Studying commuting behaviours using collaborative visual analytics

Beecham, Roger
roger.beecham.1@city.ac.uk

Wood, Jo
j.d.wood@city.ac.uk

Bowerman, Audrey
audrey.bowerman@tfl.gov.uk

Date accepted in CEUS : October 2013

Abstract

Mining a large origin-destination dataset of journeys made through London's Cycle Hire Scheme (LCHS), we develop a technique for automatically classifying commuting behaviour that involves a spatial analysis of cyclists' journeys. We identify a subset of potential commuting cyclists, and for each individual define a plausible geographic area representing their workplace. All peak-time journeys terminating within the vicinity of this derived workplace in the morning, and originating from this derived workplace in the evening, we label commutes. Three techniques for creating these workplace areas are compared using visual analytics: a weighted *mean-centres* calculation, spatial *k-means* clustering and a *kernel density-estimation* method. Evaluating these techniques at the individual cyclist level, we find that commuters' peak-time journeys are more spatially diverse than might be expected, and that for a significant portion of commuters there appears to be more than one plausible spatial workplace area. Evaluating the three techniques visually, we select the density-estimation as our preferred method. Two distinct types of commuting activity are identified: those taken by LCHS customers living outside of London, who make highly regular commuting journeys at London's major rail hubs; and more varied commuting behaviours by those living very close to a bike-share docking station. We find evidence of many interpeak journeys around London's universities apparently being taken as part of cyclists' working day. Imbalances in the number of morning commutes to, and evening commutes from, derived workplaces are also found, which might relate to local availability of bikes. Significant decisions around our workplace analysis, and particularly these broader insights into commuting behaviours, are discovered through exploring this analysis visually. The visual analysis

approach described in the paper is effective in enabling a research team with varying levels of analysis experience to participate in this research. We suggest that such an approach is of relevance to many applied research contexts.

Keywords: collaborative visual analytics; bicycle share schemes; commuting behaviour

1 Introduction

Since its introduction in July 2010 over 20 million journeys have been made through the London Cycle Hire Scheme (LCHS). Recent analyses of LCHS usage data have found daily tidal flows of bikes into and out of central London, which coincide with commuting peaks (Wood et al. 2011, Lathia et al. 2012). These flows disproportionately redistribute bikes to particular parts of the city, making many docking stations unusable - either rendered entirely full or empty of bikes. This is a problem common to most urban bike share schemes (OBIS 2011). To keep the system as balanced as possible, bikes are manually transported across the city at peak times, and in priority areas docking stations are continually replenished with bikes or bikes continually removed from docking stations. Since such load rebalancing is expensive, Transport for London (TfL), the organisation responsible for the scheme's operation, wish to better understand commuting LCHS users and their journeys.

Working with a diverse team of colleagues at TfL, three questions motivate this research:

1. What are the characteristics of people who take part in commuting based activities?
2. Where do commuting events happen?
3. Under what circumstances are journeys made during the working day?

Before these three questions can be investigated, there is a broader question:

4. How can commuting journeys and commuting LCHS cyclists be reasonably detected?

The task of identifying commuting behaviour might initially seem like a straightforward data mining exercise. For example, one means of identifying commuting journeys might be to find all instances where a LCHS cyclist completes a closed peak-time loop, where their last journey of the day happens during the evening commute and is the inverse of their first journey of the day. Recent analysis of usage data from London's underground system has found such assumptions about commuting behaviour often do not hold (Lathia et al. 2013). This might be especially true of LCHS behaviour. Usage of the scheme is perhaps more ad hoc and subject to a wider set of environmental and other variables than use of a large metro system. Moreover, whilst London underground users can generally expect access to a train at their most convenient station, competition for bikes during peak times means that LCHS cyclists may have more modest expectations: individuals may not be able to consistently collect or return bikes at their preferred station, and therefore LCHS commuting may not consist only of journeys between a single pair

of stations.

In this paper we investigate approaches to identifying commuting journeys that involve spatial analysis of individual cyclists' peak-time journeys. Our general approach is to find a broad spatial area, or set of spatial areas, representing each commuting LCHS user's workplace, and identify all journeys that end (in the morning) or start (in the evening) from this workplace area. This is our first contribution:

- Contribution 1. A new technique for deriving customers' workplace areas and labelling commuting journeys, based on a spatial analysis of travel behaviours.

We believe this technique is novel to the extent that, unlike similar studies by Lathia et al. (2013) and Agard et al. (2006) that identify individuals whose dominant temporal usage coincides with peak-times, it relies on a spatial as well as temporal evaluation of travel behaviours. Our technique might reasonably be applied to other large-scale bike share schemes.

The second contribution relates to our approach. We use visual analytics to evaluate various workplace identification techniques. This visual approach allows relatively abstract data transformations to be made intelligible to both data analysis specialists (ourselves) and domain experts (colleagues at TfL). The paper describes a process of *chauffeuring*, whereby colleagues at TfL articulate a research problem, we propose a set of solutions and, using tailored visual analytics, we collectively explore and evaluate this solution space. We believe that such an approach is particularly suited to research contexts where decisions are required from a diverse team containing analytic, policy-related and operational specialisms. Such a requirement may be common to many applied analysis contexts.

- Contribution 2. A visual analytics approach that facilitates a data-driven discourse between a diverse set of individuals.

Finally, we describe several new insights into commuting LCHS behaviour that were generated through this analysis, and that were not previously known to colleagues at TfL. These may be relevant to others with an interest in studying usage of bike-share schemes.

- Contribution 3. A set of empirically-generated findings relevant to those interested in studying bike-share cycling behaviour.

2 Related work

2.1 Inferring commuting behaviour from origin-destination datasets

Automatically collected usage datasets from shared transport systems have only recently become available to researchers, and the literature on approaches to inferring commuting behaviour from such data is not widespread. One relevant study was conducted by Lathia et al. (2013). With a month's usage data from the London underground, the authors identify the nature and extent of commuting behaviour using various data mining algorithms. First, they make reasonable assumptions about commuting travel behaviours: that commuters will make on average two journeys or more per day; that they will typically repeat the same origin-destination (OD) pair; and that commuters will have a closed loop whereby the first origin and last destination of the day should be the same. Lathia et al. (2013) subsequently find that many travellers do not fit these expected patterns. Sixty-six percent of users take less than one trip every two days and only 8% meet the expected two trips per weekday criteria. Half of all trips taken by users in the month-long study period are entirely unique OD pairs for those people. However, 50% of all users form a closed loop on Monday-Thursdays, 44% on Fridays and 37% on weekends (Lathia et al. 2013). The authors later propose clustering algorithms for automatically finding groups of travellers with similar temporal travel profiles who typically travel at particular times of day and days of the week. This approach is also taken by Agard et al. (2006) when analysing bus usage data in Quebec. The authors find a large group of bus passengers whose usage almost exclusively coincides with peak commuting times.

Since we are interested in studying whether or not LCHS members make journeys within their working day, an important aspect of this study is to identify with a degree of certainty all journey events that we think might be commutes. Whilst the unsupervised clustering algorithms proposed by Agard et al. (2006) and (Lathia et al. 2013) would enable those who apparently use the scheme almost exclusively for commuting to be distinguished from those with more varied usage characteristics, it would not enable a total, journey-level view of commuting. It is reasonable to assume that those who commute may also often use the LCHS for non-commuting, leisure-oriented or utilitarian weekend journeys. If commuting users were only defined as people whose dominant travel patterns coincide with commuting times, then we potentially miss the commuting behaviour of individuals who typically use the scheme for other purposes. In addition, that usage of the LCHS is likely to be more ad hoc than, for instance, the London underground, assumptions around commuting

activity made by Lathia et al. (2013) might be even more problematic when applied to the LCHS dataset.

An alternative approach, taken in this study, is a spatial analysis of LCHS users' peak-time journeys. We attempt to identify a broad spatial area representing each cyclist's workplace and label all peak-time journeys arriving at this workplace area in the morning and departing from this workplace area in the evening as commutes. Clearly we have no *a priori* knowledge of such workplace areas and instead we derive them from exploring spatial patterns of LCHS cyclists' peak-time journeys.

2.2 Collaborative visual analytics and *chauffeuring*

For Tukey & Wilk (1966), the aim of any data analysis is to find insights that can be easily stated and are intelligible to the individuals conducting an analysis: 'at all stages of data analysis the nature and detail of output [...] need[s] to be matched to the capabilities of the people who use it and want it' (Tukey & Wilk 1966, 697). Visual analytics refers to the application of tools and techniques for synthesising information and discovering insights from generally large and complex datasets (Thomas & Cook 2006). It can play an important role in making research outputs interpretable to individuals with a range of specialisms. For instance, Robinson (2008) notes that the use of interactive visual interfaces enables teams of analysts with different skill sets to collectively engage in difficult analysis problems. Often teams consist of a data analysis expert, who has expertise in the visual analytics tool, and a domain specialist, who has expertise in the specific themes being studied (Arias-Hernandez et al. 2011). Nunamaker et al. (1991) use the term chauffeuring to describe such a research setting. For them, chauffeuring refers to a practice in computer-supported collaborative work whereby an individual highly familiar with a computer system acts as a mediator between those who need access to that system, but do not have the technical skills to make full use of it (Nunamaker et al. 1991). In information visualization, researchers have used a slightly expanded concept of chauffeuring, whereby it is not simply the role of a so-called technician to 'drive' a visual analytics system; the technician is also an analyst engaged in the research problem (Slingsby et al. 2011).

We also use an expanded concept of chauffeuring to guide our approach: domain experts (colleagues at TfL) define a research problem; we as analysis specialists offer a set of possible solutions; and these solutions are evaluated collectively using custom-built visual analytics software. Visually representing this analysis enables data transformations and insights to be recognisable to those conducting the analysis: ourselves and colleagues at TfL. In addition, through iteratively exploring this spatial analysis visually, appropriate techniques, analysis parameters and more

specific hypotheses that might be used in confirmatory data analysis are accepted, dismissed and refined.

3 Approach

Work on this project was divided into three stages, which map closely to our understanding of chauffeuring: i. thematic problem specification, ii. development of analytics techniques and iii. hypothesis generation and insight development. We outline the tasks involved at each of these stages here, before detailing the techniques developed in answering our research problem (section 4) and subsequently the study's research findings (section 5).

3.1 Thematic problem specification

In the first instance, we met a team of five analysts and policy makers at TfL responsible for the LCHS. The various job roles of these five individuals are detailed below:

- Head of Operations
- Operations Manager
- Operations Development Manager
- Delivery Manager
- Policy Analyst

In this initial meeting, a set of analysis themes were discussed and prioritised, alongside a list of datasets that could be shared. From this, two comprehensive LCHS usage datasets (described in 4.1) were made available. A separate meeting was then scheduled with an analyst at TfL responsible for maintaining these data, and the dataset's usefulness further evaluated in the context of emerging research priorities.

Exploratory data analysis (Tukey 1977) was carried out, and at a second meeting with our five TfL colleagues a number of research hypotheses were collectively generated. Analysing the spatial and temporal structure of journeys being made, there appeared to be a very strong commuter function, also identified in separate studies of LCHS usage (Wood et al. 2011, Lathia et al. 2012). Journeys at peak-times seemed to be dominated by particular types of people: male users and those apparently living outside of London. They were also associated with particular spatial areas: major rail terminals, London's commercial centre, the City of London, and

central London. There were other patterns of peak-time travel activity. Journeys were found in parts of London typically associated with 'leisure' activities and by individuals who did not fit the dominant peak-time demographic. Discussing these findings with the five representatives at TfL, it was noted that, particularly with a bike-share scheme, we should be cautious of conflating all peak-time usage with commuting usage. Journeys will be taken during peak-times for reasons other than commuting. We suggested a more concrete research aim - of formally identifying commuting journey events. A further requirement, suggested by TfL's Operations Development Manager, was to understand the extent and circumstances under which users make bike share journeys as part of their working day, after having commuted into work in the morning, or before commuting home from work in the evening.

Out of this early exploratory discussion, then, three substantive research questions were specified:

1. What are the characteristics of people who take part in commuting based activities?
2. Where do commuting events happen?
3. Under what circumstances are journeys made during the working day?

For these questions to be considered properly, we added a broader analytical question, which is a substantial focus of this study:

4. How can commuting journeys and commuting LCHS cyclists be reasonably detected?

3.2 Development of analytics techniques

Outside of this meeting, we researched existing approaches, already discussed in section 2.1, that attempt to infer commuting activity from behavioural origin-destination datasets. Here, the challenge of deriving cyclists' workplaces from analysing spatial patterns of peak-time travel was articulated. Researching commonly used spatial data mining techniques (O'Sullivan & Unwin 2002, Shekhar et al. 2011), an initial analysis algorithm was proposed, explored and evaluated by displaying data transformations within a tailored visual analytics application (described in section 4.3.2). Problems associated with this initial technique were subsequently diagnosed and a more sophisticated algorithm proposed. Again this technique, and its underlying parameters, were explored and evaluated by visually displaying data transformations, interacting with and varying parameters. Weaknesses and problems were diagnosed and an alternative approach proposed and explored in the same way. We argue that this iterative inspection of spatial analysis

techniques and of spatial behaviours is highly effective where travel behaviours are being explored. This analysis process is documented in sections 4.2 and 4.3.

3.3 Hypotheses generation and insight development

Once a preferred technique was identified, the analysis was presented to our five colleagues at TfL at a third analysis meeting. Decisions around accepting or rejecting algorithms, and the parameters used in algorithms, were explained using the visual analytics software created in the analysis phase. As a result of these discussions, hypotheses about observed spatial patterns at the individual (cyclist) level were related to collective, local knowledge about the scheme itself and the geography of the city. This critical reflection was enhanced by the fact that the group of colleagues at TfL were responsible for different aspects of the LCHS. A further set of visual analysis software (described in section 5.4) was presented for exploring these spatial patterns more broadly at the global, scheme-wide level. Finally, outside of these meetings, early hypotheses around commuting behaviours were explored visually and tested quantitatively.

4 Analysis

In this section we explain how our preferred analysis algorithm for identifying cyclists' workplaces, and subsequently commuting journeys, was created. After describing the LCHS datasets in detail, three techniques for deriving individual commuters' workplaces are discussed. This is an iterative analysis and we set out in each sub-section how a visual analytics approach enabled discrepancies between a perceived view of an analysis technique and its practical application to be identified.

4.1 LCHS datasets

Two separate data sources were made available for this research: a full database of customers registered with the LCHS and a complete set of journey records. The customer database holds a full postcode, date of registration and gender for every customer subscribing to the scheme. We do not have access to the same information for pay-as-you-go users, who make around 35% of all LCHS journeys. In the journeys dataset, a start and end time and origin-destination (OD) pair for every journey is recorded. The two datasets can be related with a unique customer identifier (Figure 1). We have access to usage records dating from the scheme's launch on 30th July 2010 through to 14th September 2012. However, we only analyse members making

journeys within the most recent 12 months of data available to us – between 14th September 2011 and 14th September 2012. In total this amounts to 82,874 LCHS members linked to 5,048,000 journeys. The median number of journeys made by these customers over the 12-month study period is 20, with a quarter of customers making 73 or more journeys. The LCHS is designed such that short journeys are incentivised; the first 30 minutes of travel is free. Only 3% of journeys made in the 12-month study period lasted longer than 30 minutes and the average straight-line distance travelled for all journeys is 2.0km.

We augment the customer database by both leveraging external datasets and creating new derived variables from mining customers' journeys. Firstly, the full postcode in the customer database refers to the address that customers' payment cards are registered to when signing up with the scheme. This is also the address to which keys for accessing bikes are mailed. We assume that the postcode represents customers' homes and match it to two freely available geodemographic classifiers: the 2001 Census Output Area Classification (Vickers & Rees 2006) and the Indices of Multiple Deprivation (IMD) (Department for Communities and Local Government 2011). As well as their gender, we therefore have some indication of the types of communities cyclists apparently live in. We then link the postcode variable to National Grid Coordinates and calculate straight-line distances from customers' postcode unit centroid to their nearest docking station, and add this 'distance to docking station' measure as a variable in the customer database.

In order to differentiate customers by their usage behaviours, we use Recency-Frequency (RF) segmentation (Novo 2004). RF segmentation is based on the principle that Recency – how recently a customer bought or used a product – is a good predictor of how likely that person is to buy or use that product again. The same applies to Frequency; how frequently a person buys or uses a product is a good predictor of future purchase. To conduct a RF segmentation on the LCHS usage data, we first order customers according to the last journey that they made. We then break the customer dataset into five equal-sized bins and give customers a score for Recency from 1 to 5 according to their position within those bins. Frequency scores are arrived at in a similar way, but instead of considering simply the absolute number of journeys a customer makes, we divide this figure by the time that elapsed between customers' first and last journey. Once customers are ordered according to Frequency, the dataset is again broken into five equal-sized bins and Frequency scores are assigned on a 5-point scale. Customers with a Recency score of 5 and Frequency score of 5 are the most heavy scheme users, cycling both often and recently. Those with a RF score of 1-1 perhaps used the scheme occasionally when first registering, but have not used it since. We make explicit reference to these derived variables in section 5, when explaining the different behaviours our

commuter classification elicits.

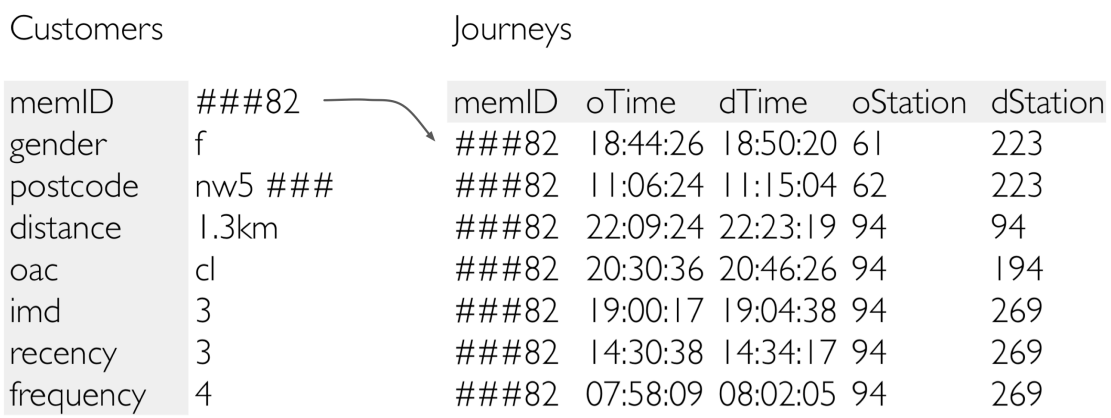

Figure 1: LCHS journeys and customers datasets. To preserve customers' anonymity, dummy values are shown.

4.2 Identifying potential commuting customers

As discussed, to find only commuting journeys, we wish to identify those journeys that regularly end within the vicinity of a customer's workplace in the morning, and regularly start from within the vicinity of a customer's workplace in the evening. We name such locations 'workplaces', but recognise that there are other attractors for repeated journeys. With no knowledge of cyclists' actual workplaces, the challenge is to arrive at an assumed workplace for each commuting member, which can be derived from members' usage records. Our first task here is to identify a subset of potential commuting cyclists.

4.2.1 A frequency based identification of potential commuting cyclists

We filter all journeys taken within the morning (6am-10am) and evening (4pm 8pm) peaks and for each customer identify the total number of these peak-time journeys. We recognise that some cyclists, particularly shift and casual workers, will also make commuting journeys outside of peak times. However, by restricting to repeated journeys within the working day, we perhaps have greater certainty of identifying genuine work-related activity. In addition, peak-time travel causes the most significant problems in terms of load rebalancing, and an important aspect of our analysis will focus on journeys that take place within these time periods.

Where a customer makes only a small number of peak-time journeys, and therefore

where we have very few points of reference, there would be insufficient data to classify with a level of certainty that member as a commuter, and their peak-time journeys as commutes. We therefore need to suggest a suitable number of repeated peak-time journeys made by cyclists, beyond which we believe observed behaviours are indicative of commuting. Ordering customers according to the number of peak-time journeys they make, and displaying this information as a frequency distribution (Figure 2), there are no obvious breaks suggesting a change in behaviours and a possible commuting function.

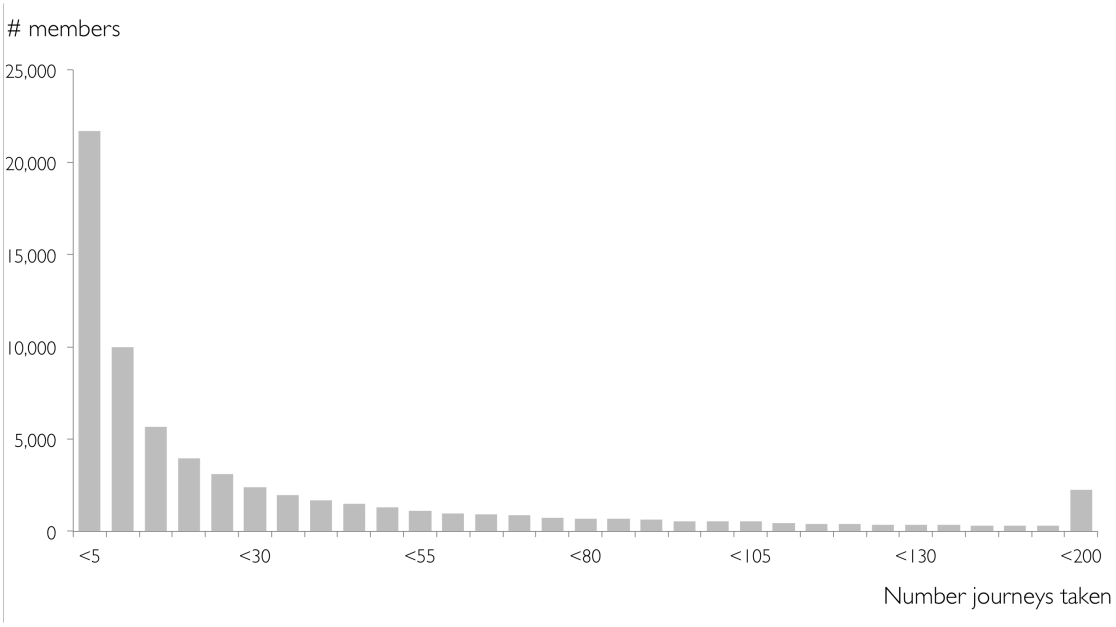

Figure 2: Frequency distribution of peak-time journeys taken by LCHS customers.

4.2.2 Visual analysis of repeated peak-time journeys

Exploring the spatial and temporal context of individuals' journeys visually, perhaps better enables judgements about the nature of cycling behaviour. Previously, we designed visual analysis software to support such exploratory analysis of the LCHS data (Beecham & Wood 2013). We created three coordinated and linked views: a timeline view showing hourly journeys by day of week; a flow map view, which gives a spatial overview of cyclists' journeys; and a customer view, allowing LCHS customers to be filtered by varying geo-demographic and behavioural characteristics. We use this same software here to evaluate the travel behaviours of members classified within each frequency bin appearing in Figure 2; we compare the space-time structure of journeys made by those who make few peak-time journeys, with members that make

many peak-time journeys. We find that after approximately 20 repeated peak-time trips, customers' usage characteristics begin to increasingly suggest a commuter function. This is expressed in Figure 3, which shows the temporal (line chart) and spatial (flow map) pattern of all journeys taken by the selected subset of members (those making between 21 and 30 journeys at peak-times). Inspecting the map view, it appears that reciprocal journeys between rail terminals and employment centres (labelled in Figure 3) become increasingly visible for those making between 21-30 peak-time journeys. By contrast, selecting individuals making less than 20 repeated peak-time journeys, the observed temporal and spatial pattern suggests a leisure function, with largely weekend journeys within London's parks more prominent.

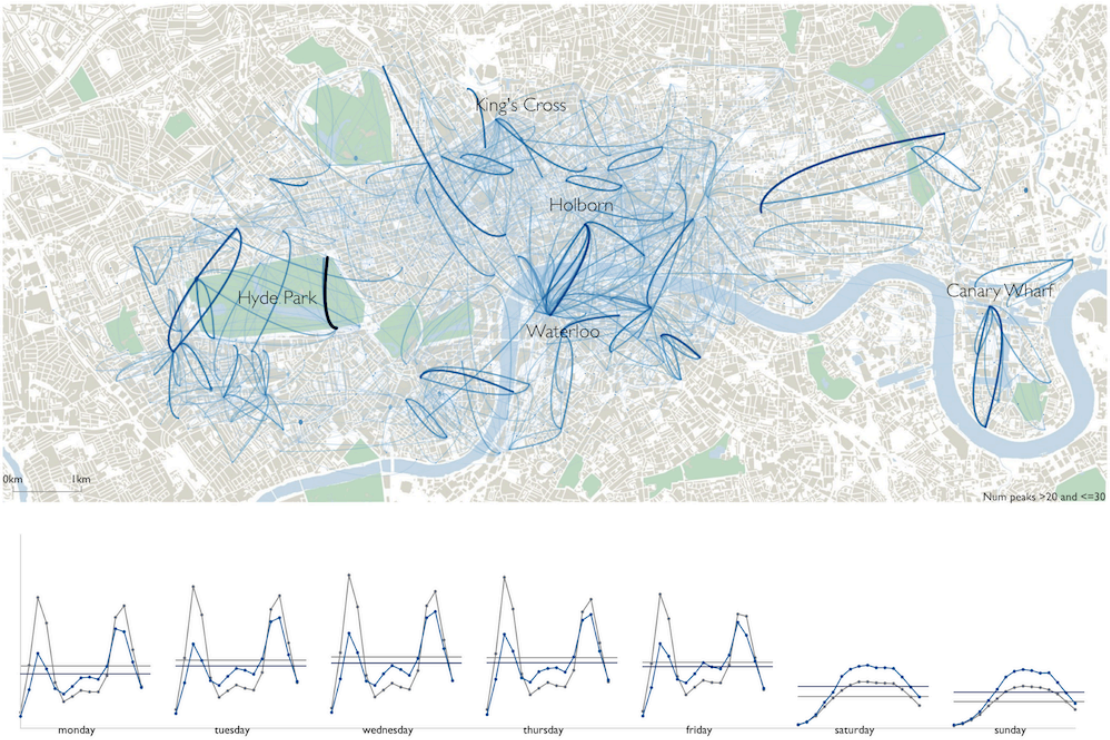

Figure 3: Journeys made by LCHS customers completing between 21-30 peak-time journeys are shown. Timeline: temporal structure of selected individuals' journeys (blue) is compared with that of the total LCHS population (grey). Map: flow lines represent journey pairs. Following Wood et al. (2011), the most common journeys are given greater visual saliency line thickness, colour and transparency varies by a flow weighting factor. The LCHS's three hubs, strategically important stations where bikes are manually re-balanced at peak-times, are labelled: King's Cross, Waterloo and Holborn. Canary Wharf, a non-central area of employment, and Hyde Park are also labelled. For members making at least 20 repeated peak-time journeys, travel behaviours begin to suggest a commuter function: 39% of cyclists making between 21-30 peak-time journeys have completed at least one journey involving a commuting hub station; for those making between 1-10 peak-time journeys this figure is significantly smaller ($p<0.001$), at 19%. Background mapping uses Ordnance Survey data Crown copyright and database right 2013.

4.3 Finding workplaces 1: Mean-centres and *k-means* methods

In total 28,551 customers, 34% of the scheme's member population, are potential commuters; making more than 20 repeated peak-time journeys in the 12 months between 14th September 2011 and 14th September 2012. From this group, we analyse the spatial distribution of peak-time travel activity, and for each cyclist aim to derive a spatial area that represents their workplace.

4.3.1 Mean-centres method

We first use a very simple technique for deriving workplace centres. After identifying all peak-time destinations during the morning and origins during the evening, we calculate the (frequency) weighted centroid of docking station locations each cyclist uses. We speculate that all inbound journeys in the morning and outbound journeys in the evening that lie within a user-defined distance of this mean-centre might be labelled commutes.

4.3.2 Visually evaluating the mean-centres method

We use visual analytics to evaluate the mean-centres classification. In doing so, we wish to analyse spatial patterns of individual cyclists' peak-time journeys, consider the (frequency) weighted centre of these locations, along with a buffer around this centre for including commuting journeys. An example of software developed specifically for this purpose appears in Figure 4. Blue dots represent docking stations (potential workplace locations) and are sized according to the number of peak-time journeys that customer has made to/from those docking stations. In order to evaluate the spatial dispersion of these locations, we draw in grey a standard deviation ellipse (Yuill 2011), the weighted-centre of which represents a customer's hypothesised workplace. In red, a buffer for filtering journeys is drawn. Through interaction, it is possible to distinguish between journeys made during the morning and evening peaks; to adjust the spatial buffer drawn around each centroid; and to iterate through and evaluate the classification for each member.

Iterating through the customer database, we find that in many cases our initial 'walking distance from centre' threshold of 500 metres (m) is too conservative; it appears to unnecessarily exclude journeys that might be reasonably linked to customers' mean-centres – customers' potential workplaces. We also find that in some instances spatial outliers exist. Here, journeys to docking stations a substantial dis-

tance from individuals' main workplace centres have the effect of displacing cyclists' weighted mean-centres. Perhaps more importantly, by iterating through various cyclists' journeys, we see that in some cases there is more than one set of workplace clusters (Figure 4). The distribution of points is either bi- or occasionally multi-centric: a cyclist may have two or occasionally three spatial clusters of peak-time 'workplace' locations. In these instances, if we were to simply use the weighted mean-centre of all peak-time commuting points, we would exclude almost all journeys an individual makes which, since they remain spatially clustered, will likely represent genuine commuting journeys.

Without representing individuals' peak-time journeys visually, these problems would perhaps not be as immediately obvious. Attending to the standard deviation of cyclists' potential workplaces, and ratios of the maximum and minimum dispersion calculated when creating standard deviation ellipses, we estimate that 20% of commuting LCHS users are likely to make journeys to or from more than one spatial workplace cluster.

Figure 4: Application for validating commuting mean-centres. Left: tri-centric distribution of workplace locations for a single member. Right: bi-centric distribution of workplace locations for a single member. Standard deviation ellipses are drawn in grey; blue dots represent commuter destinations (am) and origins (pm) and are sized by relative frequency. A suggested 500m buffer for excluding journeys appears in red. Background mapping uses Ordnance Survey data Crown copyright and database right 2013.

4.3.3 *K-means* method

A second solution, which might overcome at least one shortcoming of the mean-centres method (the fact there are multiple workplace clusters), is the *k-means* clustering algorithm. We run *k-means* analysis on each cyclist's potential set of workplace docking stations, using information from the standard deviation ellipses to specify the number of workplace clusters (k) a cyclist is likely to have. If the ratio of dispersion and standard distance exceeds a certain threshold, then we look for k=2 or k=3 cluster centres. Alternative methods such as *hierarchical* cluster anal-

ysis (HCA) (O'Sullivan & Unwin 2002) might better enable an appropriate number of clusters to be specified. However, the output of each HCA would need to be inspected at an individual cyclist level, and scaling this analysis to the full LCHS customer population would therefore be problematic.

4.3.4 Visually evaluating the *k-means* method

Running *k-means* clustering algorithms on a sample of member records, and inspecting the classification within the same set of visualization software described in the previous, we find that our rules for predicting the number of clusters are not always successful: a high level of distance deviation and large ratio of dispersion clearly may not only suggest a bi-centric distribution (Figure 5). Separate to this issue of correctly specifying an appropriate number of clusters, the same problem of extreme spatial outliers displacing cluster centres also exists.

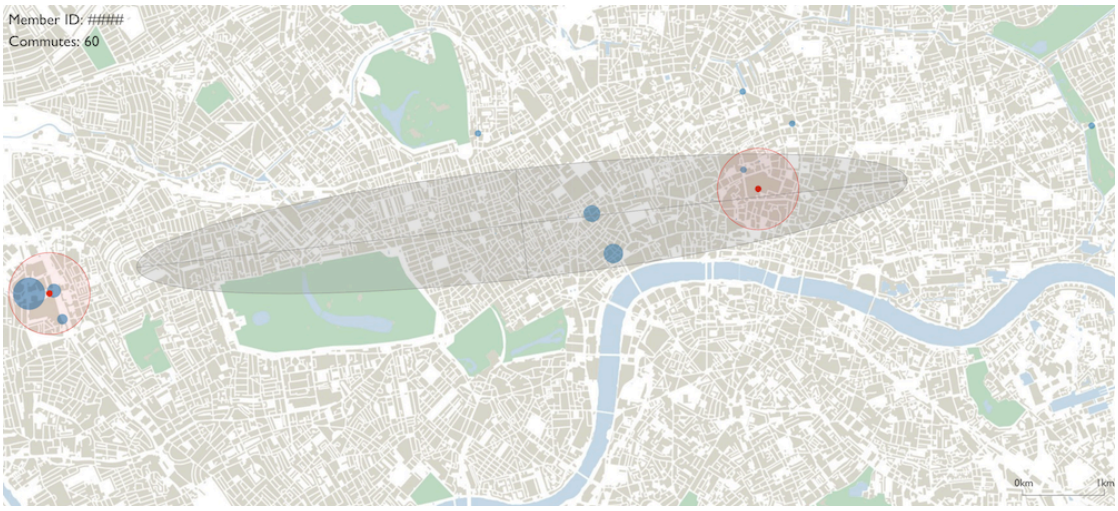

Figure 5: *K-means* analysis assuming standard distance and ratio of dispersion criteria are such that a 2-cluster solution is required and therefore k=2. A primary workplace cluster is successfully identified (far left), but we miss a second centre of activity. Background mapping uses Ordnance Survey data Crown copyright and database right 2013.

4.4 Finding workplaces 2: density-estimation method

Conscious of these two pitfalls, we finally speculate that a density-based method (O'Sullivan & Unwin 2002) might be most successful in both coping with multiple spatial clusters of workplace centres and, importantly, negating the displacement problem caused by spatial outliers. The limitation of such an approach is that we no longer have spatial coordinates representing individuals' workplace centres.

4.4.1 Density-estimation method

In our density-estimation method we make density observations at every potential workplace docking station. For each LCHS member, we identify all peak-time destination (am) and origin (pm) docking stations and, at each docking station, sum the total number of journeys made to and from that docking station, and to and from neighbouring docking stations that are within a user-defined walking distance. This user-defined distance serves as our kernel in the density estimation. Once the density estimates have been made, we propose that all journeys to docking station locations whose density counts are below a certain threshold, and therefore whose spatial areas are not visited frequently at peak times, should be labelled as non-commuting journeys.

4.4.2 Visually evaluating the density-estimation method

For the density-estimation technique to be successful, we must experiment with two parameters: the size of the kernel and density threshold that we use to exclude or include peak-time journeys. We again argue that showing this analysis visually better enables the technique to be evaluated, and Figure 6 displays a further set of software designed for this evaluation. We use very similar encodings as in our previous software. We show docking stations as dots, sized according to the number of peak-time journeys they receive. The kernels used in the density-estimation appear as transparent circles centred at each docking station; they can be resized, and the classification subsequently recalculated 'on the fly'. Docking stations that meet our acceptance criteria appear in blue; journeys to and from excluded docking stations (non-commutes) appear in red.

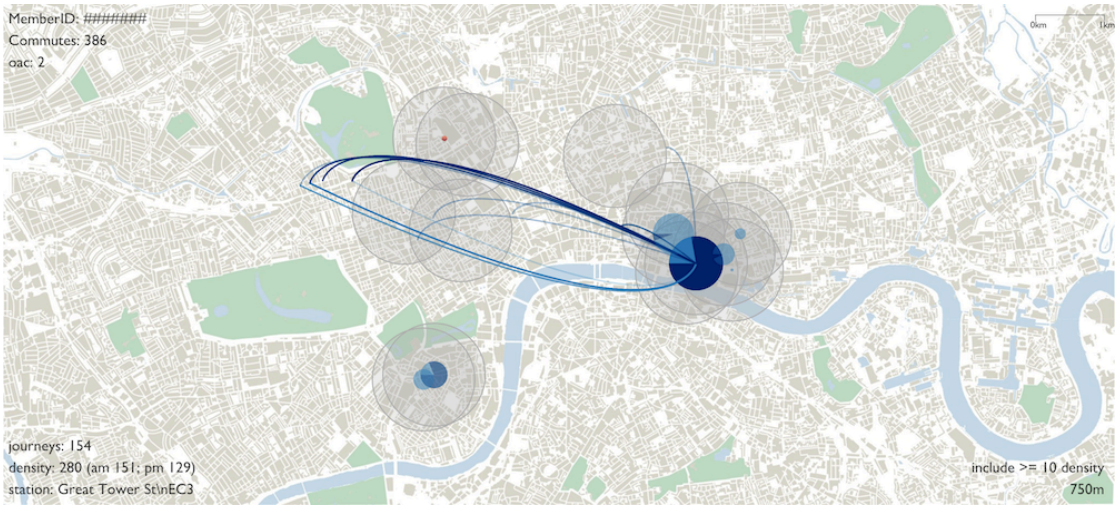

Figure 6: Density-estimation for a single cyclist is explored. Docking stations classified as commutes appear in blue, non-commutes in red. Resizable kernels are shown in grey. Morning journeys to workplaces (light blue/red) can be distinguished from evening (dark blue/red) journeys. By mousing over each workplace docking station, morning journeys arriving at, and evening journeys departing from, that docking station are displayed. By making journey lines asymmetric, journey direction is shown: the straight end represents journey origin, the curved end journey destination. Background mapping uses Ordnance Survey data Crown copyright and database right 2013.

Using this software we find that the density-estimation technique appears to address our two previous problems. It is successful at accepting secondary and tertiary clusters, whilst still excluding outlier locations: where members apparently make commuting journeys to more than one spatial workplace cluster, those journeys are included whilst more exceptional activity is correctly treated as non-commuting behaviour. Since we no longer calculate spatial centres for individuals' workplaces, the problem of spatial outliers displacing workplace centres also no longer exists.

As discussed, two analytical decisions affect the success of this technique: the size of the kernel and the inclusion density criteria. A radius of 750m around each potential workplace docking station is selected as the preferred kernel width. This kernel size is partly chosen as colleagues at TfL report anecdotal information that 750m is generally an accepted maximum between-docking-station walking distance. By visually scanning the spatial distribution of members' potential commuting docking stations, we provide some empirical support to this claim. Iteratively exploring cyclists' peak-time journeys we rarely find apparent clusters with docking stations extending beyond this 750m buffer. Second, a local density threshold of 10 or more journeys is used for excluding docking stations whose total peak-time journey frequency, combined with other docking stations within our 750m kernel, fall below this limit. This is derived from iterating through each individual's journeys, interactively varying the inclusion density parameter and visually scanning accepted and rejected workplace docking stations. With an inclusion density greater than 15, we

begin to find false negatives: docking stations that lie within a spatial cluster and should be labelled as commutes, but are coloured as red and therefore rejected. An inclusion density of less than 5 leads to false positives: docking stations that are visited relatively rarely and that are spatially distinct from an individual's dominant workplace cluster are coloured blue and therefore wrongly accepted as workplaces.

Using the described visual analytics software, we are therefore able to quickly identify problems associated with our proposed analysis techniques and suggest appropriate threshold parameters for them. When applied to the full population of LCHS cyclists, there may be instances where a cyclist's commuting behaviour is still incorrectly labelled. However, with no clear rules for quantitatively evaluating each technique, our visual software allows proposed methods to be related directly to real cycling behaviours and, as a result, empirically-derived threshold values to be specified.

5 Findings

Selecting the density-estimation technique with a kernel of 750m and an inclusion density of 10 journeys, we find that 28,075 members (34% of the member population) making journeys in the 12 month study period are commuters, and in total classified commuting journeys represent 49% of all journeys taken between 14th September 2011 14th September 2012. In this section we reflect on our three initial research questions (RQs) around understanding commuting behaviour. We also discuss insights discovered through visually exploring the analysis discussed in section 4 with colleagues at TfL.

5.1 RQ 1: What are the characteristics of people who take part in commuting based activities?

Figure 7 is a geodemographic profile of commuting cyclists identified through our density-based classification. Compared to the total LCHS member population, men and high RF users are very significantly ($p<0.001$) overrepresented amongst commuting members. So too are members who live very close to a bike share docking station. Importantly, those that live outside of London - between 15km-100km from a docking station – are also overrepresented amongst commuting members. By contrast, LCHS users that live moderate distances from a docking station - between 500m-10km away - are significantly ($p<0.001$) underrepresented amongst commuting members given their share of the total LCHS population.

From this we might speculate that immediate proximity to a docking station is an important motivating factor for commuting usage. Those living less than 500m from a docking station can access their nearest bike station by foot, and the same might apply for those living between 15km-100km from a docking station. These individuals may live too far from central London to commute using the London underground or bus system, and instead travel into London using a commuter train, which, unlike the bus or underground, arrives only at major rail terminals, and therefore possibly a distance from commuters' workplaces. From there, it might be the case that LCHS bikes are easily available and a desirable option for completing the final leg of a journey.

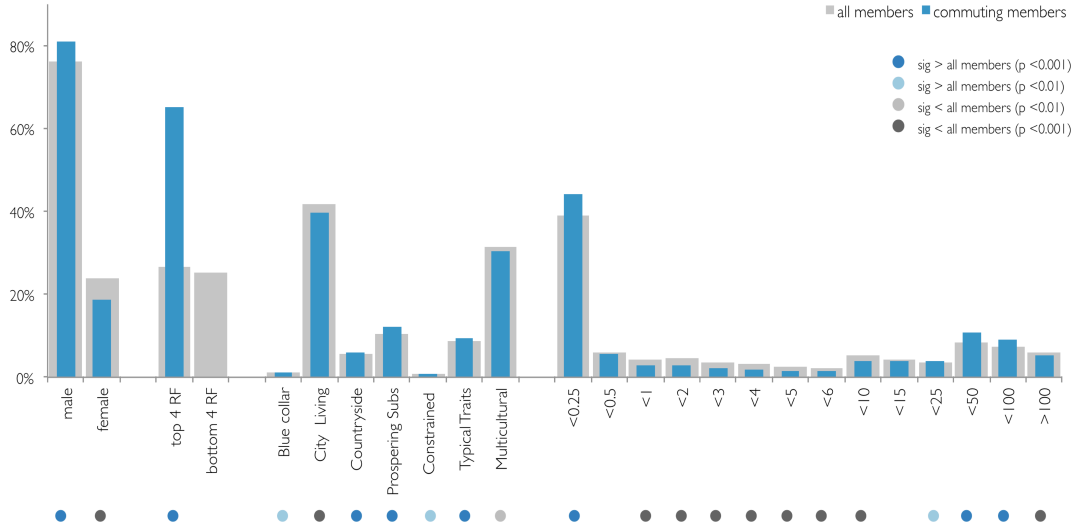

Figure 7: Geodemographic profile of commuters. Significance testing: contingency tables using the Pearson's chi-square test statistic are used to test for equality of proportions between commuters and all cyclists. Standardized residuals are used to identify which specific categories – for example, which OAC groupings – contribute most to the overall chi-square model. They are effectively z-scores, and can be used to assess category-level significance.

5.2 RQ 2: Where do commuting events happen?

The spatial patterns of journeys made by commuting members seem to support this claim regarding those commuting from outside of London. Commuting journeys tend to coincide with major rail stations, and this is particularly true for those who apparently live between 15km-100km from a docking station. Of all commuting journeys taken by members living this distance from a docking station, 22% involve journeys that start and end at hub docking stations located at two of London's major rail terminals - King's Cross and Waterloo (labelled in Figure 3). For those living less than 10km from a docking station this figure is just 3%.

Commuting journeys made by London-resident members appear more spatially diverse than for those commuting from outside of London (Figure 8). As well as the visual evidence in Figure 8, we can provide quantitative information to support this claim. Ordering cyclists according to the proportion of their commuting journeys that are unique, we find that for the median commuter amongst those living less than 10km from a docking station, 32% of commuting journeys are entirely unique OD pairs. This figure for commuters living between 15km-100km from a docking station is just 19%. The implication is that non-London commuters tend to consistently use a reduced number of docking stations when they commute. Additionally, travel times appear to be slightly more balanced between the morning and evening commute for London-resident commuters. For those living less than 10km from a docking station, 54% of commutes take place in the morning peak, whereas for those living 15km-100km from a docking station this figure is 56% (a significant difference, $p<0.001$).

Considering the initial problem introduced at the start of the paper around fleet management, our analysis confirms that it is the commuting behaviour of non-London members that is likely to put the greatest strain on the LCHS system. These individuals make more concentrated journeys in both space and time than London-resident users. If one extension of this argument is that commuting journeys made by London-resident users might be incentivised, then widening the geographic extent of the scheme into more residential parts of London may be one option. Further evidence to support this argument is that of all commuting members living less than 10km from a docking station, the majority (75%) live within just 500m of a docking station.

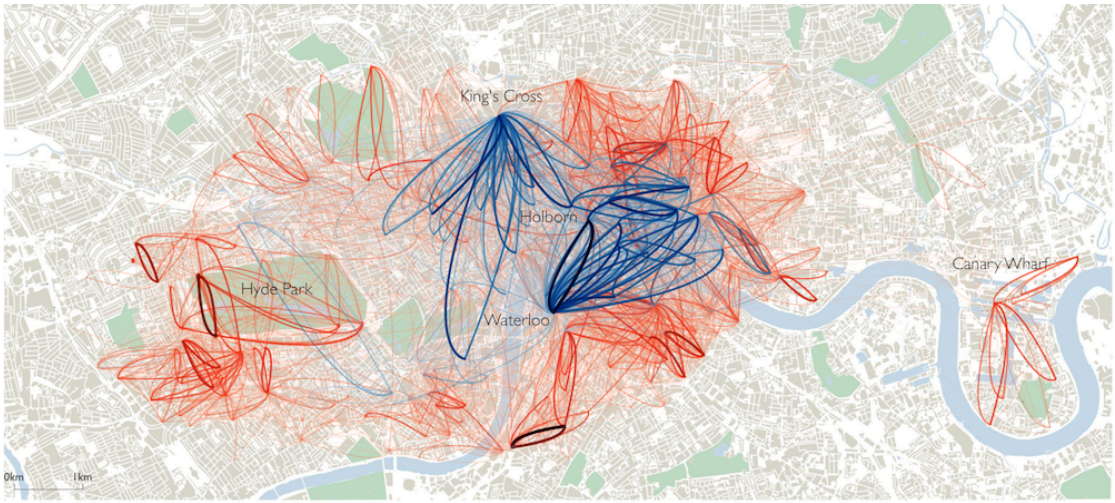

Figure 8: Map of derived commuting journeys: 'London' cyclists (living <10km from docking station) in red; 'non-London' cyclists (living <15km from docking station) in blue. As in Figure 3, flow lines are weighted according to relative journey frequency. Separate weighting factors are created for 'London' and 'non-London' commuters. Background mapping uses Ordnance Survey data Crown copyright and database right 2013.

5.3 RQ 3: Under what circumstances are journeys made during the working day?

An additional focus for this study is around journeys that are made as part of the working day – after having commuted into work in the morning and before commuting home from work in the evening. Labelling all journey events in the dataset makes this analysis possible. We identify all interpeak journeys (weekdays between 10am-3pm) and study whether, on the same day, members make a commuting journey either during the morning or evening peaks. In total 21,765 commuting members have made such interpeak working-day journeys; this represents 78% of the total commuting LCHS population.

There is some concentration of interpeak working-day journeys around London's universities: docking stations around the Bloomsbury area, where three universities are located, are a focus of interpeak working-day activity, and so too are journeys around a major university towards the south west of Hyde Park (labelled in figure 3). Spatially filtering interpeak working-day journeys, we find that the lunchtime peak is less severe in those parts of London with a concentration of universities: 22% of interpeak working-day journeys that involve docking stations within the vicinity of universities are taken between 12pm-1pm, whilst this figure for journeys within the City of London, London's commercial centre, is 26% (a significant difference, $p<0.001$). We speculate that this might reflect a higher incidence of utilitarian jour-

neys or delayed commutes taken by individuals employed or studying at universities. If this is the case, then incentivising usage within universities – by both students and university staff – may be one means of encouraging a more natural redistribution of bikes during the working day.

5.4 New insights into the geography of commuting cyclists' workplaces

In section 4, we discuss how, through visually depicting proposed analysis algorithms, we quickly identified problems with each method in the context of individual cyclists' journeys. This analysis process, and especially the design addition whereby we distinguish morning from evening peak-time journeys, also revealed interesting spatiotemporal patterns of apparent commuting travel. Discussing this analysis and software with colleagues at TfL, particularly with those working in operations, certain common behaviours were identified, which we speculate may relate to the scheme's design. As a result of these discussions, we designed a further set of visual software for collaboratively exploring the geography of classified workplaces at the scheme-wide level.

Figure 9 is an example of this application. Docking stations are again sized according to the number of inbound (in the morning) and outbound (in the evening) commuting journeys. As in section 4 we delineate between morning (blue) and evening (orange) journeys using colour, but this time we aggregate these journeys for all commuting cyclists. Essentially figure 9 is a map of 'global workplaces'. At the bottom-right, a slider allows these locations to be filtered according to the relative number of morning-evening commutes. Geodemographic variables appear as vertical bars. The bars change dynamically when data are filtered, and can be selected to identify particular subsets of the commuting LCHS population.

The figure shows one significant finding from this analysis: that 'global workplaces' with more evening than morning commutes are located entirely towards the periphery of the scheme. This is surprising since these workplaces represent the origin, not the destination, of evening commutes, and typically one associates bikes arriving at, rather than departing from, peripheral docking stations during the evening. Exploring this finding with colleagues at TfL, those responsible for LCHS operations drew attention to the fact that the spatial pattern in figure 9 relates very strongly to the initial bounds of the scheme (Transport for London 2012). Typically on weekdays, bikes are disproportionately transported from central London to such peripheral locations during the evening commute. These docking stations often remain full overnight; making it difficult for any commuters wishing to arrive at these

peripheral docking stations during the (early) morning commute. As a corollary, during the evening commute docking stations begin to fill up, and for those wishing to make outward journeys from these peripheral stations in the (later) evening, bikes are more readily available. Whilst this is only a hypothesis, it is an important finding as it is suggests that the redistribution problem might not only exist for individuals wishing to commute to very central parts of London.

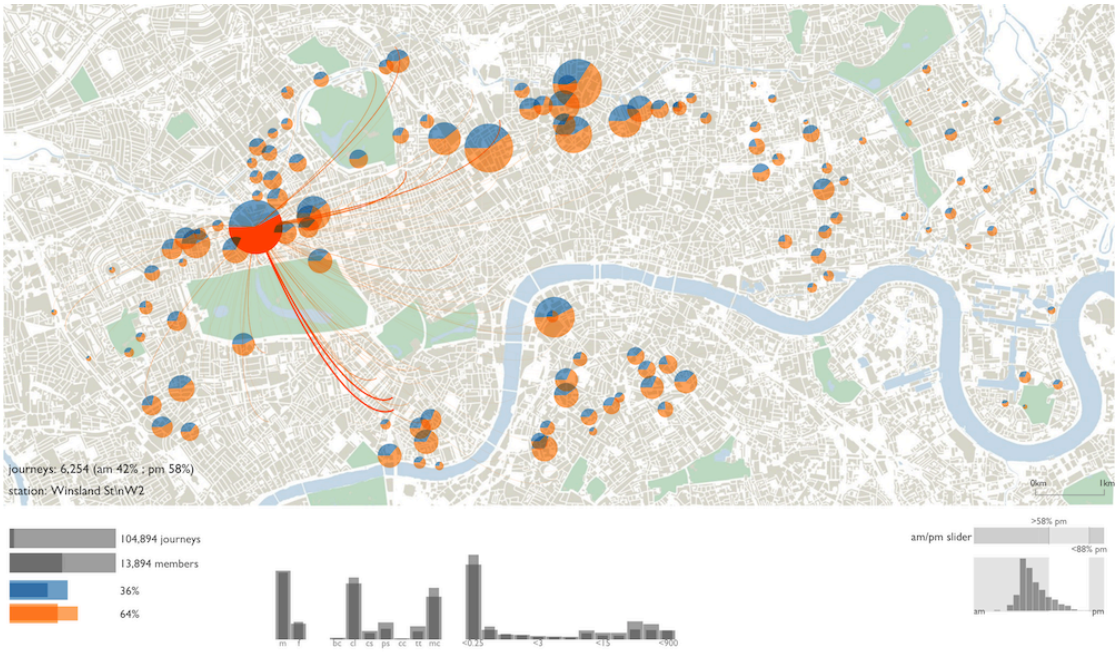

Figure 9: Application for exploring 'global workplaces'. Map: pie charts are workplace docking stations sized according to number of commutes arriving (blue) in the morning and departing (orange) in the evening. Mouse is currently held on docking station in middle left of view; its name, and number of commuting journeys is labelled and all evening commutes leaving that station are drawn on the map. Bottom: gender and geodemographic variables appear as bars; in am/pm slider, docking stations where more evening commutes depart from that station than morning commutes arrive are selected. Background mapping uses Ordnance Survey data Crown copyright and database right 2013.

6 Conclusion

We describe a method for identifying commuting behaviours from a large dataset of LCHS journeys. Unlike recent approaches that use unsupervised clustering algorithms to detect commuter-dominated behaviour, we identify commuting events by studying spatial patterns of cyclists' peak-time journeys. For each cyclist an empirically-defined workplace, or set of workplaces, is created and all journeys that arrive at this workplace in the morning and depart from this workplace in the evening are labelled as commutes. We use custom-built visual analytics software to evaluate

and explore approaches to identifying these workplace locations. Doing so enables problems associated with the techniques to be quickly diagnosed and suitable analysis parameters to be evaluated. Moreover, by displaying relatively abstract spatial analysis techniques within their spatial context, this analysis is made intelligible to analysis experts (ourselves) and domain specialists (colleagues at TfL), and also enables new hypotheses into commuting behaviours to be stated and explored.

The motivation for studying commuting behaviour is that extensive usage at peak times causes problems with maintaining a balanced bike-share system. From the derived commuting journeys we find two main types of commuting journeys: those taken by members living outside of London, who make commuting journeys particularly at major rail hub stations; and more varied commuting behaviours, both in space and time, by those living very close to a LCHS docking station. We also find evidence of interpeak journeys within the working day being made around London's universities. Visually exploring classified workplaces with colleagues at TfL highly engaged in the research problem, an important finding is of unexpected imbalances in the number of morning commutes to, and evening commutes from, derived workplaces that might relate to local availability of bikes.

These are new insights into LCHS usage, and we are currently working with colleagues at TfL on their implications for the scheme's operation. Our method for identifying individual commuting events from analysing individuals' spatiotemporal travel behaviours may be relevant to others with similar timed origin-destination datasets. Moreover, the *chauffeured* analysis approach, whereby a diverse set of colleagues at TfL articulate a research problem, we propose a set of solutions and, using visual analytics, this solution space is explored collectively, might translate to many applied research contexts.

7 References

Agard, B., Morency, C. & Trepanier, M. (2006), Mining public transport user behaviour from smart card data, *in* 'The 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)', Saint-Etienne, France.

Arias-Hernandez, R., Kaastra, L., Green, T. & Fisher, B. (2011), Pair analytics: Capturing reasoning processes in collaborative visual analytics, *in* '44th Hawaii International Conference on System Sciences (HICSS)', Minoa, Hawaii.

Beecham, R. & Wood, J. (2013), 'Exploring gendered cycling behaviours within a large-scale behavioural data-set', *Transportation Planning and Technology* pp. 1–15.

Department for Communities and Local Government (2011), The English indices of deprivation 2010: Technical report, Technical report, Department for Communities and Local Government.

Lathia, N., Ahmed, S. & Capra, L. (2012), 'Measuring the impact of opening the London shared bicycle scheme to casual users', *Transportation Research Part C: Emerging Technologies* **22**, 88–102.

Lathia, N., Smith, C., Froehlich, J. & Capra, L. (2013), 'Individuals among commuters: Building personalised transport information services from fare collection systems', *Pervasive and Mobile Computing* **9**, 643–664.

Novo, J. (2004), *Drilling Down: Turning Customer Data into Profits with a Spreadsheet - Third Edition*, 3 edn, Booklocker.com, Inc.

Nunamaker, J., Dennis, A., Valacich, J., Vogel, D. & George, J. (1991), 'Electronic meeting systems to support group work', *Communications of the ACM* **34**(7), 40–61.

OBIS (2011), Optimising bike sharing in european cities, Technical report, OBIS Handbook.

O'Sullivan, D. & Unwin, D. (2002), *Geographic Information Analysis*, John Wiley Sons, New Jersey, USA.

Robinson, A. (2008), Collaborative synthesis of visual analytic results, *in* 'IEEE Symposium on Visual Analytics Science and Technology', Columbus, Ohio, USA.

Shekhar, S., Evans, M., Kang, J. & Mohan, P. (2011), 'Identifying patterns in spatial information: A survey of methods', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**(3), 193–214.

Slingsby, A., Dykes, J. & Wood, D. (2011), 'Exploring uncertainty in geodemographics with interactive graphics', *IEEE Transactions on Visualization and Computer Graphics* **17**(12), 2345–2554.

Thomas, J. & Cook, K. (2006), 'A visual analytics agenda', *IEEE Computer Graphics and Applications* **26**(1), 10–13.

Tukey, J. W. (1977), *Exploratory Data Analysis*, 1 edn, Addison-Wesley, London.

Tukey, J. & Wilk, M. (1966), Data analysis and statistics, an expository overview, *in* 'International Workshop on Managing Requirements Knowledge', Los Alamitos, CA, USA.

Vickers, D. & Rees, P. (2006), 'Introducing the area classification of output areas', *Population trends* (125), 15–29.

Wood, J., Slingsby, A. & Dykes, J. (2011), 'Visualizing the dynamics of London's bicycle hire scheme', *Cartographica* **46**(4), 239 – 251.

Yuill, R. (2011), 'The standard deviational ellipse: An updated tool for spatial description', *Geografiska Annaler. Series B, Human Geography* **53**(1), 28–39.