVM learning process and classification, being a supervised learning framework.

4.3.1. Multi-segment instance classification

As previously mentioned, a SVM constructs a hyper-plane in a high-dimensional space to achieve the largest separation between different classes, where the higher the dimension, the better the separation. Consequently, SVM maps original finite-dimensional space into a much higher-dimensional space to increase the separation. In this work, we enter the classification with more than one dimension in order to have a far better separation to start with.

Since we only have one dimension to begin with (speed or acceleration), we aim to simulate a multi-dimensionality to study sequences of GPS trajectory movements rather than each segment on its own (e.g. the stop and go motion of a car due to traffic). Therefore, the data is divided into equal-sized instances of several segments as demonstrated in Fig. 3b. This simulates the multi-dimensionality of the data in the learning process which is an advantage of SVM where it eliminates the need for feature extraction, as mentioned in the previous subsection. The main reason for using instances is that it is more meaningful to study a certain stage of a trip than one single segment value; this exposes the learning process to consequent GPS data that represent the variability in one's manner when undertaking each transportation mode.

For that purpose, the data is divided into two thirds for learning and one third for validation purposes. Data instances are then formed out of the learning data. The data instances then enter the SVMs learning process using the stationary Gaussian kernel with a radial basis kernel function (RBF) using the multi-class method.

Fig. 4 shows several window (instance) sizes that were tested. A suitable size from 3 to 8 segments was identified to be the most adequate. As might be noted, the classification gives better results for longer data instances. However, a longer sequence of mixed transportation modes could introduce higher complexity, since the probability of having several modes within one instance is introduced, which will over-complicate the classification problem. Therefore, we chose to use the small-sized instance that still contains a decent number of segments to represent a realistic sequence; in this case three.

The multi-segment instance classification achieves around an 80% inference accuracy using either speed or acceleration. This is shown in the confusion matrix in Table 5, where the red colour lightness varies according to accuracy of the classification for the diagonal axis (darker lightness (e.g. car) reflects higher classification accuracy than brighter lightness (e.g. bus)). The rest of lightness variance in Table 5 reflects the confusion in classification between different classes, with a darker lightness reflecting a higher confusion (e.g. nearly 40% of bus mode class is classified as car). There appears to be a good discrimination between the train mode and the rest, yet having a great confusion with the bus mode. The other classes seem to perform well, except the bus and tube modes, since the latter often consists of only one segment and therefore, it is merged into stages that are dominated by other modes. The classification, however, is non-realistic due to the assumption that the track is segmented into similar-mode stages.

4.3.2. Moving window SVM classification

In order to allow going into the segment level rather than merging different modes into the same stage, we applied a fixed-length moving window on the whole track; sliding that window segment-by-segment along the track's speed values once and once more for acceleration. Every time the window slides, a classification of that instance of data is performed. Figs. 3c and 5 illustrate this process, where a moving window classifies each 3-sized instance moving segment-by-segment along the track.

Table 5
Confusion matrix for classification of instances of 3 segments.

Classification	Actual						Total count
	Bus	Car	Cycle	Train	Tube	Walk	
Bus	27.03%	6.30%	26.01%	52.88%	11.11%	0.35%	180
Car	39.86%	76.72%	11.56%	2.88%	55.56%	2.12%	523
Cycle	25.00%	9.35%	57.80%	0.00%	18.52%	0.09%	192
Train	0.68%	0.38%	0.00%	44.23%	0.00%	0.00%	49
Tube	4.05%	4.96%	0.58%	0.00%	3.70%	0.27%	37
Walk	3.38%	2.29%	4.05%	0.00%	11.11%	97.17%	1125
Total count	148	524	173	104	27	1130	2106

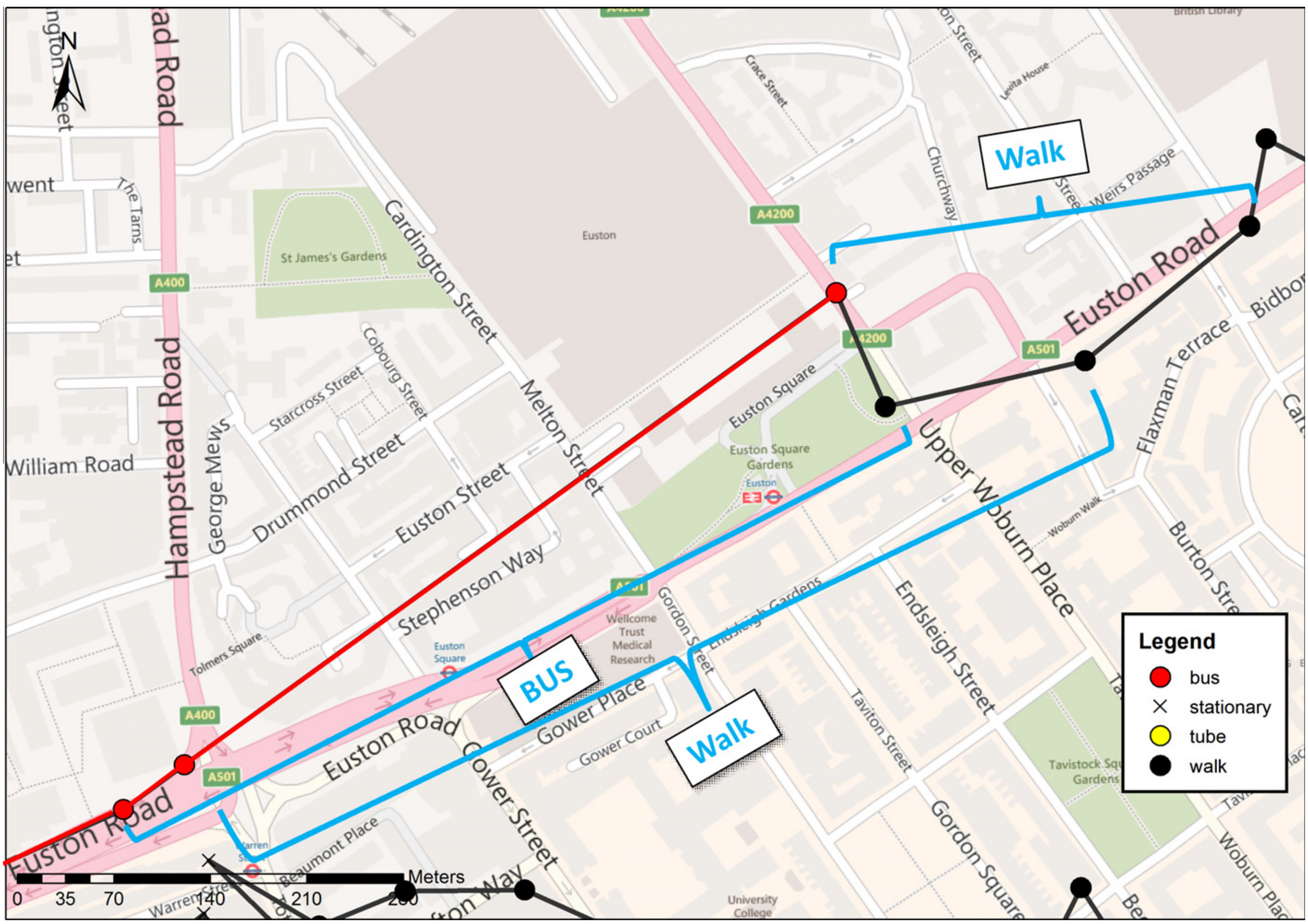


Fig. 5. A moving window classifying each 3-segment instance moving segmen t-by-segment along the track.

Once the classification is performed on the instances, the classification is passed over to the segment level and the change points in the track are identified. The change points are initially identified as any two consequent instances with different modes; the first mode being *a* and the second *b*. Then, the algorithm mines into the last instance with the mode *a* and assigns the classification of the first and second segments as *a*, and the third's as *b*. The same happens with the first instance with the *b* mode, passing the classification of the first segment as mode *a* and the second and third segments as mode *b*.

4.4. Verification by track segmentation

The framework then applies a verification process to each classified arc. It does this by applying two processes iteratively. The first of these two processes runs through each change point in the segment level and assesses the probability of mode *a* and *b* following each other according to a transition matrix (Table 6). This matrix is based on Zheng et al. (2008) and is compiled from this research's data. The matrix contains the different probabilities of switching between every two modes, which is a good indication of the natural flow of modal mixes.

As could be noted from Table 6, almost all modes are followed by a *walk* mode. Therefore, the algorithm then segments the track into several stages, where every two different modal stages are separated by a *walk* stage. However, some stages will have two or more modes. In this case, the most dominant mode will be assigned to the whole stage, unless in the case of two modes, the ratio is less than 1:2 between the segments of *a* and *b* or vice versa. This creates a continuous flow of modes along different periods of the track.

Table 6
Transition matrix between modes showing probabilities of different modal mixes occurring (%).

Transportation modes	Bus	Car	Cycle	Train	Tube	Walk
Bus		0.9	0	0	0	99.1
Car	0		0	2.2	0	100
Cycle	0	0		1.3	0	97.8
Train	7.1	0	7.1		0	85.7
Tube	0	0	0	1.5		98.5
Walk	29.5	37.3	11.8	3.2	18.2	

4.5. Integrating speed and acceleration results

As we demonstrated earlier in Section 3, we use two IVs to conduct this classification; namely speed and acceleration. Therefore, we run the classification framework once for speed and once for acceleration assessing the performance of each of the variables in the process. We integrate the results of the best mode results, obtained from one variable, with the best from the other. This relies on the fact that each variable would be a better discriminator for some modes over the others. Section 5 describes this integration in details along with the results obtained from each variable.

4.6. Underground segment identification

In London, the occurrence of the "*tube*" (underground) mode could be overground, in which case would be identifiable at the SVM classification process from speed, or underground, where an alternative process needs to be in place to identify these instances. Around 45% of the *tube* network is underground, however, 75% of the trips are done within the central zones where the network is underground, which means that in many cases there will not be any GPS fixes attainable at several areas and will not be identified using the SVM classification stage.

A statistically-driven time-distance threshold between *tube* fixes is tested for travel occurring above a certain distance without coverage though not exceeding a time threshold. A further verification test is applied where the suspected *tube*-entrance and *tube*-exit fixes are tested for their proximity to any *tube* stations. These, along with the overground classified *tube* segments constitute all the *tube*-classified segments within the framework.

5. Results

Building on our previous work, we increased the participant numbers to 81 (as recommended in Section 2), and considered acceleration with the speed for this classification problem. The results of the moving window algorithm using speed reveal an accuracy of 72% and 83% using acceleration. This demonstrates a considerable improvement over both the previous accuracies of the multi-segment instances, without applying a moving window approach. It also has the advantage of classifying on a segment level-basis, rather than only classifying instances. Tables 7 and 8 show the confusion matrices of this classification using speed and acceleration respectively. Some speed classification errors could be noted, such as the *car* mode with other transportation modes, while some modes, such as the *walk* mode, seem to be better classified using speed.

In order to get a better accuracy measure for the classification, we perform an inter-reliability analysis using the Kappa statistic to determine consistency among coders. Cohen's Kappa is generally thought to be a more robust measure than simple percentage agreement calculation, since K takes into account the agreement occurring by chance (Carletta, 1996). The Kappa coefficient (K) measures pairwise agreement among a set of coders making category judgments, correcting for expected chance agreement, and hence is thought to be a good measure of any classification's accuracy. Kappa is calculated from Eq. (13).

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{13}$$

where $P(A)$ is the proportion of times that the coders agree; $P(E)$ is the proportion of times that the coders we expect them to agree by chance.

As illustrated in Table 9, the inter-rater reliability for speed was found to be 0.586 ($p < 0.001$), 95% CI (0.578, 0.594) and for acceleration 0.743 ($p < 0.001$), 95% CI (0.735, 0.751). That is to say, K val-

Table 7
Confusion matrix of the moving window algorithm based on speed.

	Actual						Total count
Classification	Bus	Car	Cycle	Train	Tube	Walk	
Bus	31.96%	6.84%	5.09%	0.00%	6.15%	4.98%	1348
Car	43.36%	63.32%	7.35%	12.34%	39.75%	8.43%	4742
Cycle	16.28%	1.57%	85.31%	0.97%	18.03%	6.75%	2354
Train	0.60%	20.41%	0.13%	81.01%	2.05%	0.83%	1815
Tube	1.88%	4.38%	1.10%	5.69%	30.33%	0.12%	387
Walk	5.93%	3.48%	1.03%	0.00%	3.69%	78.90%	8290
Total count	1333	4708	1552	932	244	10167	18936

Table 8
Confusion matrix of the moving window algorithm based on acceleration.

	Actual						Total count
Classification	Bus	Car	Cycle	Train	Tube	Walk	
Bus	42.31%	0.66%	8.12%	0.00%	18.85%	2.59%	1030
Car	18.23%	87.26%	10.50%	12.77%	18.44%	4.77%	5163
Cycle	21.83%	1.10%	76.87%	1.93%	11.07%	1.91%	1775
Train	0.15%	5.40%	0.06%	84.01%	0.00%	0.35%	1076
Tube	3.53%	2.95%	0.90%	1.29%	33.20%	1.41%	436
Walk	13.95%	2.63%	3.54%	0.00%	18.44%	88.97%	9456
Total count	1333	4708	1552	932	244	10167	18936

Table 9
Symmetric measurements for cohen's Kappa values for speed and acceleration.

IV		Value	Asymp. Std. error[a]	Approx. T^{b}	Approx. sig.
Speed	*Measure of agreement Kappa*	0.586	0.004	139.899	0.000
	N of valid cases	18936			
Acceleration	*Measure of agreement Kappa*	0.743	0.004	167.029	0.000
	N of valid cases	18936			

[a] Not assuming the null hypothesis.
[b] Using the asymptotic standard error assuming the null hypothesis.

Table 10
Difference between confusion matrices of the moving window algorithm based on results from acceleration and speed.

	Actual					
Classification	Bus	Car	Cycle	Train	Tube	Walk
Bus	10.35%	−6.18%	3.03%	0.00%	12.70%	−2.39
Car	−25.13%	23.94%	3.16%	0.43%	−2131%	−3.66
Cycle	5.55%	−0.47%	−8.44%	0.97%	−6.97%	−4.84
Train	−0.45%	−15.02%	−0.06%	3.00%	−2.05%	−0.47
Tube	1.65%	−1.42%	−0.19%	−4.40%	2.87%	1.29
Walk	8.03%	−0.85%	2.51%	0.00%	14.75%	10.07
Total count	1333	4708	1552	932	244	10167

ues reflect a moderate agreement for speed and a substantial agreement for acceleration, according to rule of thumb values of Kappa (Landis & Koch, 1980).

5.1. Type I and II errors

Table 10 shows the difference between the accuracies obtained from classification using acceleration and speed classification. The red values express the excellence of acceleration classification over speed and vice versa for the blue values. As previously noted, it seems very obvious that some modes, such as the *car* mode, are better identified using acceleration and less misclassified as other modes. On the other hand, other modes, such as *walk*, are less confused for using speed. On the other hand, the rest of the modes have little difference in results from using either IVs. An interesting confusion, however, occurs between the *train* and the *car* results, where speed appears to have a better performance by not classifying *train* as *car* (8% better), while acceleration performs better without confusing *car* instances with a *train* classification (20% better).

It could also be noted from the main diagonal of Table 10 that *bus*, *car* and *walk* are better classified using acceleration, while *cycle*, *train* and *tube* are higher classified using speed. This does not mean that each group of modes should be classified using their respective IV, but it only suggests that they are over classified using these specific IVs. The trick here is to select the IV that better discriminates the classified mode from the rest. That is to be

Table 11
Differences between of acceleration and speed in terms of Type I and II errors.

Mode	Type I error difference	Type II error difference	Type I & II errors difference
Bus	−10.35%	7.16%	−3.19
Car	−23.94%	−46.52%	−70.45
Cycle	8.44%	−5.76%	2.68
Train	−3.00%	−18.05%	−21.06
Tube	−2.87%	−3.08%	−5.95
Walk	−10.07%	24.44%	14.37%
Average	−6.97%	−6.97%	−13.93%

Blue (+ve) valuG5 demonstrate Speed'5 excellence.
Red (-ve) values demonstrate Acceleration's excellence.

achieved by selecting the IV that achieves a higher classification for each mode, while not over-classifying that specific mode and hence, decreasing the accuracy of the other modes. This would also have the advantage of accounting for the effect of the sample size of each mode. This could be achieved by testing whether a certain variable on average dominates the row and column of each mode in Table 10 (actual and classified mode), while if acceleration dominates in the column level (Type I error) but speed dominates in the row level (Type II error) (such as *walk*), that would mean that acceleration is only over-classifying that specific mode. This calculation results in Table 11, where each mode is assessed for the Type I and II errors' excellence of one IV over the other, given that red represents an excellence of acceleration and blue for speed.

5.2. Integration results

As can be noted from Table 11, acceleration seems to produce better results than speed for most transportation modes with the exception of *walk* and *cycle*, achieving an average supremacy of nearly 14% accuracy over speed. The inter-rater reliability for the raters was found to be $K = 0.802$ ($p < 0.001$), 95% CI (0.794, 0.810), which reflects almost perfect agreement. We adopted these results into our final integrated result of the inference, resulting in an accuracy of 88%. Table 12 shows the confusion matrix of this integration, demonstrating a better separation specifically for the *car*, *train* and *walk* modes.

Some modes appear to be performing better than others. We could note from Table 12 that the *car*, *train* and *walk* modes are discriminated very well using this classification. In contrast, the *cycle* mode seems to be classified moderately while the *bus* and *tube* modes still require enhancement. This could be carried out using a network matching process to both the *bus* and *tube* networks. This further work is currently in process for enhancing the classification of these two latter modes.

6. Conclusions

In this work we discuss the classification problem of inferring transportation mode from sparse GPS data. We first provide the

Table 12
Confusion matrix of the moving window algorithm based on integrating acceleration and speed.

	Actual						Total count
Classification	Bus	Car	Cycle	Train	Tube	Walk	
Bus	58.29%	1.02%	11.73%	0.00%	19.67%	1.54%	1212
Car	15.75%	88.47%	9.21%	12.77%	10.25%	2.57%	4923
Cycle	10.58%	1.08%	75.19%	1.93%	6.15%	1.41%	1535
Train	0.15%	4.76%	0.06%	84.01%	0.00%	0.11%	1021
Tube	3.00%	1.72%	0.58%	1.29%	45.49%	0.55%	309
Walk	12.23%	2.95%	3.22%	0.00%	18.44%	93.82%	9936
Total count	1333	4708	1552	932	244	10167	18936

means for assessing the significance of each potential independent variable that could be used for this process. We provide a statistical evaluation using the data collected by this research within Greater London as a case study. The outcome of this process provides evidence that speed and acceleration are the favourable candidates to undergo this classification problem showing a great discriminatory power in this context. However, each of these variables is also proven to be fit for identifying certain modes; *car* mode being better identified using acceleration and *walk* using speed as examples. We also provide a brief summary of the sample size calculation process required for the context of this study.

Building on previous attempts and on the results of the statistical evaluation, we provide in this study a novel approach for inferring the transportation mode from sparse GPS data without any extra information. In contrast to existing techniques, our approach uses one consistent framework based on Support Vector Machines (SVMs) to classify each segment into its respective transportation mode. Unlike many previous attempts, the framework tends to study the whole pattern of the trajectory motion during the whole trip using the advantage of being an offline process. The framework does this by first classifying several consequent segments together (named as an instance) with a certain window size, and sliding this window along the whole track classifying each instance. The most adequate window size was found to be of 3 segments length. The classification is then assigned to the segment level to each of the segments participating in each instance.

In order to preserve the cohesiveness of the classification of the track, we segment the track into stages of different modes, each two stages separated by a *walk* stage, except for certain scenarios where we then apply a transition matrix to assess the modal mix occurrence probability. A statistically-driven time-distance threshold between *tube* fixes is then applied for travel occurring above a certain distance without coverage, though not exceeding a time threshold to identify the *tube* mode.

Our model achieves relatively good accuracies using either speed or acceleration. However, building on the findings of the statistical evaluation and the SVM classification, results from the classification using both speed and acceleration are combined together. This is based on the fact that each variable is better at classifying certain modes. Finally, an accuracy of 88% is achieved from the combined result at segment level with a Kappa statistic reflecting almost perfect agreement. A good segmentation is also achieved between different modal stages; which enhances the accuracy of the classification.

Further work shall be carried out attempting to further separate similar modes, such as *bus* and *tube* modes using network matching from the rest. The accuracy of the device being used and its firmware also appeared to have some effect on the classification results that could be explored in further work. Finally, different rate of GPS data collection could be compared to perform this classification.

Acknowledgments

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.

References

Anderson, T., Abeywardana, V., Wolf, G., & Lee, M. (2009). National travel survey GPS feasibility study: Final report. Department of Transport, UK, Version 4.0, December 2009.

Bolbol, A., & Cheng, T. (2010). GPS data collection setting for pedestrian activity modelling. In *GISRUK 2010: Proceedings of Geographical Information Science Research UK Conference 2010*. UCL, London.

Bolbol, A., Cheng, T., & Paracha, A. (2010). GeoTravelDiary: Towards online automatic travel behaviour detection. In *WebMGS: 1st International Workshop on Pervasive Web Mapping, Geoprocessing and Services*. Politecnico di Milano, Como, Italy.

Bolbol, A., Cheng, T., Tsapakis, I., & Chow, A. (2012). Sample size calculation for studying transportation modes from GPS data. In *Proceedings of Transport Research Arena*. Athens, Greece, April 23–26.

Bricka, S., & Bhat, C. (2006). Comparative analysis of global positioning system-based and travel survey-based data (pp. 9–20). *Transportation Research, Record No*. 1972.

Cambridge systematics (1996). *Scan of recent travel surveys*. US DOT Report DOT-T-97-08, June.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2), 249–254.

Chung, E., & Shalaby, A. (2005). A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology, 28*(5), 381–401.

Crammer, K., & Singer, Y. (2000). On the learnability and design of output codes for multiclass problems. *Computational Learning Theory*, 35–46.

Edwards, D., Griffin, T., Hayllar, B., Dickson, B., 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.

Everitt, B. S., & Dunn, G. (1991). *Applied multivariate data analysis*. London, England: Edward Arnold.

Klecka, W. R. (1980). *Discriminant analysis*. CA, Sage: Beverly Hills.

Landis, J. R., & Koch, G. G. (1980). *Xu. Biometrics, 33*, 159–174.

Liao, L., Fox, D., & Kautz, H. (2007). Extracting places and activities from GPS traces using hierarchical conditional random fields. *The International Journal of Robotics Research, 26*(1), 119.

Manzoni, V., Maniloff, D., Kloeckl, K., & Ratti, C. (2011). *Transportation mode identification and real-time co2 emission estimation using smartphones, SENSEable City Lab*. Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, Technical Report.

Mitchell, T. (1997). *Machine learning*. McGraw-Hill.

Ortúzar, J. D., & Willumsen, L. G. (2011). *Modelling transport* (4th ed.). London, England: John Wiley & Sons Inc..

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2010). Using mobile phones to determine transportation modes. *Transactions on Sensor Networks, 6*(2).

Schüssler, N., & Axhausen, K. W. (2009). Processing GPS raw data without additional information. *Transportation Research Record: Journal of the Transportation Research Board, 2105/2009*(January), 28–36.

Smith, M. E. (1979). Design of small sample home interview travel surveys. *In Transportation Research Record: Journal of the Transport Research Board, 1972*, 29–35.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. 978-0-387-77241-7. New York: Springer-Verlag.

Stenneth, L., Wolfson, O., Yu, P. S., & Xu, B. (2011). Transportation mode detection using mobile phones and GIS information. ACM SIGSPATIAL GIS '11, November 1–4, 2011. Chicago, IL, USA.

Stopher, P. R. (2008). Collecting and processing data from mobile technologies. In *The 8th international conference on survey methods in transport*, Annecy, France.

Stopher, P., Clifford, E, Zhang, J., & FitzGerald, C. (2008a). Deducing mode and purpose from GPS data. In *The 87th annual meeting of the transportation research board*, Washington, DC, United States.

Stopher, P., Kockelman, K., Greaves, S., & Clifford, E. (2008b). Sample size requirements for multi-day travel surveys: Some findings. In *The 8th international conference on travel survey methods*, Annecy, France.

TFL (2011). *Network operating strategy NOS issue 5 public consultation*. Transport For London, May.

Thompson, D. C., Rebolledo, V., & Thomson, R. S. (1997). Bike speed measurement in a recreational population: Validity of self-reported speed. *Injury Prevention, 3*, 43–45.

Tsapakis, I., Turner, J., Cheng, T., Heydecker, B. Emmonds, A., & Bolbol, A. (in press). Effects of tube strikes on journey times in the transport network of London. *Transportation Research Record*, TRB, National Research Council, Washington, DC.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Wang, H., Calabrese, F., Di Lorenzo, G., & Ratti, C. (2010). Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Proceedings of the 13th international IEEE conference on intelligent transportation systems*, Washington, DC, USA, 2010. IEEE Computer Society.

Weidmann, U. (1993). *Transporttechnik der Fussgaenger*. ETH Zuerich, Schriftenreihe IVT-Berichte 90, Zuerich (In German).

Xu, Y. (2010). Effective GPS-based panel survey sample size for urban travel behaviour studies. Thesis (PhD). Georgia Institute of Technology: School of Civil and Environmental Engineering.

Zheng, Y., Chen, Y., Li, Q., Xie, X., & Ma, W. (2010). Understanding transportation modes based on GPS data for Web applications. *ACM Transaction on the Web, 4*(1), 1–36.

Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008). Learning transportation mode from raw GPS data for geographic applications on the Web. In *Proceedings of WWW* 2008 (pp. 247–256). ACM Press.