Classification of LiveBus Arrivals User Behaviour

Natalia Selini Hadjidimitriou[a], Marco Mamei[a], Mauro Dell'Amico[a], Ioannis Kaparias[b],

[a] Department of Sciences and Methods for Engineering, University of Modena and Reggio Emilia, Italy

[b] Faculty of Engineering and the Environment, University of Southampton, United Kingdom

Contact:

Natalia Selini Hadjidimitriou: selini@unimore.it

Abstract

With the increasing use of Intelligent Transport Systems, large amounts of data are created. Innovative information services are introduced and new forms of data are available, which could be used to understand the behaviour of travellers and the dynamics of people flows. This work analyse the requests for real time arrivals of bus routes at stops in London made by travellers using Transport for London's LiveBus Arrivals system. The available dataset consists of about one million requests for real time arrivals for each of the 28 days under observation. These data are analysed for different purposes. LiveBus Arrivals users are classified based on a set of features and using K-Means, Expectation Maximization, Logistic regression, One-level decision tree, Decision Tree, Random Forest, and Support Vector Machine (SVM) by Sequential Minimal Optimization (SMO). The results of the study indicate that the LiveBus Arrivals requests can be

classified into six main behaviours. It was found that the classification-based approaches produce better results than the clustering-based ones. The most accurate results were obtained with the SVM-SMO methodology (Precision of 97%). Furthermore, the behaviour within the six classes of users is analysed to better understand how users take advantage of the LiveBus Arrivals service. It was found that the 37% of users can be classified as interchange users. This classification could form the basis of a more personalised LiveBus Arrivals application in future, which could support management and planning by revealing how public transport and related services are actually used or update information on commuters.

Keywords: Intelligent Transport System, Public Transport, Real Time Information, Clustering algorithms, Supervised classification, Data Mining, Data-driven behaviour.

Introduction

In order to offer public transport services that meet citizens' needs, public authorities and transport planners need to offer new services that allow users to easily access information about bus services. According to Lyons & Harman (2002), information addressed to public transport users is of fundamental importance to not only attracting new users to the service, but also to keep the existing ones loyal. One of the latest innovative services in this context gives users the possibility to know, in real time, the arrival of a bus at a stop. This service is provided using Automatic Vehicle Location (AVL) data, which enables the localisation of a bus in real-time. On this basis, an algorithm forecasts the arrival time of a bus at a stop using the information on the current position on the road network. In London, this information is provided at bus stop by the Countdown service. The same information can also be retrieved using a PC or a smartphone. The service provided by the local public transport authority, Transport for London (TfL), is called LiveBus Arrivals.

Besides its intended use of providing information to travellers, however, the LiveBus Arrivals service, just like other systems of its kind, also offers a valuable "side-product", as the requests' database have the potential to provide an insight into customer behaviour. There is, hence, great interest among, private and public stakeholders to develop new processes of exploiting available data, so as to understand the behaviour of customers and, based on this knowledge, improve their products and services. The present study, therefore, proposes a classification of users of the LiveBus Arrivals service and a new way of extracting information on users' behaviour from available data. The result of such a classification reveals how the users interact with the

application and, as such also, with the public transport system generally. Consequently, the analysis could support the introduction of new application functionalities and novel services based on the devised user profiles. Furthermore, the results can reveal how the public transport system is actually used so that it can provide useful statistics and guidelines for service planning and management. For instance, a high number of requests for real time information for a particular bus stop may reveal that the service offered in a section of the bus network needs to be improved (e.g. because of frequent delays caused by overcrowding during specific time windows). Finally, results could help automate, enrich or complement the census data on the origin and destination (OD) flows between different London boroughs.

Along these lines, the contribution of this work is twofold:

- A classification of LiveBus Arrivals users is performed on the basis of the frequency and type of requests in the database. For example, this could include users querying the same stop repeatedly over a short period of time, which could indicate that they are interested in a particular stop and a particular bus line, or users querying different stops and at various times, which may suggest that they explore alternative routes and itineraries. Several classifiers are trained to classify users in the data in the resulting classes.
- The result of the classification is then analysed both from a temporal and from a spatial perspective to identify the main peak hours during which different user profiles access the system (thus also revealing when the public transport system is mainly used and by which user classes). From a spatial perspective, on the other hand, the results of the classification can describe the outflows of workers who commute by bus from each

borough of London. The analysis is also complemented with census information about the mobility in the city.

Overall, the results of this analysis could provide guidelines to improve the LiveBus Arrivals service and, more in general, management and planning of public transportation and related services.

The remainder of the paper is organised as follows. Section 2 outlines the existing literature on the impact of real time information on the users and on data analysis techniques to study the behaviour of travellers. Section 3 consists of the description of the available dataset. Section 4 outlines the different classes of the LiveBus Arrivals users identified thanks to the analysis of raw data, provides an example of their representation in a Geographic Information System.
(GIS) and shows the implemented methodology to automatically classify the entire dataset. Section 5 presents the classification results and discusses the results showing how they could have an impact in the LiveBus Arrivals service and in the management of public transport. Finally, Section 6 concludes and reports the future research directions.

2. Related work

We report hereafter the most relevant contributes from the literature, related to our study and organized into four main categories.

2.1 The impact of Real Time Information

Much recent research has focused on the evaluation of the impact of Real Time Information (RTI) on traveller satisfaction and on their perception of service quality. The importance of predicted service was already evidenced by Zeithaml et al. (1993) who analysed different types of expectations and investigated several categories of service components by conducting focus groups.

The study distinguished between three types of customer expectations (desired service, adequate service and predicted service) and underlined the importance of predicted service in influencing how the customer evaluates the gap between the desired level of service and the level of service they are willing to accept (adequate service). Fonzone et al. (2016) describe the importance of Information and Communication Technologies in the era of big data. They evidence that the provision of RTI affects travellers' route and stop choices during the entire duration of the journey, especially in case of service disruption. The relation between the availability of information on travel time and travellers route choices has been explored in a laboratory experiment described in Tanaka et al. (2014). They found that information on travel time significantly affect travellers' decision making process.

2.2. The impact of RTI on public transport

In the public transport sector, the causal relationships between the factors that affect behavioural intentions have been studied by Lai and Chen (2010) who applied a structural equations model using the results of a survey carried out in Taiwan on a newly introduced transit service. They found that service quality had a positive effect on the perceived value of the service, overall satisfaction, involvement and behavioural intentions.

Several studies have assessed the impact of RTI on public services' quality using different methodologies. Brakewood et al. (2014), for instance, carried out a survey to evaluate the benefits of using a real time information service on bus arrivals on waiting time perception, service satisfaction and behaviour change. They provided to a sample of travellers a real time information service and compared a set of indicators with the ones measured on a controlled group. Among the main findings, it was noticed that the experimental group experienced a decrease of the perceived waiting time of nearly 2 minutes compared to the control group. The performance and reliability of RTI systems on expected waiting time has been measured by Cats and Loutos (2015). The comparison between the predictions and the real position of the vehicle allowed them to conclude that RTI systematically yields to underestimate the expected waiting time by 6.2% on average.

The casual relations between the real time information service, travel behaviour and psychology have been analysed by Zhang et al. (2008) using panel data who concluded that, although an increase of ridership cannot be expected immediately after the introduction of a RTI service, a change in the behaviour can be expected in the long term. Similarly, Watkins et al. (2011) carried out a survey to analyse the waiting time perception of bus riders. They developed a model to measure the perceived waiting time and found that the use of RTI allows reducing, not only the perceived waiting time by about 2 minutes, but also the actual waiting time because users can check the time of arrival before being at the stop. Furthermore, Tang and Thakuriah (2012) use longitudinal data to analyse changes in bus usage before and after the introduction of a bus tracking system. For this purpose, they implemented a linear mixed effect model and found

that the use of a bus tracker did increase the use of bus in the city of Chicago although this change was small.

2.3. Mobility Patterns with Intelligent Transport Systems

Several studies focused on the possibility to replace manual surveys via automated data collection and analysis. The analysis of people dynamics over time is important because it may help to understand the reactions due to external factors such as a disaster, a football match, the introduction of new public transport services, as explained by Ratti et al. (2006).

Recently, the intensive use of mobile phones has offered the possibility to continuously track people and study their behaviour. Taking advantage of this opportunity, Becker et al. (2011a) classified phones usage to explain city dynamics. In a second publication of the same year, Becker et al. (2011b) analysed drivers' route choices using mobile phone data and presented an algorithm to match the data to the road network with the final aim of measuring the traffic volumes. The results of the analysis were successively validated using the statistics issued by the transportation authority.

As can be seen from the literature, several studies have focused on the analysis of OD flows in public transport and have explored the possibility to use mobile phone data to identify the weak links of the road network (i.e.Holleczek et al. (2014) or GPS coordinates to analyse and describe the patterns that characterise people behaviour (i.e. Jiang et al. (2009) and Liu et al. (2012). Positioning data have been deployed to have information on the current and historical position and predict the place a user will visit (i.e. Noulas et al. (2012) or to analyse travellers' behaviour between different origins and destinations of the city based on mobility purposes such as work,

shopping, etc. (i.e. Liu et al. (2013) and Yuan et al. (2012)). The work proposed by Ahas et al. (2010), for instance, employed mobile phone data from the city of Tallinn to classify travellers (i.e. housewives, working wives, commuters, etc.) based on different rhythms and city zones (i.e. city centre, suburbs, etc.). Similarly, Sevtsuk and Ratti (2010) analysed the mobility in the city of Rome at different hours, days and weeks using the volume of calls activity. They focused on longitudinal activity patterns and found that differences in mobility patterns could be explained by demographic and built environment variables.

The requests made by the users on the real time arrivals at a bus stop have been, until now, analysed mainly to understand the influence of this type of information on the use of bus transport (Tang and Thakuriah (2012)) and on the estimation of the perception of waiting time (Dziekan and Kottenhoff (2007)).

2.4. Limitations and further research

Using mobile phones for the analysis of people dynamics has several limitations. A study on the variation of non-vehicular mobility within the city using mobile phone traces performed by Calabrese et al. (2013) identified two main limitations: the lack of knowledge on mobile phone users and the localisation error that affects the statistical results. For this reason, automatically collected data may be deployed to complement existing surveys or may be validated using data collected with standard methodologies.

The only work that analysed LiveBus Arrivals requests has been presented by Hardy (2012) who showed the statistics on the number of requests sent by users by time of the day, location and type of channel used to retrieve the information (Personal Computer, a smart phone or the Short

Message Service (SMS)). The study underlined that the demand for real time information was higher during peak hours meaning that the service was mostly used when waiting time was longer.

The analysis of existing literature showed that the majority of works have been focused on the development of methodologies aimed at analysing the behaviour of travellers based on two main sources of information: mobile phones and GPS data. Thus, no studies have focused on analysing data for the purpose of jointly studying the mobility patterns of travellers and the RTI application usage. Based on this motivation, the next section describes the dataset employed in this study.

3. Dataset

The dataset used in the present work consists of all requests made by users to the LiveBus Arrivals service that provides information on the real time arrivals of buses at stops in London. The period under examination is between the 16th July 2012 and the 12th August 2012. The size of the raw data is about 100 MB for each observed day and consists of requests made either by smartphones, browsers or third-party applications.

The dataset consists of a text file for each day. Each record in these files includes the IP (Internet Protocol) address, the time and the day of the request, the HTTP (HyperText Transfer Protocol) request, the web address, the mobile phone model, its configuration and vendor ID. An example of a data record is shown in Figure 1. The example shows a request for real time arrivals at Chatteris Avenue (stop number 72979), located in the borough of Harrow. Using a parser which recognises the name of the webpages containing the different information, it is possible to extract the number of the stop or bus.

Figure 1: Raw data. The query encodes a request for real time arrivals at stop number 72979. The line shows the IP address, the timestamp, the HTTP request, the web address, the mobile phone model together with its configuration and vendor ID.

No information on the specific user is available neither it is possible to unequivocally identify the same user on different days of the observed period, nor therefore the database contains no sensitive data. These data are imported into the relational database, called Postgresql 9.3, and the final dataset consists of about 77.7 million requests for RTI, 31.8 million of which are requests for stop arrivals and 45.8 million are requests for information on bus routes. On a daily average, about 1 million requests have been sent by the users of the LiveBus Arrivals service. The total number of interaction blocks is 2.2 million for the entire period.

A screenshot of extracted data is provided in Figure 2 in which four different users have sent requests for real time arrivals, either for a bus line or for arrivals at a stop.

Figure 2: Extracted Data. Multiple interactions of the same IP within a time limit are arranged in interaction blocks.

For instance, three repeated requests for the time of arrival at Charing Cross Station (75034) have been asked by a user (31.107.126.168) on an early morning (6:34). The IP address identifies the single user when the requests are sent using a PC. When the requests are sent using the same mobile model and IP address they probably belong to the same user. The Mobile

Network Operator may, in fact, assign different IP addresses to users when they access the web or when they change their position. In the elaboration of the dataset, requests sent by the same IP address within one hour are considered as being sent by the same user. It is, naturally, acknowledged that there is a possibility that two or more travellers have the same mobile phone, with the same configuration and vendor ID, using the same wireless local area network and looking for information in the same time range. Nevertheless, the probability of such an occurrence is very low; the present study's assumption that the users are identified correctly can be considered a solid one.

After having extracted the information from the raw data, the time elapsed between the time of request is computed for each group of requests made by the same user. The geographical coordinates are assigned to each bus stop to visualise them in the GIS.

Figure 3: Complete information. Some interaction blocks allow to fully understanding user mobility (i.e., boarding, alighting and bus route).

Some interaction blocks can give complete information on user mobility: the code number of stops and the bus route (Figure 3). Figure 2 displays the table with the list of requests, including the timestamp, for real time arrivals of a bus (bus number 496 to Harold Wood), at two different stops, Harold Wood Station (51285) and The Brewery (52277), and a graphical representation of the requests. These types of requests allow unambiguous interpretation on the boarding (or alighting) stops.

4. Classes of LiveBus Arrivals users

The main goal of the present work is to segment LiveBus Arrivals users into multiple classes on the basis of their application's usage patterns. User segmentation can provide better customisation for the LiveBus Arrivals application and can provide feedback on how the public transport system is used (Zeithaml et al. (1993); Brakewood et al. (2014)).

Five working days of the dataset are analysed, as described in Dell'Amico et al. (2014) where the requests are divided in two main categories: Complete and Partial information. Based on this analysis, six main profiles that can represent users' interaction with the system are identified, called: *Waiting*, *Stop Explorer*, *Route Explorer*, *Deep Explorer*, *Interchange* users and *Outliers*.

The selection of the six classes of users is based on descriptive statistics. For instance, users classified as Waiting are characterised by observations for which only one stop and no information on buses are requested. Based on preliminary analysis it can be found that at least the 15% of the observations have these characteristics. In the following paragraphs, the six classes of users are described more in detail.

4.1. Waiting users

Waiting users are those who either send a single request for a bus stop or continuously refresh the system, probably until the bus arrives. The result is a single or a set of consecutive requests for real time information on the same stop. An example of such types of requests is shown in the first three rows of Figure 2.

4.2. Stop Explorer users

The second class consists of users who inquire about the time of arrival at stops located near to each other. When a user is interested in arrivals at neighbouring stops, there could be three possible reasons: 1. the same bus is passing by the neighbouring stop and the user intends to walk there without missing the bus; 2. the user is checking for information on another stop to transfer or board another bus; 3. the user wants to use the bus only for very few stops.

Some requests relate to users requesting LiveBus Arrivals information at the opposite directions of the same stop. Therefore it may be that they cannot easily identify which is the correct route direction. Bus route directions are usually identified with the stops at the end of the line which, in some cases, may not be clear.

4.3. Route Explorer users

Sometimes users request information on a bus line to know the route direction or to have an overview of the list of stops. In this case no real time information on arrivals at stop is requested and so no data on the trip or on the approximate location of the user is available. A Route Explorer is only looking for a bus route without sending any additional request.

4.4. Deep Explorer users

The fourth class is called Deep Explorer because the user is interested not only in the time of arrival at a particular stop but also in a particular bus route. In this case, it is possible that the traveller is cross-checking if a bus line passes by a stop, and subsequently the user requires further information on the bus route.

As a side note, when passengers ask for live arrival information of a bus at a specific stop, the information is partial because there are no data on the destination but only on the stop of departure, on the bus route and its direction. Figure 4, for instance, shows that the user has asked for bus route W8 arrivals at Browning stop (47051), the route direction is Herefield Close. In this example, although the destination is unknown, there are only five stops until the end of the line.

Figure 4: Request for live arrivals of a bus route at a stop. The user has asked for bus W8 arrivals at Browning stop (47051), the route direction is Herefield Close.

4.5. Interchange users

Passengers often use the LiveBus Arrivals service to get information on transfer stops. An example of bus-to-bus transfer is shown in Figure 5 where repeated requests have been sent to get live arrivals at Worsley Bridge Road (direction to Catford) and 4 requests have been sent for the same stop but for the opposite direction (to Downham). The stop direction is determined by its code number. Worsley Bridge Road is connected to Catford Road by bus route 181, while Beckenham Hill is connected to Catford by bus 54 or by rail. The graphical representation of the requests allows understanding that the user has considered two possible routes to arrive at destination (Catford Road). The decision on which route to choose may well have been influenced by real time arrivals, such that the user may have boarded either the 181 to Catford or the 181 to Beckenham Hill and then bus 54 to Catford.

Figure 5: Bus-to-Bus transfer: alternative routes. Twelve repeated requests have been sent to get live arrivals at Worsley Bridge Road (50511) to Catford (55572) and four requests have been sent for the same stop but for the opposite direction (58558). Worsley Bridge Road is connected to Catford Road by bus route 181, while Beckenham Hill (71415) is connected to Catford by bus 54 or by rail.

4.6. Outliers

The dataset finally includes a set of requests that cannot be interpreted using the available information. For instance requests that are sent over one hour for several stops located in different London boroughs. There are other types of requests which are not possible to interpret. For instance some users have requested information on live arrivals at multiple stops located within 10 km from each other, where several bus routes serve each of the requested stops.

The intent of this type of request is probably to gather information on all bus routes transiting nearby. However these data are considered outliers because they cannot provide any precise information on user behaviour.

5. Methodology for the classification of LiveBus Arrival users

In this section, the intuition on the classes of observed behaviours is formalised in order to create a quantitative classifier. Specifically, the methodology to classify different behaviours of LiveBus Arrivals users is described. Such a classification task is challenging since ground truth classes are not available and so it is difficult to set up a standard supervised classification

approach based on training and a testing set only. To tackle this issue there are two complementary approaches:

- The application of a clustering algorithm to construct feature vectors describing user behaviour and the subsequent attachment of a label (e.g., Waiting users) to each cluster on the basis of domain knowledge.
- The identification of class labels on the basis of domain knowledge for a subset of the feature vector, followed by the training of a supervised classifier on these labelled data. Use the classifier to assign a label to the remaining feature vectors.

While the two approaches are similar from a functional viewpoint, there are some differences can be highlighted: (i) the clustering-based approach is a more consolidated approach and is the basis of unsupervised analysis, (ii) vice versa, supervised algorithms (e.g., Support Vector Machine) tend to be more refined than clustering ones, (iii) from the perspective of this work, it is simpler and more tractable to identify class labels for individual users than to entire clusters.
The domain expert in charge of assigning the labels can in fact better analyze the behaviour of individual users and establish more precise rules for classifying some of them.
In this work both approaches are applied using different clustering and classification algorithms.

5.1. The features

LiveBus Arrivals data are processed to extract a feature vector describing users' interactions with the application. For each block of observation the following features are extracted:

- The first two features are called respectively *bus* and *stop*. These variables are computed as the inverse of the number of unique buses and stops requested. The result is a set of variables that range from zero to one.
- Feature 3 is called *AVG delta time* and it is computed as the average time difference between each request in seconds. This variable provides information on the frequency a certain user sends multiple requests for LiveBus arrivals.
- Feature 4 is called *stops within 500 m* because it quantifies the number of requested stops that are located within 500 meters from each other. To compute this feature, the matrix of Euclidean distances is computed between all the stops. Finally, the numbers of stops that are located within 500 metres from each other are counted. The procedure consists of comparing stops in pairs and counting the number of adjacent stops (stop located within 500 metres). Hence, in the case of two near stops, the result will be 1; in the case of three near stops out of four, the result will be 3. This feature allows identifying both the type of requests for which the stops are located within 500 metres or far from each other. When stops are located far from each other, a possible explanation is that the users are looking for an interchange stop. In the case of nearby stops, the user is likely searching for different travel solutions or is checking for more precise information on the bus direction.

To compute these features, the dataset has been pre-processed and imported into MATLAB such that, consecutive requests made by the same user are grouped into an array containing separate tables. The result of the feature computation is a matrix where each row identifies an observation and each column a feature.

5.2. Features for Different User Profiles

On the basis of domain knowledge, the features of the different user profiles can be inferred:

- The *Waiting users* send one or multiple requests for the same stop during a short period of time. In case of multiple requests, the average time elapsed between consecutive requests is greater than zero. These types of observations are characterised by values of the feature *stop* equal to 1 and values of feature 3 *AVG delta time* generally low meaning that the time elapsed between consecutive requests is on average low.
- The *Stop Explorer users* request information on two or more stops that are located within 500 metres from each other. In this case, the value of feature *stop* is equal to or less than 0.5 and the value of the feature *stop within* is between 1 and 20.
- The *Route Explorer users* are after a particular bus route and can also repeat the request more than one time. They are only interested in the bus, so that the feature bus is 1 and all the other features, except for AVG delta time, are 0.
- The *Deep Explorer users* are after a particular stop and bus so that the features bus and stop are always 1 and AVG delta time is greater than 0.
- The *Interchange users* are looking for transfer stops. In this case, if the fraction of requested stops is equal to 0.5, the user can request one or more bus lines and then the number of near stops is 0.
- The *Outliers* identify groups of requests that cannot explain any user behaviour. For instance the user requests an inexplicable large number of different stops and most of them are located near each other.

Table 1: Training Set example

Therefore, the classification mechanism is instrumented on the following basis:

- In the case of the clustering-based approach, user-profiles labels are assigned to the extracted clusters on the basis of the above rules.
- In the case of the supervised-classification approach, the above rules are used to identify class labels for a subset of the feature vectors. A supervised classifier is trained on these labelled data and is used to assign a label to the remaining feature vectors. Table 1 shows an example of *training set* made by hand on the basis of the previous analysis.

6. Results

Experiments are conducted with both the clustering and the supervised classifier approach. Since no labels are assigned to the observations of the entire dataset, the classification is evaluated with a 3-fold cross-validation approach on the training set (the only labelled data available). The cross validation method consists of the separation of the entire dataset in three equal groups of elements with 1/3 of the dataset acting as the validation set while the remaining part is the training set. The accuracy of the classification is defined as the fraction of the number of correctly predicted data over the number of observations of the testing data.

Experiments are conducted with the following settings:

1. For the clustering approach: K-Means and EM algorithms Duda et al. (2000) are tested, setting parameters to obtain six clusters (as there are six user profiles to be identified). Then, the minimum error mapping of classes to clusters is found by recursively considering all possible class to cluster assignments. Computations are conducted with the Weka library (*www.cs.waikato.ac.nz/ml/weka/*).
2. For the supervised classification approach, the following are tested: Logistic regression, OneR (i.e., one-level decision tree), Decision tree (C4.5), Random Forest (Duda et al. (2000)), and SVM-SMO (Vapnik (1995); Platt(1998)). In order to decide for the most appropriated kernel function and SVM regularization C and kernel γ parameters, we run a simple grid search optimization ending up in using: sigmoid/poly kernel, C=2048 and γ=16. Computations have been conducted with the Weka library.

Table 2: Classification Results using 3-fold cross-validation and multiple approaches

Table 2 illustrates results in terms of precision, recall and F-score for the different algorithms being tested. As can be seen, the classification-based approaches tend to produce better results than the clustering-based ones and tree-based approaches and SVM-SMO generate excellent results. It is however important to emphasise that all these results are obtained in the absence of a solid ground truth, so classification performance might be biased by our domain-knowledge-based class assignments. As SVM-SMO produces the best results (even though by a small margin), the following analysis is based on the results of this classification.

6.1. Discussion of the results

A first analysis of the results can be used to assess how the LiveBus Arrivals application is used. 30% of the requests have been sent by *Waiting users*, the 32% by *Explorers* while 37% of the requests can be associated to *Interchange users*. Furthermore the dataset includes a very small percentage of *Outliers* (0.02%) as shown in Table 3.

Table 3: Classification Results - Weekdays

The distribution of requests among the different classes of users suggests that the service is divided in almost exactly three categories. Therefore, an improved service could include the possibility to retrieve RTI on bus-to-bus transfer. A high percentage of users are classified as *Waiting users* as such an improved LiveBus Arrivals application could foresee the possibility to have automatically updated information.

In addition, the results are not limited to the LiveBus Arrivals application itself, but could give insights on how the public transport service is used. Specifically, the temporal pattern of users accessing the application can reveal the pattern with which users ride public transportation (see Section 6.2). The spatial distribution of requests reveals where people use public transportation (see Section 6.3). Finally, all these data can be compared with census statistics about the mobility in the city (see Section 6.4). The main limitations of the study deals with the different perspectives of the analysis of mobility such as:

- The randomness of human mobility does not allow determining with certainty users behaviour based on their requests for information on real time arrivals.
- The sample of data is limited to the users who deploy the LiveBus Arrivals system so that it is not possible to include other classes of users.
- It is difficult to perform a standard supervised classification because true classes are not available.

6.2. Classes on time

Figure 6 shows the distribution of requests for each of the five identified profiles of the LiveBus Arrivals users over time. Overall it is possible to identify two main peak hours (6-8 AM and 5-7 PM) and different behaviour during the weekend when the number of requests decreases significantly for all types of users (Figure 7).

Figure 6: Number of requests over time for each class of user (Monday - Friday)

Figure 7: Number of requests over time for each class of user (Saturday - Sunday)

Overall it can be seen that the temporal behaviour of requests conforms to common-sense expectations about the five user classes, thus partially validating the analysis. Temporal analysis gives also data-driven information on how public transport is used that could be helpful in better managing and organising transport schedules. For instance, the analysis of the number of requests for real time arrivals from a specific bus stop could provide an indication of its congestion during the day. Canada Water Station, for instance, is the bus stop with the highest

number of requests. Figure 6 shows that the majority of the requests are sent during the afternoon thus indicating that the stop is highly congested during afternoon peak hours (between 5 and 6 PM). The behaviour of requests for RTI at Canada Water Station may indicate that during the afternoon the congestion is higher and travellers are waiting longer compared to the morning. Furthermore, this may suggest performing a deeper analysis of the real situation to evaluate the need of an increased frequency of the services.

6.3. Spatial distribution of requests

The number of requests for each bus stop made by the LiveBus Arrivals users can be visualised using the GIS. The classification of different types of users and the possibility to aggregate the requests by borough provide an aggregated overview of the requests' origin. However these results cannot provide information on the destination of the users which is fundamental to be able to supply the service based on the real demand for transport.

Figure 8: Number of requests by hour of the day - Canada Water Station

Figure 9 shows a representation of the number of requests made by all users during the entire period under observation. Darker colours indicate high numbers of requests. The map shows that there are several areas in which the LiveBus Arrivals service is used more intensively compared to others and this result could be the basis to start a deeper analysis of the public transport usage by district areas. Specifically, public transport managers could decide to increase transport offers in darker regions, while attempting to understand why the system is less used in lighter regions.

It is also possible to compare the outflows of commuters who go to work by bus provided by the census (see Figure 10). The intensity of RTI requests and the number of commuters who go to work by bus shows similar behaviour in terms of areas of origins.

Figure 9: Number of requests for each borough between 6 and 8 AM

Figure 10: Number of outflows of workers who travel by bus from each borough

6.4. Estimation of commuters using buses

In this section the results of the classification are used in an attempt to validate the City of London Travel to Work census (2011). This survey provides data on the number of City workers who travel by bus by place of residence, where inflows and outflows are available for each borough of the City of London.

The survey consists of 22,000 interviews posed to commuters who travel by bus.

The analysis is focused on the main peak hours during which the higher number of requests is sent by the users. The aggregation of the different classes of LiveBus Arrivals requests by London borough and by hourly time ranges is performed by running a spatial query on the set of labelled observations over each borough. The query is run for all classes, excluding the Route Explorers and the Outliers.

To evaluate the ability to use LiveBus Arrivals to estimate the number of commuters that travel by bus in the city of London, the aggregated number of requests is weighted with the number of residents of each borough. The weight is computed as the ratio of the residents of each borough

with the total population of the City of London. All boroughs are considered in the analysis so that the total number of residents corresponds to the overall population of London.

More specifically, the correlation between the weighted number of LiveBus Arrivals requests and the outflows of workers commuting by bus from each borough is computed. The highest value of the correlation (84%) is found for the *Waiting users* in the time range included between 5 and 7 PM and the number of outflows as shown in Figure 11.

Figure 11: Waiting users: correlation between outflows and weighted number of requests. The strong positive correlation highlights the relation between the outflows of commuters who travel by bus from each borough of the city of London and the number of requests sent from each borough.

The correlation for all classes of users is showed in Table 4 where the lowest correlation is 78% for the *Deep Explorer* users between 5 and 7 PM. Strong positive correlation partially validates the approach which consists on the possibility of enriching (or even substituting) census surveys with this kind of automated analysis.

Table 4: Correlation between weighted number of requests and travels by bus

7. Conclusion

This work presented a new application of the information on the requests on real time arrivals made by the users of the bus service in London. The data have been analysed by taking few

samples and different classes of users' behaviour have been described and represented using GIS. A set of features that characterise six different classes of users have been computed to run experiments with both the clustering and the supervised classification approach. It was found that the classification-based approaches tend to produce better results than the clustering-based one and that tree-based approaches and SVM-SMO generate excellent results. Following the results of the experiments, the analysis was based on the classification obtained with the SVM-SMO.

After having classified the LiveBus arrivals users, the last part of this work is focused on the evaluation of the possibility to use these requests as a proxy for the estimation of the number of outflows of workers commuters who travel by bus. Peak hours have been identified by plotting the number of requests for each hourly time range and by grouping them by day of the week. The peak hours reflect the most probable behaviour of bus users during different times of the day. For instance, in the morning users know the number of stops they intend to board so that they send repeated requests for real time arrivals at the same bus stop or interchange with another bus. Finally, the correlation between the number of requests for live time arrivals in each borough of the city of London and the number of outflows of workers commuting by bus from each borough have shown, in all cases, a high correlation. Therefore it is possible to conclude that requests for real time arrivals may be taken into account to update the census data on commuters by bus.

The use of existing data to describe the behaviour of the bus transport users can be considered as a company-driven approach. This process intends to determine how users take advantage of the LiveBus Arrivals system thus supporting the possibility to add additional features to the system. For instance, it was found that 37% of the users can be classified as Interchange users meaning that adding a functionality that provides RTI on bus-to-bus transfer may be welcome by the

travellers. A more specific analysis at borough level can provide additional information on the behaviour of travellers. By combining the number of requests for each bus stop to additional information such as bus reliability, it may be possible to obtain a new and robust indicator of waiting times. This reliability indicator may help to identify where exactly the bus route needs to be improved by, for instance, increasing the service during specific time ranges.

Acknowledgement

The authors would like to acknowledge networking support by the COST Action TU1004 and Transport for London Buses for providing the data.

Annex

K-Means and EM Clustering are popular unsupervised learning approaches.

K-Means is an iterative refinement technique. Given an initial set of *k* random centroids (i.e., points in the same domain of observations to be clustered): $c_1^{(1)},\ldots,c_k^{(1)}$ – superscript (1) indicates first iteration, the algorithm proceeds by alternating between two steps. At each iteration *t*:

1. Assign each observation x_p to the closest centroids

$$S_i^{(t)} = \{ x_p : dist\left(x_p - c_i^{(t)} \right) \le dist\left(x_p - c_j^{(t)} \right) \ \forall j \in [1,k] \}$$

2. Compute the new centroids' locations

$$c_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_p \in S_i^t} x_p$$

The algorithm has converged when the assignments no longer change.

The EM (expectation maximization) technique extends K-Means by computing probabilities of cluster memberships. The goal of the EM algorithm then is to maximize the overall probability or likelihood of the data, given the (final) clusters.

Logistic regression is a classification method that predicts the probabilities of the outcomes of a categorically distributed variable, given a set of independent variables (features). Logistic regression creates a scoring function by linearly combining features for a given observation with weights that are selected to maximize the likelihood of a set of (training) examples.

Decision Tree, OneR, Random Forest are classification methods based on the idea of generating a tree-like model that predicts the value of a class based on several input features. A *decision tree* is generated by recursive partitioning: *(i)* A feature attribute is selected to split on, *(ii)* disjoint ranges of attribute values are formed, *(iii)* a tree is returned with one edge or branch for each subset. Each branch has a descendant sub-tree (recursively) or a class value. *OneR* is a decision tree with a single attribute split. *Random Forest* combines multiple decision trees (that are trained using different data and focusing on different attributes) to improve performance and reduce over-fitting.

Support Vector Machines (SVM) is a supervised learning approach for regression and classification problems. SVM aims to construct one or a set of hyperplanes to partition observations. An example in the bi-dimensional space is reported in Figure 12. The objective is

to maximise the distance between the closest observations (support vectors) and the separating hyperplane. The larger the margin between the closest observation and the hyperplane, the better the classifier is able to generalize.

Figure 12: Hyperplanes and support vectors. Source: adapted from Burger (1998)

The decision problem consists in finding the optimal separator of observations.

The perpendicular projection of the observation onto the hyperplane is provided by the dot product between the vector **w** and the observation $\mathbf{x}_p$ as follows:

$$\mathbf{w} * \mathbf{x}_p \geq b$$

In the bi-dimensional space, the decision on which hyperplane the observation lies on, depends on which of the two following inequalities is satisfied:

$$\mathbf{w} * \mathbf{x}_p + b \geq +1 \qquad for \qquad y_p = +1$$

$$\mathbf{w} * \mathbf{x}_p + b \leq -1 \qquad for \qquad y_p = -1$$

The two constraints are then combined into one and, according to the Lagrangian formulation of the problem; the constraints are subtracted from the objective function:

$$L_P \equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_p \alpha_p \left[y_p (\mathbf{x}_p * \mathbf{w} + b) - 1 \right]$$

The problem, therefore, consists of solving a convex quadratic programming problem, where L_P has to be minimized with respect to **w** and *b*. The main innovation of the SVM relies on the use of a kernel function which simplifies the computation by avoiding the need of explicitly mapping the data to a high dimensional feature space.

References

Ahas, R., Aasa, A., Silm, S. & Tiru, M. (2010). Daily rhythms of suburban commuters movements in the tallinn metropolitan area: Case study with mobile positioning data. Transportation Research Part C: Emerging Technologies, 18(1), 45-54.

Becker, R. A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, E. & Volinsky, C. (2011a). Route classification using cellular handoff patterns. Proceedings of the 13th International Conference on Ubiquitous Computing, 123-132.

Becker, R.A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A. & Volinsky, C. (2011b). A tale of one city: Using cellular network data for urban planning. Pervasive Computing, IEEE, 10, 18-26.

Brakewood, C., Barbeau, S. & Watkins, K. (2014). An experiment evaluating the impacts of real-time transit information on bus riders in Tampa, Florida. Transportation Research Part A: Policy and Practice, 69, 409-422.

Burgers, C. J. C. (1998). A tutorial on *support vector machines* for pattern recognition. Data Mining and Knowledge Discovery, 2, 121-167.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J.J. & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies, 26, 301-313.

Cats, O. & Loutos, G., (2016). Real-Time Bus Arrival Information System: An Empirical Evaluation. Journal of Intelligent Transportation Systems, 20(2), 138-151.

Dell'Amico, M., Hadjidimitriou, S. & Kaparias, I., (2014). A descriptive study on public transport user behaviour from live bus arrivals. In 2nd Conference on Sustainable Urban Mobility, Volos (Greece).

Duda, R. O., Hart, P. E. & Stork, D. G, (2000). Pattern Classification. Wiley.

Dziekan, K. & Kottenhoff, K., (2007). Dynamic at-stop real-time information displays for public transport: effects on customers. Transportation Research Part A: Policy and Practice, 41(6), 489-501.

Fonzone, A., Schmöcker, J.-D. & Viti, F., (2016). New services, new travelers, old models? Directions to pioneer public transport models in the era of big data. Journal of Intelligent Transportation Systems, 20(4), 311-315.

Hardy, N., (2012). Provision of bus real time information to all bus stop in London. In Proceedings of the 19th Intelligent Transportation Systems World Congress.

Holleczek, T., Yu, L., Lee, J.K., Senn, O., Ratti, C. & Jaillet, P. (2014). Detecting weak public transport connections from cellphone and public transport data. In Proceedings of the 2014 International Conference on Big Data Science and Computing.

Jiang, B., Yin, J. & Zhao, S., (2009). Characterizing the human mobility pattern in a large street network. Physical Review E, 80(2).

Lai, W. & Chen, C., (2010). Behavioral intentions of public transit passengers the roles of service quality, perceived value, satisfaction and involvement. Transport Policy, 18(2), 318-325.

Liu, F., Janssens, D., Wets, G. & Cools, M., (2013). Annotating mobile phone location data with activity purposes using machine learning algorithms. Expert Systems with Application, 40(8), 3299-3311.

Liu, Y., Kang, C.G., Gao, S., Xiao, Y. & Tian, Y., (2012). Understanding intra-urban trip patterns from taxi trajectory data. Journal of Geographical Systems, 14:463-483.

Noulas, A., Scellato, S., Lathia, N. & Mascolo, C., (2012). Mining user mobility features for next place prediction in location-based services. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, 1038-1043.

Platt, J.C., (1998). Advances in Kernel Methods - Support Vector Learning, chapter Fast training of support vector machines using sequential minimal optimization. MIT Press.

Ratti, C., Williams, S., Frenchman, D. & Pulselli, R.M., (2006). Mobile landscapes: Using location data from cell phones for urban analysis. Environment and Planning B: Planning and Design, 33, 727-748.

Sevtsuk, A. & Ratti, C., (2010). Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. Journal of Urban Technology, 17(1), 4160.

Tanaka, M., Uno, N. & Shiomi, Y., (2014). Experimental Study of Effects of Travel Time Distribution Information on Dynamic Route Choice Behavior. Journal of Intelligent Transportation Systems, 18(2):215-226.

Tang, L. & Thakuriah, P., (2012). Ridership effects of real-time bus information system: A case study in the city of Chicago. Transportation Research Part C: Emerging Technologies, 22, 146-161.

Vapnik, V.N., (1995). The Nature of Statistical Learning Theory. Springer-Verlag.

Watkins, K.E., Ferris, B., Borning, A., Rutherford, G.S. & Layton, D., (2011). Where is my bus? Impact of mobile real-time information on the perceived and actual wait time of transit riders. Transportation Research Part A: Policy and Practice, 45(8), 839-848.

Yuan, Y., Raubal, M. & Liu, Y., (2012). Correlating mobile phone usage and travel behavior a case study of Harbin, china, computers. Environment and Urban Systems, 36(2), 118-130.

Zeithaml, V.A., Berry, L.L. & Parasuraman, A., (1993). The nature and determinants of customer expectations of service. Journal of the academy of Marketing Science, 21(1), 1-12.

Zhang, F., Shen, Q. & Clifton, K., (2008). Examination of traveler responses to real-time information about bus arrivals using panel data. Transportation Research Record: Journal of the Transportation Research Board, 2082.

Table 1: Training Set example

Class	bus	stop	AVG delta time	stops within 500m
Waiting	0	1	723	0
Waiting	0	1	39	0
Waiting	0	1	50	0
Waiting	0	1	359	0
Route Explorer	1	0	0	0
Route Explorer	1	0	8.6	0
Route Explorer	1	0	0	0
Route Explorer	1	0	20	0
Stop Explorer	0	0.16	739	14
Stop Explorer	0	0.14	1	17
Stop Explorer	0	0.16	537	9
Stop Explorer	0	0.14	2.25	13
Interchange	0.3	0.5	8.8	0
Interchange	0.5	0.5	78.7	0
Interchange	1	0.5	38.5	0
Interchange	1	0.5	43.5	0
Deep Explorer	1	1	36.1	0
Deep Explorer	1	1	67.4	0
Deep Explorer	1	1	61.4	0
Deep Explorer	1	1	133.3	0
Outlier	0	0.0002	3.89	6898
Outlier	1	0.0006	29.2	1104
Outlier	0	1	4.7	11486
Outlier	1	0.0008	42.7	560

Table 2: Classification Results using 3-fold cross-validation and multiple approaches

Classifier	**Precision**	**Recall**	**F-Score**
K-Means	0.765	0.745	0.747
EM	0.674	0.617	0.621
Logistic	0.94	0.933	0.931
OneR	0.735	0.817	0.761
Decision Tree	0.962	0.95	0.949
Random Forest	0.963	0.95	0.949
SVM-SMO	0.965	0.948	0.952

Table 3: Classification Results - Weekdays

Type of users	Percentage
Waiting	29.70%
Stop Explorer	3.70%
Route Explorer	16.70%
Deep Explorer	11.40%
Interchange	36.60%
Outliers	0.02%

Table 4: Correlation between weighted number of requests and travels by bus

Time range	Waiting	Deep Explorer	Stop Explorer	Interchange
6-8 AM	79.40%	80.00%	78.10%	79.40%
5-7 PM	84.60%	77.50%	81.70%	81.70%

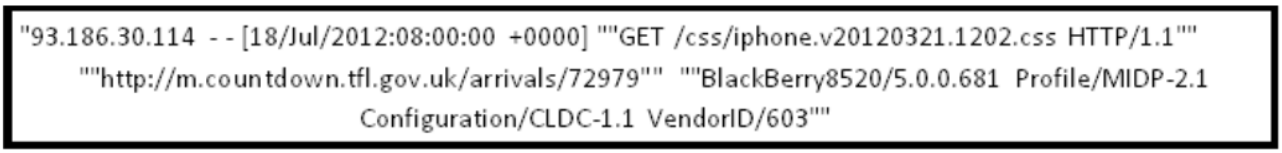
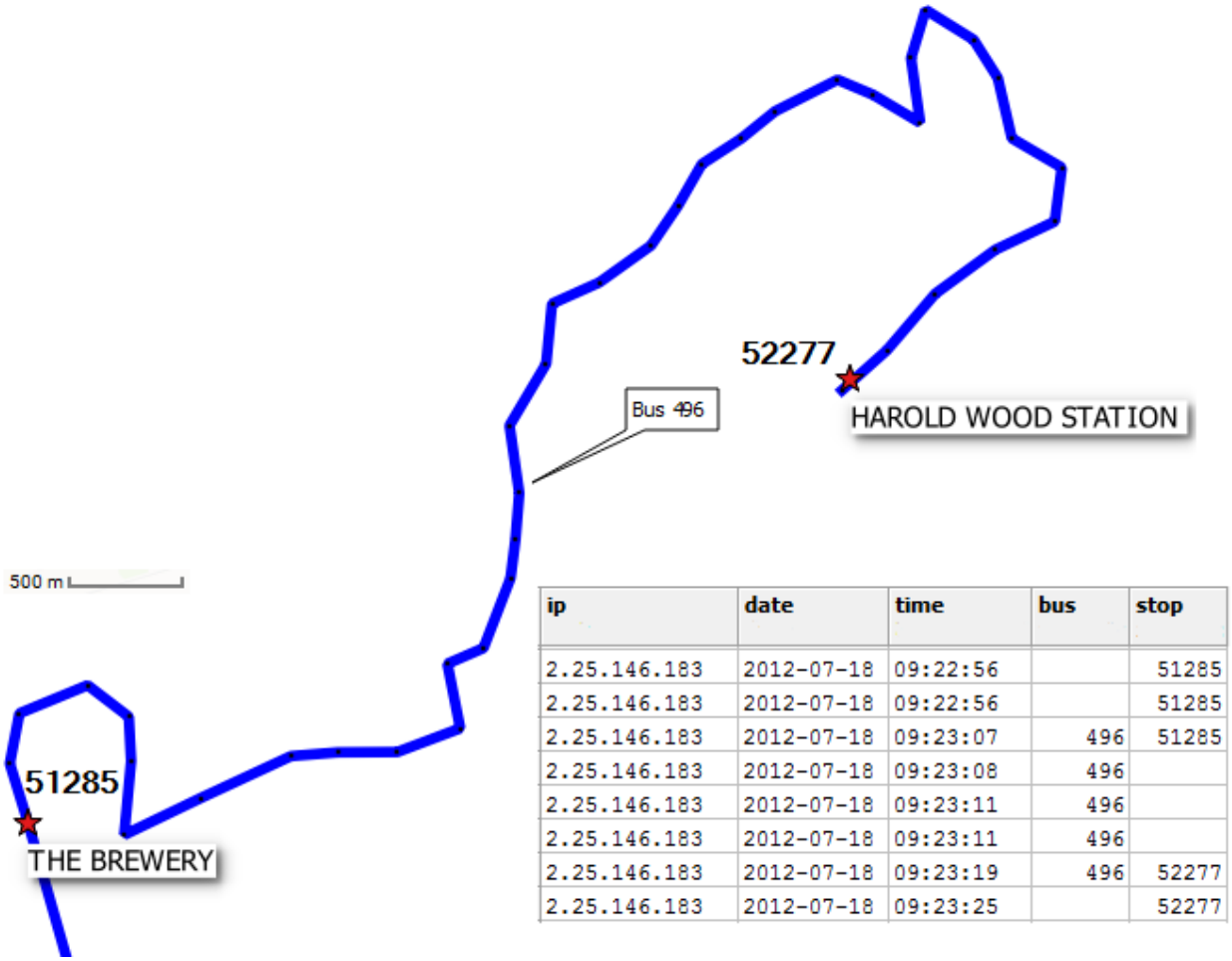


Figure 1: Raw data. The query encodes a request for real time arrivals at stop number 72979. The line shows the IP address, the timestamp, the HTTP request, the web address, the mobile phone model together with its configuration and vendor ID.

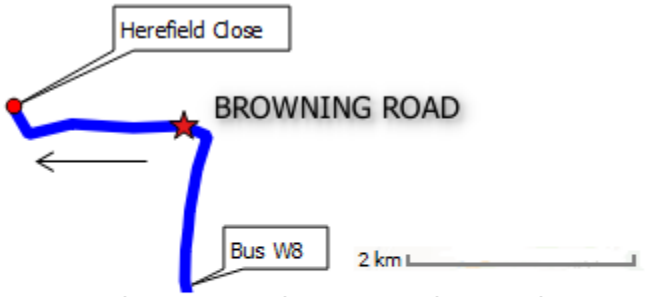
ip	date	time	bus	stop
31.107.126.168	2012-07-17	06:34:14		75034
31.107.126.168	2012-07-17	06:34:16		75034
31.107.126.168	2012-07-17	06:34:20		75034
212.183.128.37	2012-07-17	07:34:16	228	
212.183.128.37	2012-07-17	07:34:19	228	
212.183.128.37	2012-07-17	07:34:19	228	
212.183.128.37	2012-07-17	07:34:20	228	
93.186.20.24	2012-07-17	12:34:08		71221
93.186.20.24	2012-07-17	12:34:09		71221
93.186.20.24	2012-07-17	12:34:17		50180
93.186.20.24	2012-07-17	12:34:17		50180
93.186.20.24	2012-07-17	12:34:20		74980
93.186.20.24	2012-07-17	12:34:21		74980
31.113.208.39	2012-07-17	06:59:14		72252
31.113.208.39	2012-07-17	07:00:45		72252
31.113.208.39	2012-07-17	07:00:46		72252
31.113.208.39	2012-07-17	07:02:26		72252
31.113.208.39	2012-07-17	07:02:46	356	
31.113.208.39	2012-07-17	07:02:48	356	
31.113.208.39	2012-07-17	07:03:38	356	48628
31.113.208.39	2012-07-17	07:04:12		48628
31.113.208.39	2012-07-17	07:04:14		48628
31.113.208.39	2012-07-17	07:04:35		48628
31.113.208.39	2012-07-17	07:04:36		48628
31.113.208.39	2012-07-17	07:04:45		48628
31.113.208.39	2012-07-17	07:04:46		48628

Figure 2: Extracted Data. Multiple interactions of the same IP within a time limit are arranged in interaction blocks.



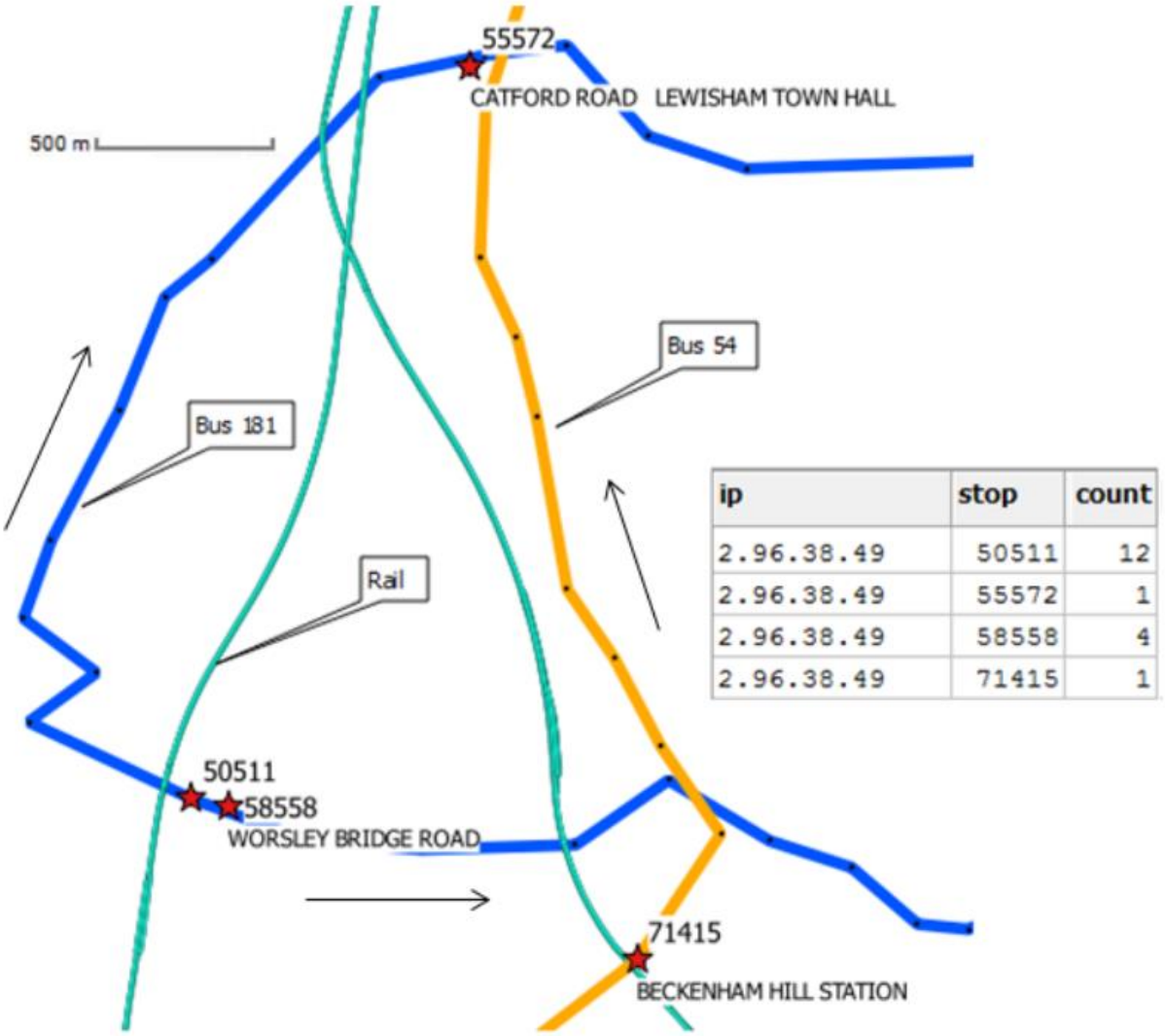
ip	date	time	bus	stop
2.25.146.183	2012-07-18	09:22:56		51285
2.25.146.183	2012-07-18	09:22:56		51285
2.25.146.183	2012-07-18	09:23:07	496	51285
2.25.146.183	2012-07-18	09:23:08	496	
2.25.146.183	2012-07-18	09:23:11	496	
2.25.146.183	2012-07-18	09:23:11	496	
2.25.146.183	2012-07-18	09:23:19	496	52277
2.25.146.183	2012-07-18	09:23:25		52277

Figure 3: Complete information. Some interaction blocks allow to fully understanding user mobility (i.e., boarding, alighting and bus route).



ip	date	time	bus	stop
2.25.67.198	2012-07-18	06:16:32	8	
2.25.67.198	2012-07-18	06:16:33	8	
2.25.67.198	2012-07-18	06:16:39	8	47051

Figure 4: Request for live arrivals of a bus route at a stop. The user has asked for bus W8 arrivals at Browning stop (47051), the route direction is Herefield Close.



ip	stop	count
2.96.38.49	50511	12
2.96.38.49	55572	1
2.96.38.49	58558	4
2.96.38.49	71415	1

Figure 5: Bus-to-Bus transfer: alternative routes. Twelve repeated requests have been sent to get live arrivals at Worsley Bridge Road (50511) to Catford (55572) and four requests have been sent for the same stop but for the opposite direction (58558). Worsley Bridge Road is connected to Catford Road by bus route 181, while Beckenham Hill (71415) is connected to Catford by bus 54 or by rail.

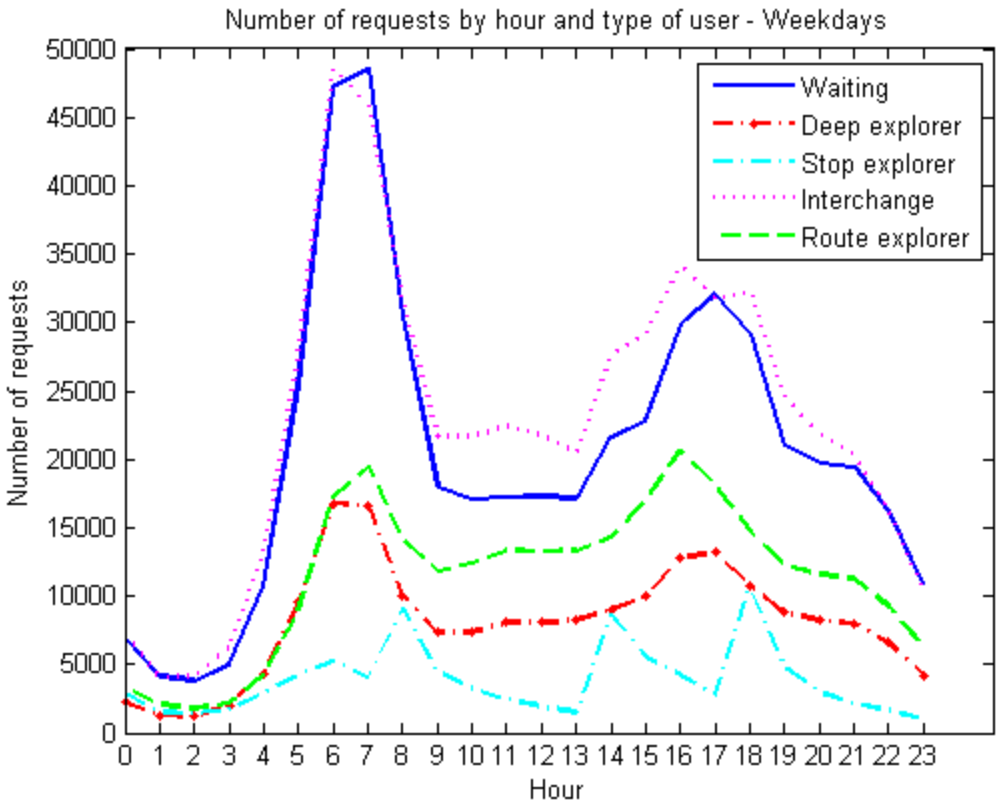


Figure 6: Number of requests over time for each class of user (Monday - Friday)

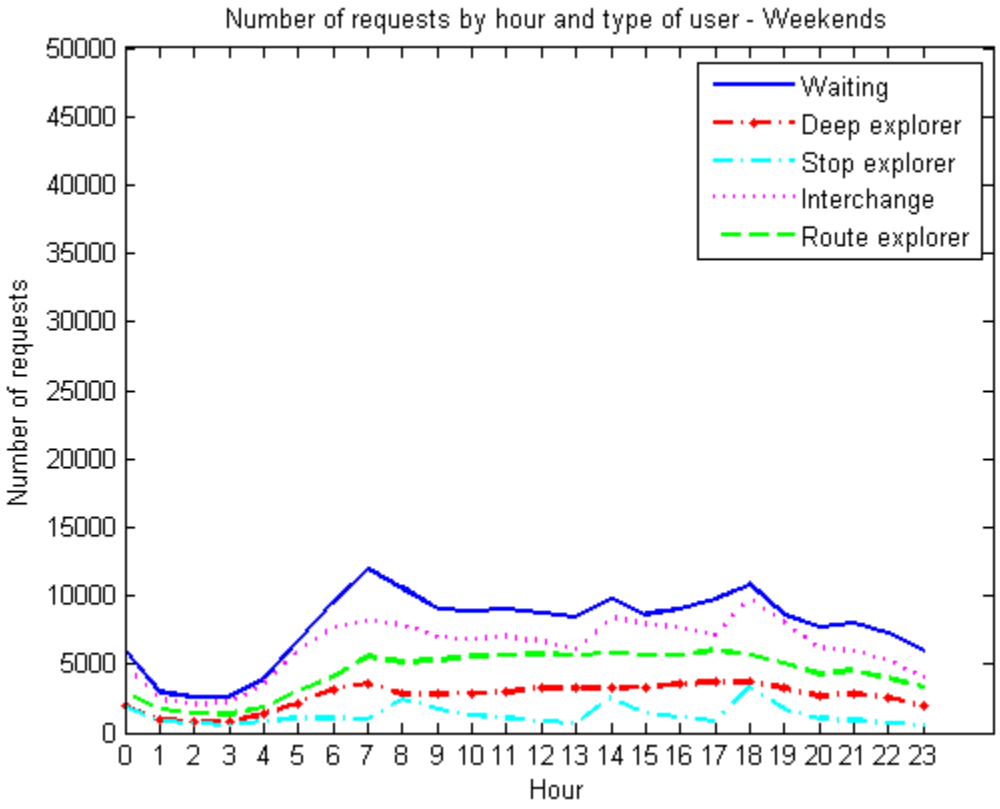


Figure 7: Number of requests over time for each class of user (Saturday - Sunday)

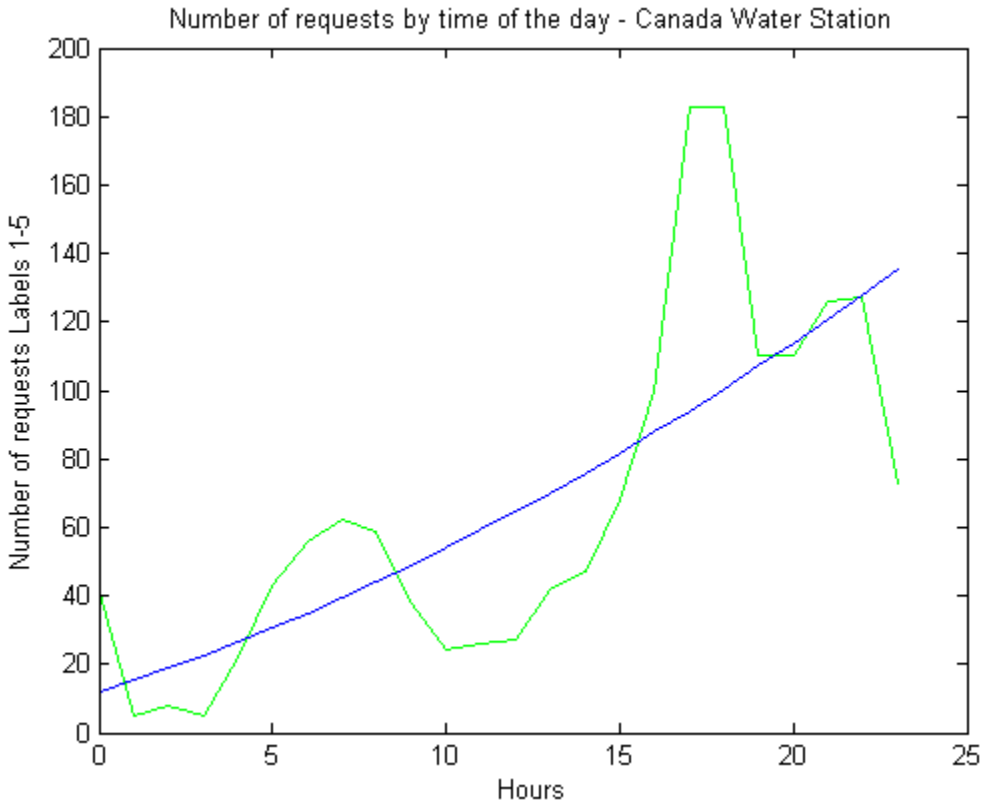


Figure 8: Number of requests by hour of the day - Canada Water Station.

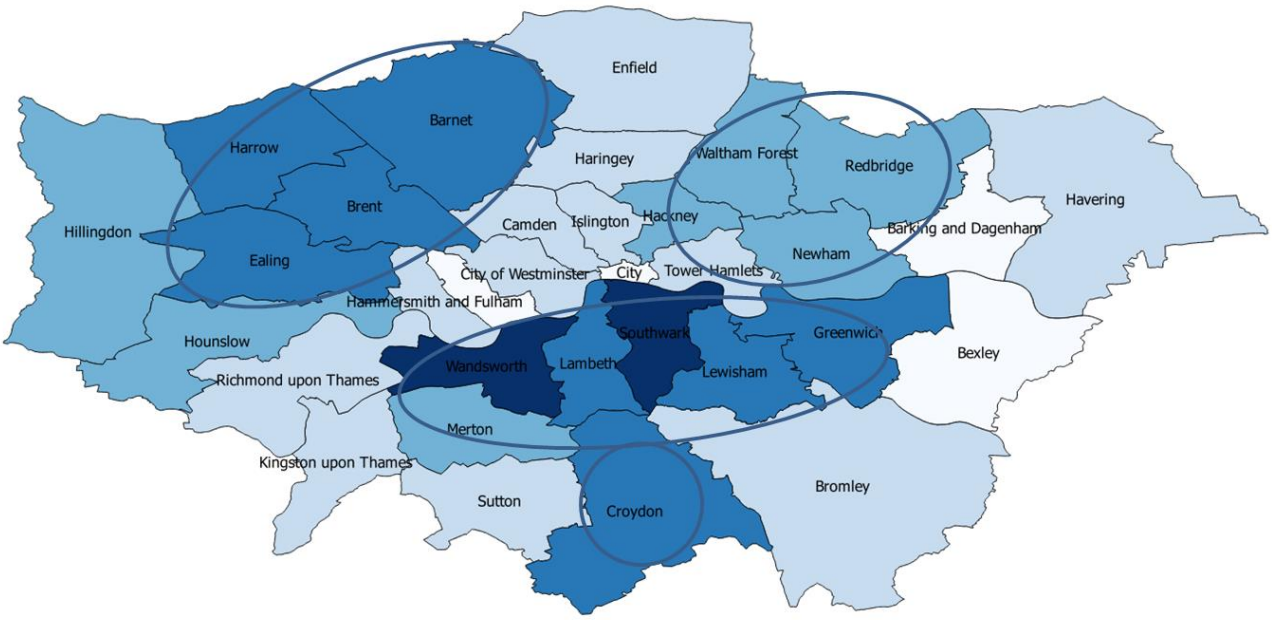


Figure 9: Number of requests for each borough between 6 and 8 AM.

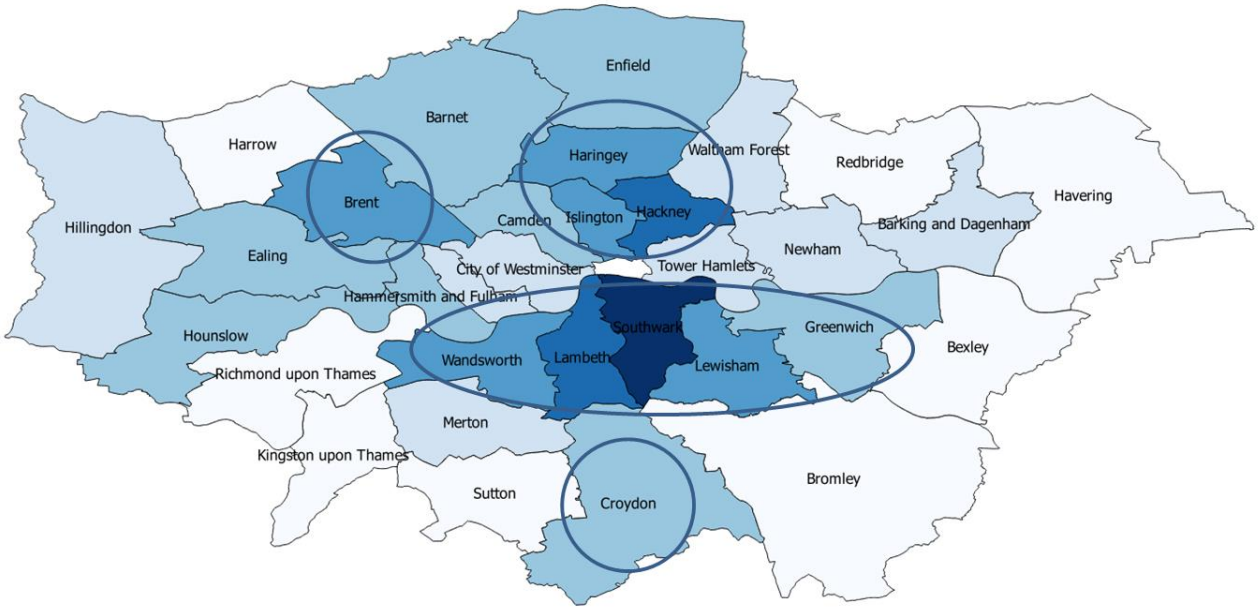


Figure 10: Number of outflows of workers who travel by bus from each borough.

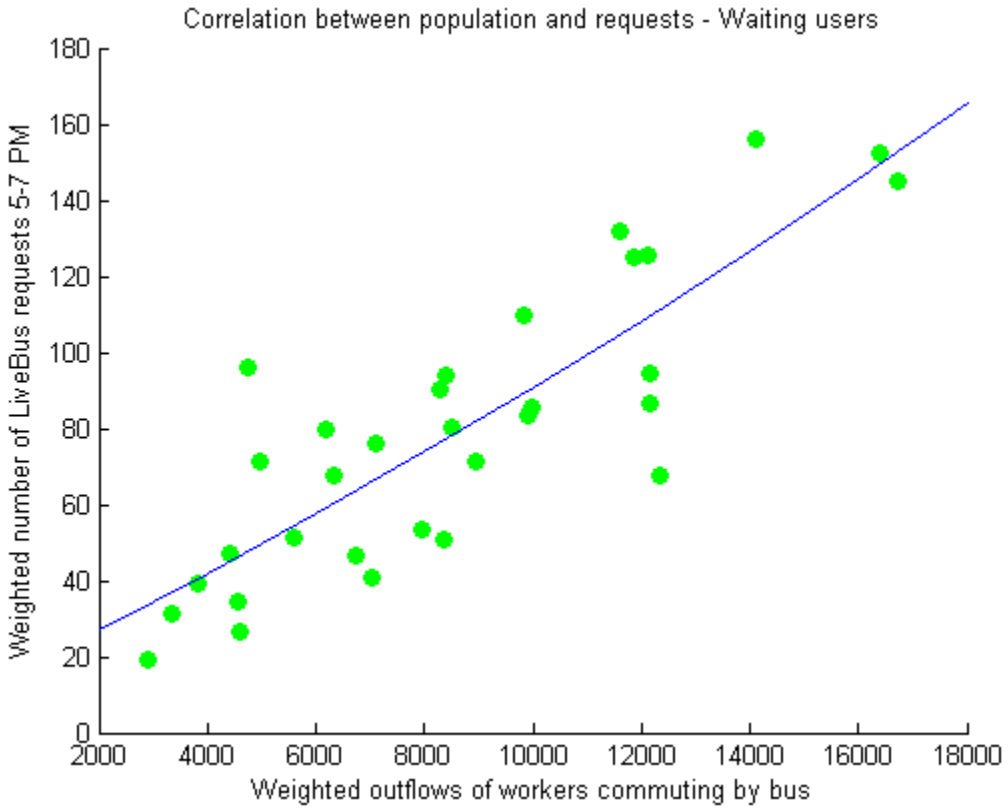


Figure 11: Waiting users: correlation between outflows and weighted number of requests. The strong positive correlation highlights the relation between the outflows of commuters who travel by bus from each borough of the city of London and the number of requests sent from each borough.

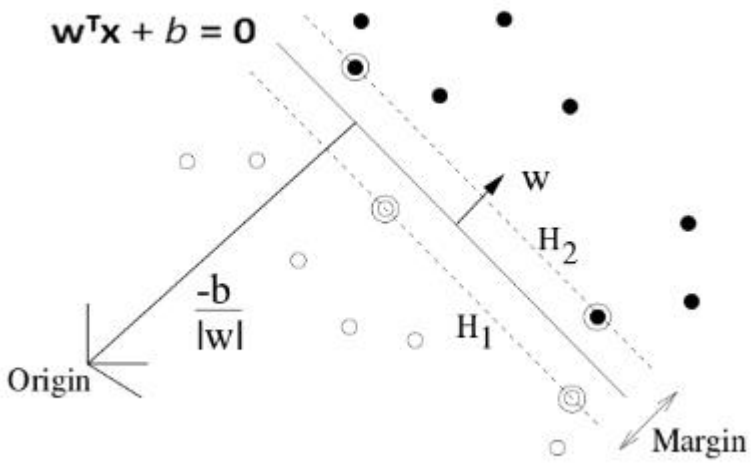


Figure 12: Hyperplanes and support vectors. Source: adapted from Burger (1998)

