
Article:
Bwambale, A, Choudhury, CF orcid.org/0000-0002-8886-8976 and Hess, S orcid.org/0000-0002-3650-2518 (2019) Modelling departure time choice using mobile phone data. Transportation Research Part A: Policy and Practice, 130. pp. 424-439. ISSN 0965-8564

https://doi.org/10.1016/j.tra.2019.09.054

Modelling departure time choice using mobile phone data

Andrew Bwambale
Choice Modelling Centre
Institute for Transport Studies
University of Leeds
34-40 University Road, LS2 9JT, Leeds, United Kingdom
Email: ts13ab@leeds.ac.uk

Charisma F. Choudhury
Choice Modelling Centre
Institute for Transport Studies
University of Leeds
34-40 University Road, LS2 9JT, Leeds, United Kingdom
Email: C.F.Choudhury@leeds.ac.uk

Stephane Hess
Choice Modelling Centre
Institute for Transport Studies
University of Leeds
34-40 University Road, LS2 9JT, Leeds, United Kingdom
Email: S.Hess@its.leeds.ac.uk

Submission Date 27 September 2019

Abstract

The rapid growth in passive mobility tracking technologies has led to departure time choice studies based on GPS data in recent years (e.g. Peer et al., 2013). GPS data however typically has limited sample sizes and is affected by technical issues like signal losses and battery depletion leading to gaps in the data. On the other hand, the rapid growth in mobile phone penetration rates has led to the emergence of alternative passive mobility datasets such as Global System for Mobile communication (GSM) data. GSM data covers much wider proportions of the population and can be used to infer departure time information. This motivates this research where we investigate the potential use of GSM data for modelling departure time choice. We describe practical approaches to extract relevant information from GSM data and propose a modelling framework that accounts for the fact that the desired departure times are unobserved. We assume that the preferred departure times vary randomly across the users and apply the mixed logit framework to jointly estimate the distribution parameters of the preferred departure times and the sensitivities to schedule delay. Comparison of the model results and time valuation metrics derived from the GSM data with similar metrics derived from the GPS data of a subset of the users reveals that the fewer time gaps in the GSM data lead to reliable model outputs. The proposed framework can be used for mobile phone and other passive data sources with unobserved preferred departure times.

Keywords: Time of travel, GSM data, GPS data, schedule delay, time valuation

1 Introduction

The modelling of time-of-travel choices has over the years emerged as an important and challenging issue worth consideration under travel demand management through policy measures such as congestion pricing and flexible working hours (Hess et al., 2007b). Time-of-travel choices are principally a trade-off between enduring longer travel times during peak demand periods to ensure punctual arrivals at the target destinations versus avoiding the peak periods and opting for earlier or later arrivals to reduce the travel times (Small, 1982), although in a toll road setting there may be additional differences in toll in the peak.

In most practical applications, time-of-travel choice problems have been expressed as scenarios where an individual is faced with a finite number of discrete departure time periods and chooses the alternative with the highest utility (Cosslett, 1977, Small, 1982). Departure time choice models, which are basic functions of the factors affecting departure time decisions are thus important tools for predicting travel demand and evaluating alternative measures for managing this demand.

Departure time choice models have largely been developed using traditional stated preference datasets (e.g. Thorhauge et al., 2017, Ramos et al., 2017, Arellana et al., 2013, Arellana et al., 2012, Hess et al., 2007b, Hess et al., 2005, De Jong et al., 2003, Daly et al., 1990, Bates et al., 1990) and revealed preference datasets (e.g. Arellana et al., 2013, Bhat, 1998a, Bhat, 1998b, Small, 1987, Small, 1982, Abkowitz, 1981). The former are prone to hypothetical bias and behavioural incongruence while the latter are generally expensive to obtain, prone to reporting errors, and typically involve small samples. This problem is particularly common in developing countries, where stringent budget constraints for transport studies act as a barrier for large-scale data collection and there is very limited research on modelling departure time choice (e.g. Ramos et al., 2017, Arellana et al., 2013). Another reason for the use of stated preference data has been that the correlations inherent in revealed preference data make it difficult to capture the trade-offs between changes in departure time and other variables.

The last few decades have been characterised by rapid growth in technologies that enable the passive collection of individual mobility trajectories. This has led to departure time choice studies based on GPS data from smartphones (e.g. Peer et al., 2013). Although the use of smartphone apps has reduced costs, such studies remain expensive and thus usually involve small samples.

However, the rapid growth in mobile phone penetration rates worldwide (GSM Association, 2017) has led to the emergence of network-generated passive mobility datasets such as Call Detail Records (CDRs)[1] and Global System for Mobile communication (GSM)[2] data. These datasets can anonymously cover much wider proportions of the population using their current mobile handsets, without additional expenses such as recruiting of respondents and procuring of smartphones. Such mobile phone datasets have been successfully used in various transportation planning applications (Çolak et al., 2015, Iqbal et al., 2014, Jiang et al., 2013, Isaacman et al., 2012, Schlaich, 2010). However, a review of the literature shows that there is no study using such data to model departure time choice decisions. This motivates this research where we investigate the potential of GSM data for departure time choice modelling. It may be noted that GSM data is deemed to be more appropriate for capturing departure time choices due to its semi-continuous nature as opposed to CDR location data, which is typically discontinuous.

Since GSM data generation only requires the users' mobile phones to be active, the regular location area updates by the network operator make it possible to capture most of the trips made.

[1] CDR data typically consists of the time stamped locations of the responding tower that handles a call/text/web access request from a user as well as the details of the request (type, sender/receiver, etc.).

[2] GSM data reports the IDs of all the GSM cells traversed by an active mobile phone (i.e. a phone-set with a valid sim that is switched on) at regular time intervals.

However, it is important to highlight the limitations of GSM data in the context of departure time analysis. The coarse location resolution of GSM data makes it impossible to capture intra-cell movements as well as the actual arrival or departure times from points within the cells. Instead, it is only possible to observe the cell boundary crossing times, especially where the GSM cells are recorded at short time intervals (e.g. 60 seconds in this study). It is worth noting that the differences between the actual departure and the (post-departure) cell boundary crossing times as well as the differences between the actual arrival and the (pre-arrival) cell boundary crossing times reduce as the GSM cell sizes become smaller. This is the case for most metropolitan areas where GSM cellular networks are dense, with small cell sizes that can go as low as 100 metres (e.g. De Groote, 2005). This implies that the cell boundary crossing times would still be within minutes from the actual departure or arrival times.

The above points motivate us to explore the potential of GSM data alongside (or instead of) GPS data for modelling departure time choice to inform policy measures related to big data adoption for transport studies. We use the Nokia Mobile Data Challenge (MDC) dataset (Laurila et al., 2012, Kiukkonen et al., 2010), which includes both GSM and smart phone GPS data and enables us to get an idea about the strengths and weaknesses of the GSM data in terms of extracting information for departure time analysis. Departure time choice models are then developed using advanced discrete choice modelling techniques. We focus on modelling departure time choices during peak periods as these are most critical in transport planning and operation. The study also proposes a theoretical approach for dealing with the absence of information on the desired times of travel in passively collected data. The proposed approach is unique in that it allows us to understand the sensitivities as well as the valuations attached to schedule delay despite the passive nature of the data. Furthermore, we propose a practical approach for imputing missing travel time data for some of the time intervals in the analysis period.

The remainder of the paper is arranged as follows; section 2 presents a brief review of relevant literature, section 3 describes the data used for this study and the associated challenges, section 4 presents the modelling framework, section 5 presents the model results, while section 6 presents the summary and conclusions of the study.

2 Literature review

Departure time choice decisions generally involve a trade-off between the travel time and the schedule delay associated with a given time period. However, estimating the schedule delay requires knowledge of the desired times of travel. Most stated preference datasets for departure time choice modelling collect information on the desired times of travel which makes it easy to estimate the schedule delay terms. However, this is not usually the case for revealed preference data, especially passively collected data such as mobile phone data. Peer et al. (2013) is an exception where users were asked to report their desired times of travel. Previous studies have tried to address this issue in different contexts as summarised below.

Hess et al. (2007a) propose the use of time period specific constants to capture the aggregate scheduling preferences (among other effects) in the absence of the desired times of travel. However, Ben-Akiva and Abou-Zeid (2013) argue that the time period specific constants only capture the schedule delays if they are specified differently for each socio-economic group based on the assumption that individuals in the same socio-economic group have the same desired times of travel. This however results in the explosion of constants in the model specification, an issue that can be addressed with functional forms to approximate the alternative specific constants (Hess et al., 2005). However, another important point to highlight is that relying solely on constants to capture scheduling makes it difficult to understand the continuous sensitivity to delay.

On a different note, Koppelman et al. (2008) propose an approach where the schedule delay for a particular departure time period is estimated as the weighted mean of all the possible schedule

delays with respect to the different time periods, where the weights are estimated from a time-of-day distribution of the observed departure times represented by a trigonometric function. However, a potential issue with this approach is that it assumes a strong correlation between the schedule delays and the observed time-of-day distributions, which may not be the case. A slightly related approach is proposed by Kristoffersson and Engelson (2018) who apply reverse engineering techniques that rely on a previously estimated departure time choice model to derive conditional departure time probabilities (given the preferred departure time), which are then combined with the observed departure time distributions for groups of O-D pairs to derive the weights for each preferred departure time period using ordinary least squares. However, a potential drawback with this approach is that previous models may be non-existent, and where they exist, there may be serious consequences with regard to model transferability.

Finally, Brey and Walker (2011) propose a hybrid choice framework in which the preferred times of travel are assumed to be latent and varying across individuals, and parameterise the probability density function as a mixture of normal distributions. However, in their framework, the latent preferred times of travel are explained using the trip and the travellers' characteristics, and are measured against the stated preferred times of travel (indicator variables) obtained from a survey, which is not possible in this case study. In this study, we propose a simple alternative approach which is described in Section 4 of this paper under the modelling framework.

However, besides using utility theory based models, it is important to note that departure time choices can be analysed using alternative methods involving the application of rule-based and data mining algorithms (e.g. Ettema et al., 2005, Xiong and Zhang 2013). It is however difficult to apply these algorithms in disentangling the causal relationships between departure time and the influencing factors and use them for forecasting in different policy scenarios.

3 Data

This study uses the Nokia Mobile Data Challenge (MDC) dataset collected as part of the Lausanne Data Collection Campaign (LDCC) between 2009 and 2011 (Laurila et al., 2012, Kiukkonen et al., 2010). The subsequent sections describe the study area, the mobile phone data, and the processes undertaken to extract relevant information for departure time analysis.

3.1 Study area

The main study area is Lausanne, located in southwestern Switzerland, however, the spatial coverage of the data covers the entire country.

Lausanne has a dense GSM cellular network with small cell sizes (see Schulz et al., 2012 for details). The small cell sizes make the area generally suitable for the current study as the actual departure or arrival times, which are unobservable for GSM data would still be within minutes of the observed cell boundary crossing times.

Another key aspect of Lausanne is that over 68% of the residents are working commuters, and over 90% of these use motorised transport modes, which are usually affected by peak period delays e.g. due to traffic congestion (ThemaKart, 2017). The travel times in Lausanne typically increase by 44% and 63% during the morning and the evening peak periods, respectively (TomTom, 2016).

3.2 Data description

The MDC dataset contains several types of records such as demographic data, GSM data, GPS data, call logs, and bluetooth data etc. However, this study only uses the demographics, the GSM, and the GPS data, which are described in the subsequent sections.

3.2.1 Demographic data

The MDC data is one of the few available mobile phone datasets with user demographic details, however, the sample size is small given that participation was voluntary. The available data

comprises of 83 full-time workers. The other available demographics for each of these include the gender and the age-group as summarised in Table 1.

Table 1: Demographic data summary statistics

Characteristic	Description	Number	Proportion (%)
Gender	Female	23	27.71
	Male	60	72.29
Age-group	Under 28 years	24	28.92
	28 years and above	59	71.08

It is important to note that although demographic data is available in this case, such data is usually unavailable in most mobile phone datasets due to privacy reasons. Previous studies have focused on the subject of demographic prediction and how this can be incorporated into transport modelling frameworks (see Bwambale et al., 2017 for details). However, since this is not the main focus of this paper, we directly incorporate the reported demographics into the models.

3.2.2 GSM and GPS data

The GSM data reports all the GSM cells traversed by each user's mobile phone at an interval of approximately 60 seconds. The data contains approximately 24.8 million records generated by the full-time workers. Each record is described by a user ID, a unique internal ID of the GSM cell, the unix timestamp and time zone. Table 2 presents an excerpt of the GSM data.

Table 2: Excerpt of the GSM data

User ID	GSM Cell ID	Unix timestamp	Time zone
5451	686	1251762486	-7200
5451	686	1251762546	-7200
5451	1785	1251762606	-7200
5451	1785	1251762663	-7200

The GPS data (timestamped latitudes/longitudes) was collected concurrently with the GSM data using the users' smartphone GPS receivers, which allows for cross-comparison of the two datasets. Despite the higher time resolution of the GPS data versus the GSM data (i.e. 10 seconds versus 60 seconds), the GPS data contains only 5.2 million records.

GSM data was collected as long as the user's mobile phone was switched on albeit that signal losses were possible, while for GPS data, the facility needed to be enabled. The average data collection period per user was 278 and 205 days for the GSM and the GPS data respectively. We use the concept of active time to refer to the time when a user's phone (or GPS service) is switched on, where this is assumed to be the case as long as the time interval between successive records does not exceed 10 minutes. For GSM data, the proportion of active time was on average 73.44% across users, compared to 5.00% for the GPS data. The corresponding median values and lower quartiles were 77.11% and 66.56% respectively for the GSM data compared to 4.37% and 3.16% respectively for the GPS data.

Time gaps in the GPS data may be caused by GPS disabling (e.g. due to battery issues) or signal loses in urban environments, buildings and tunnels (NCO, 2018, Gong et al., 2012, Chen et al., 2010). However, it is important to note that GPS technology has since improved and most of these issues may not be encountered in more recent GPS datasets. This observation of course does not affect the findings in terms of the potential use of GSM data. The methodology to extract meaningful information from both datasets is explained in the next section.

3.3 Data processing

The data processing methodology is presented in Figure 1. In this section, we briefly describe the key aspects of each major step.

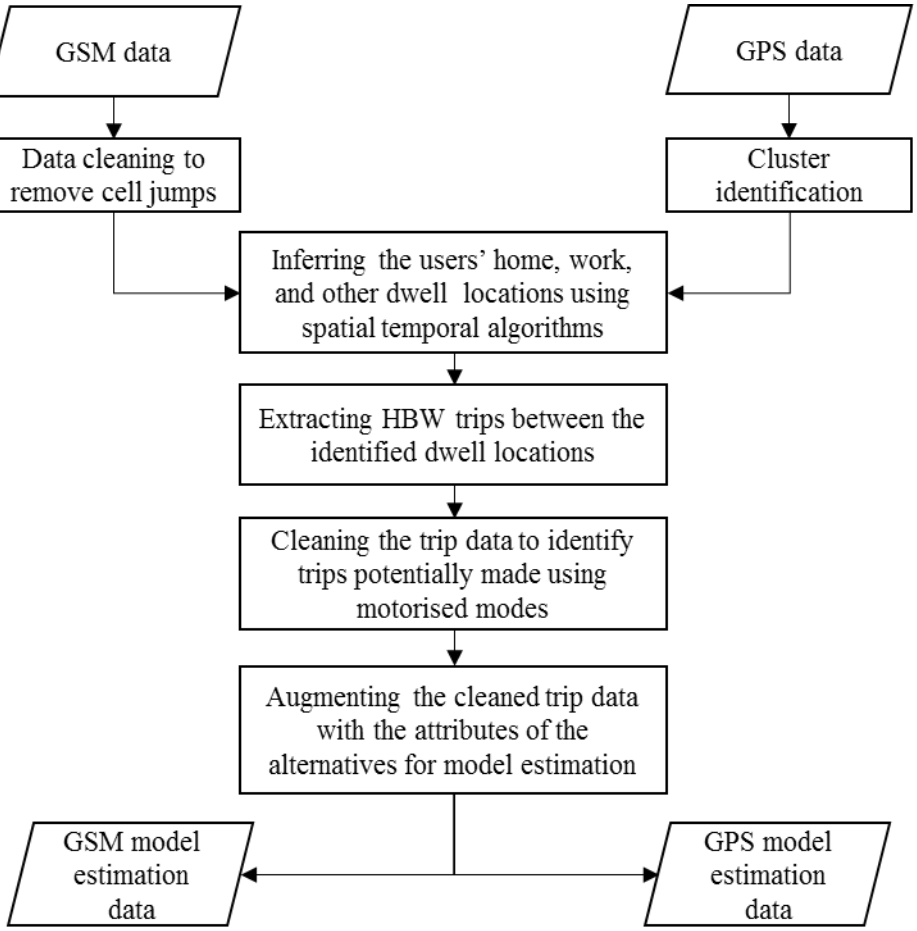


Figure 1: Summary of the data processing methodology

3.3.1 Data preparation

a. GSM data

GSM data is noisy in nature as it sometimes contains cell jumps that do not represent actual movement. The noise is mainly caused by cell tower call balancing operations aimed at optimising the quality of calls, which makes mobile operators assign mobile phones to neighbouring cells even when these phones are not physically located within those cells (Çolak et al., 2015). To mitigate cell jumps, the ordered GSM cell sequence of each user was analysed to calculate the time periods between intermittent observations of the same cell, and those with time periods less than 10 minutes were treated as cell jumps, thus relabelling the cell observations between them accordingly (Iqbal et al., 2014). The cleaned GSM cell sequences were then analysed to extract the users' dwell locations as described in the next section.

b. GPS data

The technical issues associated with GPS data as highlighted in Section 3.2.2 may sometimes not lead to total signal loss, but rather, may lead to inaccurate GPS locations. Furthermore, GPS location accuracy is affected by factors such as the quality of the GPS antenna in the smartphone, and the density of GPS satellites at the current location (NCO, 2018). Due to these factors, it is likely that different GPS points in the vicinity of one another could be linked to the same dwell location and this requires the application of spatial clustering techniques to identify the GPS point

clusters. We conducted complete-linkage hierarchical clustering (Everitt et al., 2011, Murtagh, 1985) with a threshold distance of 300 meters as used in previous studies (Çolak et al., 2015, Jiang et al., 2013). This is detailed in Appendix A.

3.3.2 Identification of home and work locations

A home location was defined as the GSM cell or GPS point cluster in which a user was observed for the longest time between midnight and 6 am on a particular day, while a work location was defined as any location other than the home location in which a user spent the longest time between 8 am and 5 pm on a particular working day. All the other GSM cells or GPS point clusters in which a user was seen to spend more than 10 minutes were described as 'other' dwell locations (Çolak et al., 2015, Jiang et al., 2013). We analysed each day separately to capture the changes in each user's home and work locations across the observation period.

An important point to note is that the original GSM data did not have the coordinates of the tower positions due to privacy reasons. The data only reports the IDs of the GSM cells without showing their positions as described in Section 3.2.2. Although the data alone is able to show the cell sequences and dwell times, it does not show the relative positions of these cells. To address this problem, the GPS points from all the users observed within 30 seconds of a GSM cell were extracted, and the mean latitudes and longitudes calculated for each cell. Although this was not possible for all the GSM cells, 62% of the inferred home cells, 70% of the inferred work cells, and 61% of the 'other' dwell locations were successfully matched. The data matching process was critical as it enabled the estimation of the distances and the travel speeds between the dwell cells, which we use to identify trips potentially made using motorised modes. It should be noted that under normal circumstances, GSM data should report the coordinates of the towers linked to the cells, in which case data matching would not be necessary.

3.3.3 Extracting HBW trips between the dwell cells

The focus of our analysis is home-based work (HBW) trips. Unlike direct trips, which have no en-route activity, trips with intermediate dwell locations (i.e. those labelled as 'others') could have en-route activities that last very long to the extent that such trips can no longer be categorised as clear HBW trips. In this study, we specified an upper limit of one hour on the total duration across all the intermediate stops and included only the trips satisfying this criterion in our HBW model.

During the extraction of HBW trips, we checked whether each user's phone or GPS receiver was active both on departure and at arrival. For departure, we checked the time difference between the last observation in each departure dwell location and the first observation outside that location, while for arrival, we checked the time difference between the first observation in each arrival dwell location and the preceding observation outside that location. In this study, we specified a threshold of 2 minutes to ensure that we capture reasonably accurate trip start and end times without losing significant portions of the samples[3]. This, for example, helped us avoid situations where a user's phone was switched off during the trip, and switched back on several hours after arrival. The trips meeting all the above conditions were then taken through the subsequent stages as described in the next sections.

3.3.4 Cleaning the trip data to identify trips potentially made using motorised modes

From a policy perspective, the focus is usually placed on motorised traffic, which is the main source of traffic congestion. However, one of the general limitations of mobile phone data is its anonymous nature, and therefore, the modes of transport used by the users are not known. A few previous studies have explored the possibility of detecting travel modes from mobile phone data (e.g. Qu et al., 2015, Doyle et al., 2011), however, as this is not the main focus of this study, we apply simple heuristics from the literature to infer the trips potentially made using motorised

[3] 2 minutes corresponds to the 99th percentile time difference between subsequent GPS and GSM records in the full datasets excluding time-gaps above 10 minutes.

modes. Observing a median speed above 15 kilometres per hour for a trip length above 5 kilometres is considered a good indicator that a user generally uses motorised transport for that trip chain (Hydén et al., 1999). Details of the applied heuristics are presented in Appendix B.

3.3.5 Augmenting the cleaned trip data with the attributes of the alternatives

Although it may seem convenient to assume that a user's choice set only comprised of the departure time intervals ever observed for the user across the different days in the sample, such an assumption is unrealistic since the failure to observe certain time intervals does not necessarily mean they were not considered. Although some factors like individual work flexibility have a bearing on the time periods to be considered, given the current data limitations, it seems more reasonable and safer to assume that all the departure time intervals were available and potentially considered (as opposed to making ad hoc assumptions). This implies a need to calculate the attributes for those time periods for which no actual trips were observed, a process described in this section.

The morning and the evening peak periods were divided into 15-minute intervals. The average travel times associated with each of these intervals were estimated for each of the user's trip chains using timestamps in their cleaned GSM and GPS data. The estimated travel times were then combined with time-period specific congestion factors to impute the travel times for the unobserved time intervals. The time-period specific congestion factors were estimated with the aid of the Google Maps direction tool, which predicts the average travel times between a given O-D pair at different departure or arrival times (Google Maps, 2018).

To reduce this task to manageable proportions, we divided Lausanne into 10 representative zones bounded by the major roads, thereby generating 90 O-D pairs as shown in Figure 2. For each O-D pair, we extracted the travel times associated with each 15-minute interval between 5 am and 11 am (for the home-to-work commute) and 3 pm to 9 pm (for the work-to-home commute). The average travel times for each interval across all the O-D pairs were then determined.

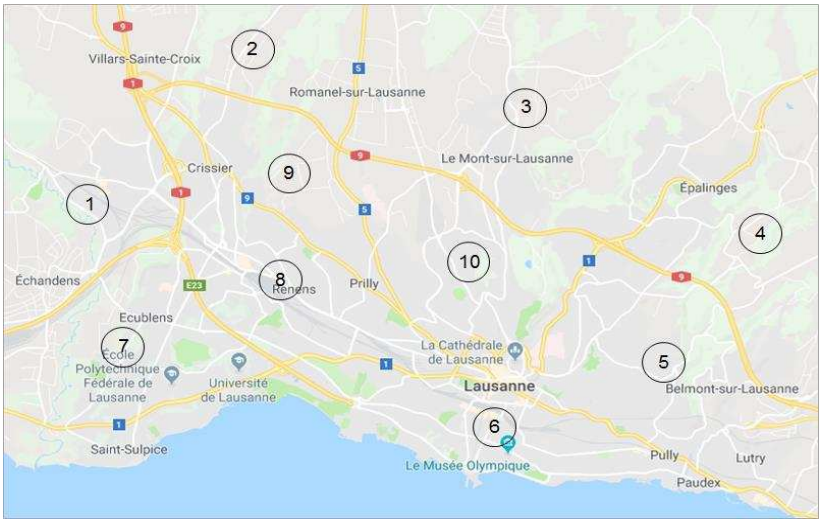


Figure 2: Sample zones for travel time analysis (Google Maps, 2018)

For each analysis period, we computed the time-period specific congestion factors by first establishing the interval with the shortest average travel time, and calculating the ratios of the travel times for each interval versus the minimum travel time for the analysis period as illustrated in Figure 3.

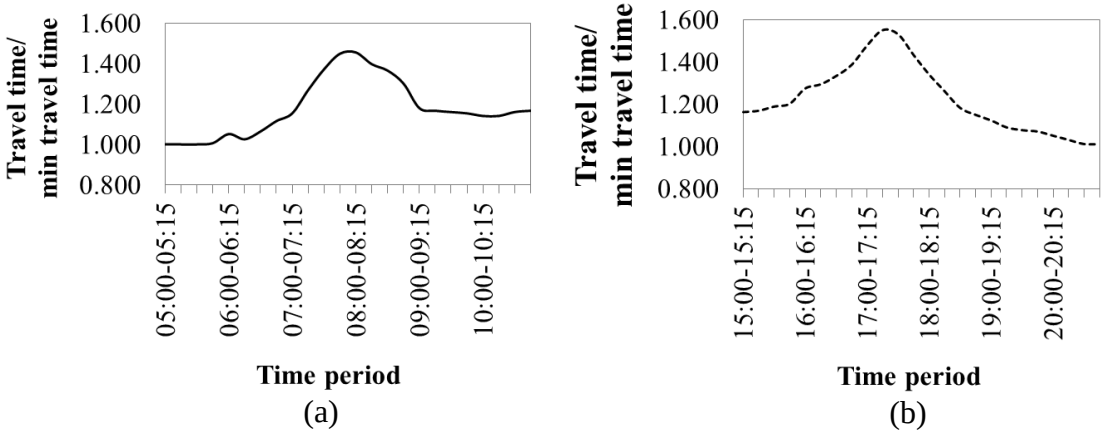


Figure 3: Travel time variation (a) Morning peak, (b) Evening peak

Given the typical ratios for each time interval, we used the observed average travel times for each user (i.e. those based on the cleaned GSM and GPS data) to estimate the minimum travel times for each of their trip chains, and applied the appropriate time-period specific congestion factors to impute the travel times for each time interval. We used the imputed travel times for both the chosen and the unchosen alternatives as this mitigates possible endogeneity bias which could arise from interrelationships with other underlying factors (Calastri et al., 2017, Sanko et al., 2014).

3.4 Comparison of the GPS and the GSM processed data

Due to differences in the temporal coverage of the GSM versus the GPS data (see Section 3.2.2), the inferred home and work locations were different for some users as illustrated in Figure 4. As observed, most of the inferred GPS and GSM home/work locations are within 5 kilometres of each other, an indicator that most belong to the same cell. However, we also have scenarios where the inferred dwell locations for the same day are over 40 kilometres apart. In such cases, GSM data, which has a higher temporal coverage, is expected to be more reliable. However, since our focus is on parameter level comparison in the model development stage, we retain the dwell locations for each data type as extracted.

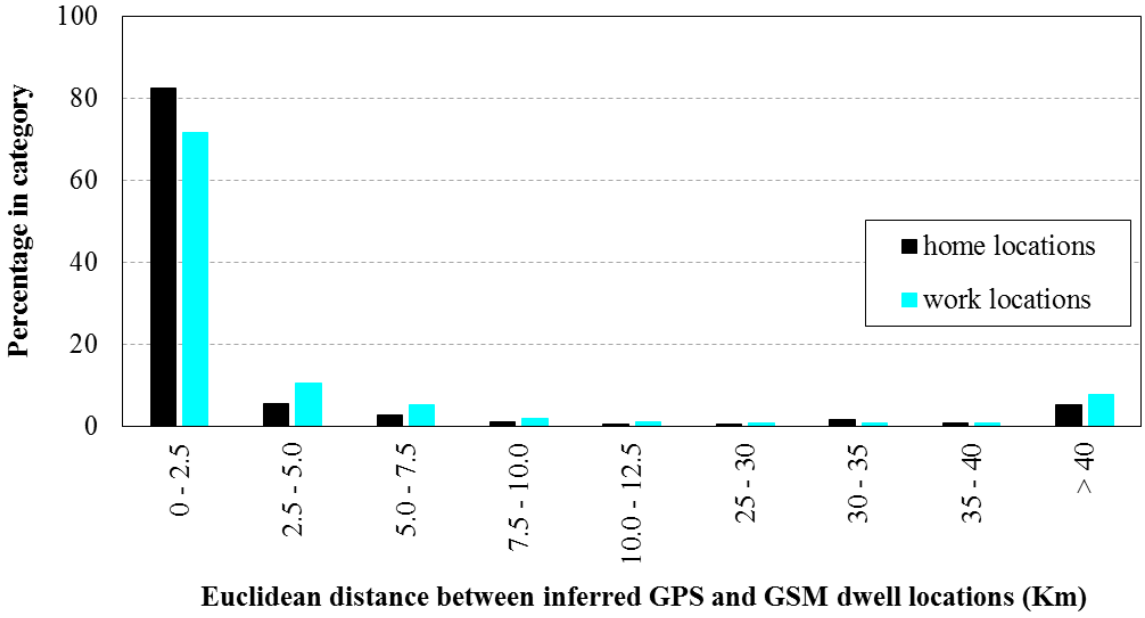


Figure 4: Differences between the GPS and GSM inferred home/ work locations

Furthermore, it is observed that the extracted departure time distributions are different across the two datasets, with the home-to-work commute having more pronounced differences (see Figure

5). This is because the time gaps in the night GPS traces are more than those in the daytime GPS traces, potentially because the users disable their GPS receivers more at night. This is probably the reason why the peak period for the home-to-work commute is not clearly defined. On the other hand, the peak periods for the GSM data are clearly observed for both the home-to-work, and the work-to-home commute. This is in line with the expected behaviour of full-time workers.

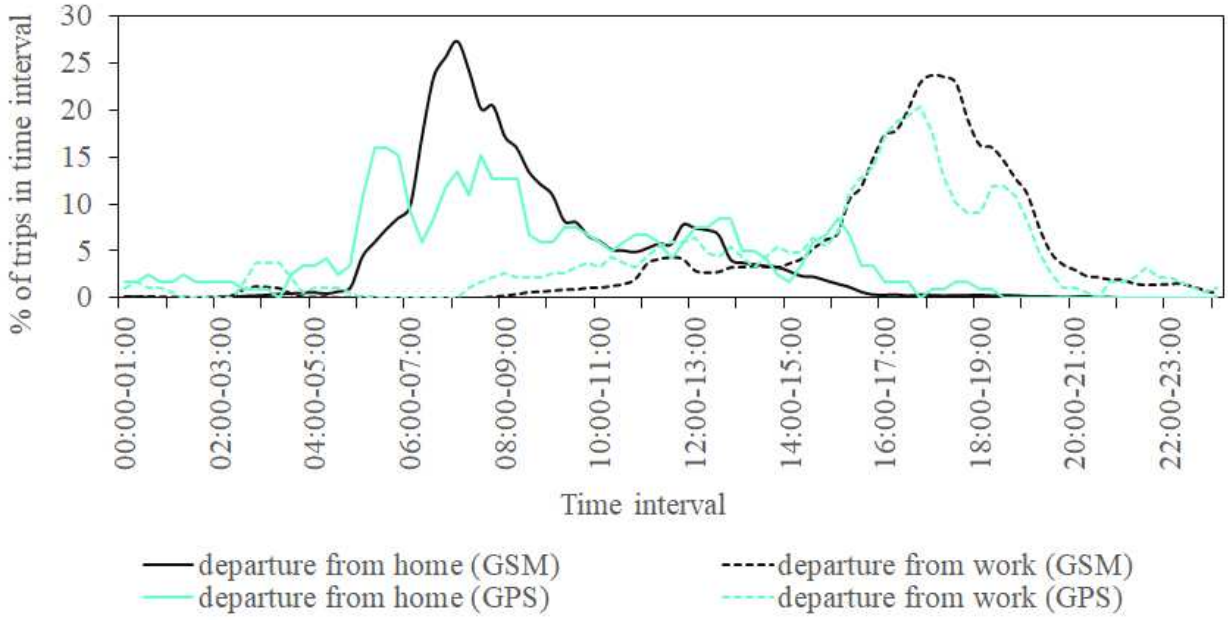


Figure 5: Commuter trip frequency distribution

4 Modelling framework

Our analysis is based on the random utility framework (Marschak, 1960), and theoretical insights from the scheduling model by Small (1982). Let U_{ntk} be the utility for individual n derived from departing in time period t in choice situation k. This can be expressed as;

$$U_{ntk} = V_{ntk} + \varepsilon_{ntk} \quad (1)$$

Where V_{ntk} and ε_{ntk} are the systematic and the random parts of utility, respectively. Based on scheduling theory, V_{ntk} is usually expressed as a function of the travel time, the corresponding schedule delays, and any other key attributes as follows;

$$V_{ntk} = \beta_{t-time}T_{ntk} + \beta_{l-dummy}L_{ntk} + E_{ntk}(\beta_{e-time}SDE_{ntk}) + L_{ntk}(\beta_{l-time}SDL_{ntk}) + \cdots \quad (2)$$

Where T_{ntk} is the travel time associated with time period t in choice situation k for individual n. E_{ntk}, L_{ntk}, SDE_{ntk}, and SDL_{ntk} are the earliness dummy, the lateness dummy, the amount of earliness, and the amount of lateness associated with time period t for individual n in choice situation k. The βs are the corresponding model parameters to be estimated. It may be noted that the earliness terms (E_{ntk} and SDE_{ntk}) and the lateness terms (L_{ntk} and SDL_{ntk}) are mutually exclusive.

Estimating the amount of earliness or lateness associated with a particular departure time period requires information on the desired times of travel, which is not available in this case as we are relying on passively collected datasets. However from a theoretical perspective, we assume that every individual makes efforts to depart at his/her desired time to minimise scheduled delay. By re-writing Equation (2), this may be expressed as follows;

$$V_{ntk} = \beta_{t-time} T_{ntk} + \beta_{l-dummy} L_{ntk} + E_{ntk}(\beta_{e-time}[PDT_n - D_t]) + L_{ntk}(\beta_{l-time}[D_t - PDT_n]) + \cdots \quad (3)$$

Where PDT_n is the preferred departure time for individual n, and D_t is the midpoint of departure time interval t in terms of the hours since midnight (e.g. for departure time interval 8:00 am – 8:15 am, D_t is 8.125 hours, which corresponds to 8:07:30 am).

In the absence of the preferred departure times of the users, it is reasonable to assume that these vary randomly across the individuals following a certain statistical distribution. The objective we are trying to pursue is to estimate the mean and standard deviation of this statistical distribution. The use of statistical distributions helps us to avoid the assumption that the preferred departure times follow the same trend as the observed departure times as used in Koppelman et al. (2008).

However, estimating the above specification with different schedule delay early and lateness terms presents serious identification and optimisation issues. This is because the earliness and lateness dummies depend on the preferred departure time, which is also being estimated at the same time. As the optimiser tries to find the preferred departure time, the dummies keep alternating between 0 and 1, thereby resulting in a function that is not continuously differentiable. This prompts us to deviate from Small's typical model formulation (e.g. Cosslett, 1977, Small, 1982) by investigating alternative schedule delay functions that are first of all behaviourally intuitive, and continuously differentiable.

From a behavioural perspective, the schedule delay function needs to reflect reductions in the schedule disutility as the observed departure times approach the preferred departure times (from both the earliness and lateness sides) and must peak at points where the delay is zero. Furthermore, the function needs to have an indifference region around the preferred departure time to reflect the fact that the rate of increase in the schedule disutility is small around the preferred departure times, and increases as the observed departure times spread further away from the preferred departure time. After testing alternative functional forms (e.g. the logistic and the parabolic functions), we selected the parabolic function, which gave consistent and intuitive results. The systematic utility is now expressed as follows;

$$V_{ntk} = \beta_{t-time} T_{ntk} + \alpha(PDT_n - D_t)^2 + \cdots \quad (4)$$

Where α is a parameter to be estimated, representing the sensitivity to delay. For the schedule function above to be behaviourally intuitive, the parameter α is expected to have a negative sign. A potential issue with this function is that it does not capture the damping effect of schedule delay on marginal disutility as argued in previous studies (e.g. Koppelman et al., 2008), which calls for further research to address this issue.

Assuming the βs, the αs and PDT_n (for individual n) are known, and the random part of utility ε_{ntk} is independently and identically distributed across the choice situations, the alternatives, and the individuals, the departure time choice probability for individual n can be estimated using the multinomial logit (MNL) model (see McFadden, 1974 for details). However in this case, we have several choice situations for the same individual across different days, thus, we need to capture the panel effect while calculating the choice probabilities.

Let $P_{n,k}(t|\beta, \alpha, PDT_n)$ denote the logit probability that individual n chooses departure time period t in choice situation k, conditional on β, α and PDT_n. Furthermore, let $\hat{t}_{n,k}$ be the departure time chosen by individual n in choice situation k, such that $P_{n,k}(\hat{t}_{n,k}|\beta, \alpha, PDT_n)$ gives the logit

probability of the observed choice for individual n in choice situation k, conditional on β, α and PDT_n. The logit probability of individual n's observed sequence of choices is;

$$P_n(\beta, \alpha, PDT_n) = \prod_{k=1}^{K} P_{n,k}(\hat{t}_{n,k}|\beta, \alpha, PDT_n) = \prod_{k=1}^{K} \frac{\exp(V_{n\hat{t}k}|\beta, \alpha, PDT_n)}{\sum_{t^* \in C_n} \exp(V_{ntk^*}|\beta, \alpha, PDT_n)} \quad (5)$$

Where C_n is the choice set. It is important to note that for users with more than one trip chain, we compare the attributes of the same trip chain across the different time periods while computing the choice probabilities. That is, each trip chain represents a different choice scenario.

However as earlier mentioned, the preferred departure times PDT_n are not observed, and are assumed to vary randomly across individuals. Suppose PDT_n is independently and identically distributed over the individuals with density $f(PDT|\Omega)$, where Ω is a vector of the parameters of this distribution, such as the mean and standard deviation, the mixed logit probability is given by;

$$P_n(\beta, \alpha, \Omega) = \int_{PDT} \left[\prod_{k=1}^{K} P_{n,k}(\hat{t}_{n,k}|\beta, \alpha, PDT_n) \right] f(PDT|\Omega)\, dPDT \quad (6)$$

The integration over the density of PDT is done over all the individual's choices combined, since the same PDT applies to all the choice situations. The log-likelihood (LL) function for the observed choices is;

$$LL(\beta, \alpha, \Omega) = \sum_{n=1}^{N} \ln\left(\int_{PDT} \left[\prod_{k=1}^{K} P_{n,k}(\hat{t}_{n,k}|\beta, \alpha, PDT_n) \right] f(PDT|\Omega)\, dPDT \right) \quad (7)$$

An important consideration is the choice of distribution to be used. Due to our limited knowledge of the individuals' preferences, coupled with the fact that we have not conducted any surveys to determine the distribution of the preferred departure times, we assume a truncated normal distribution bounded between the limits of the analysis period (i.e. the morning or evening peak periods). Since the integral in Equation 7 has no closed form, it is estimated using simulation methods. The simulated log-likelihood (SLL) is expressed as follows;

$$SLL(\beta, \alpha, \Omega) = \sum_{n=1}^{N} \ln\left(\frac{1}{R} \sum_{r=1}^{R} \left[\prod_{k=1}^{K} P_{n,k}(\hat{t}_{n,k}|\beta, \alpha, PDT_n) \right] \right) \quad (8)$$

The PDT distribution parameters (i.e. the mean and standard deviation) are estimated alongside the other model parameters by maximising the simulated log-likelihood using 300 Halton draws per user (Bhat, 2001). During parameter estimation, there may be a possibility of confounding between random PDT and random schedule delay sensitivity α, however, this is mitigated by applying the same parameter α to ${PDT_n}^2$, ${D_t}^2$, and $-2PDT_nD_t$ (see Equation 4).

5 Model results

This section discusses the process of variable specification, the estimation results, as well as the policy insights derived from the estimation results.

5.1 Variable specification

The variables available for possible inclusion in the departure time utility equation are; travel time, latent schedule delay, trip chain characteristics, and user demographics. However, each of these variables was defined in a particular way for different reasons as explained in the subsequent paragraphs.

For travel time, we tested the logarithmic specification to allow for damping effects (Daly, 2010) and found no gains in model fit compared to the linear specification. This could be attributed to the small ranges of travel time across the alternatives of each user. Therefore, we adopted a linear specification.

|The schedule delay function was entered into the model as specified in Equation 4. We investigated the possibility of different PDT distribution parameters for different demographic groups and could not obtain significant gains in model fit for either dataset.

The trip chain characteristics were incorporated into the model using the number of intermediate stops, and time-period specific parameters were specified to capture the differential impact on utility across the time periods. It may be noted that the duration at the intermediate stops is already incorporated into the travel time.

A number of interactions of the schedule delay and the travel time parameters with the user demographics were tested. For the GPS data, we could not obtain intuitive results for all the interactions tested due to the small sample size per demographic group in the final sample, so we specified generic parameters. On the other hand, for the GSM data, we successfully interacted the travel time and the schedule delay parameters with age-group alone and age-group by gender, respectively. The final systematic utility specifications for the GSM and the GPS data are given by Equations (9) and (10), respectively;

$$V_{ntk} = \beta_{time-age} T_{ntk} + \alpha_{del\_age\_gender} {SD_{nt}}^2 + \beta_{stops\_t} N_{stops} \quad (9)$$

$$V_{ntk} = \beta_{time} T_{ntk} + \alpha_{del} {SD_{nt}}^2 + \beta_{stops\_t} N_{stops} \quad (10)$$

Where $SD_{nt} = (PDT_n - D_t)$, and N_{stops} is the number of intermediate stops in the trip chain. The βs and αs are the model parameters to be estimated.

5.2 Estimation results

We present the estimation results for the home-to-work commute and the work-to-home commute models for both the GSM and the GPS data in Table 3 for comparison purposes. It may be noted that there are differences in the sample sizes of the different models and/or commute directions when compared to the values reported in Table 1, which is attributed to the data cleaning rules we applied to exclude trips which are very short, very long trips and/or made by non-motorised modes (see Appendix B for details). As observed, most of the parameter estimates are statistically significant at the 95% level of confidence. In the subsequent sections, we discuss each aspect of the results in details.

5.2.1 Distribution parameters for departure time distribution

We specified a truncated normal distribution for the preferred departure time (for reasons explained in the paragraph after Equation 7). For this distribution, the estimated mean and standard deviation are those of the underlying normal distribution. To calculate the true means and standard deviations, the estimated parameters were adjusted as follows;

$$\mu = \hat{\mu} + \left[\frac{\phi(\mathrm{A}) - \phi(\mathrm{B})}{\Phi(\mathrm{B}) - \Phi(\mathrm{A})}\right]\hat{\sigma} \quad (9)$$

$$\sigma = \hat{\sigma}\left[1 + \frac{\mathrm{A}\phi(\mathrm{A}) - \mathrm{B}\phi\mathrm{B}}{\Phi(\mathrm{B}) - \Phi(\mathrm{A})} - \left(\frac{\phi(\mathrm{A}) - \phi(\mathrm{B})}{\Phi(\mathrm{B}) - \Phi(\mathrm{A})}\right)^2\right]^{1/2} \quad (10)$$

Where, $\hat{\mu}$ and $\hat{\sigma}$ are the estimated mean and standard deviation respectively of the underlying normal distribution, μ and σ are the true mean and standard deviation respectively of the truncated distribution. $\mathrm{A} = (a - \hat{\mu})/\hat{\sigma}$, $\mathrm{B} = (b - \hat{\mu})/\hat{\sigma}$, where a and b are the lower and upper bounds respectively of the truncated distribution. For the home-to-work commute, these are set to 5 and 11, respectively, while for the work to home commute, these are 15 and 21 respectively.

From Table 3, it is observed that the mean preferred departure times for the home-to-work commute are 7.9742 and 8.0105 (approximately 8:00 am), while those for the work-to-home commute are 17.7240 (approximately 05:45 pm) and 17.2903 (approximately 05:15 pm) in the GSM and the GPS models, respectively. Although flexible working time is not unusual in Switzerland, normal business hours generally start between 08:00 am and 08:30 am and end between 05:00 pm and 06:30 pm (Switzerland Tourism, 2018), which is consistent with our findings.

Furthermore, it is observed that the corresponding standard deviations are slightly higher for the home-to-work commute when compared to the work-to-home commute. The lower amount of variation during the work-to-home commute is probably the reason behind the higher evening traffic congestion in Lausanne and other major Swiss cities (TomTom, 2016).

5.2.2 Sensitivity to schedule delay

Generally, individuals prefer to depart at particular times due to certain constraints at both the origin and the destination. Thus, any deviations from the desired times of travel are expected to cause disutility, hence the negative parameter signs for the schedule delay terms reported in Table 3. For GSM data where we have different schedule delay parameters for different demographic groups, where we note that female workers are more sensitive to shifting departure time compared to male workers in the same age-group. This is the case during both the home-to-work and the work-to-home commute. The higher sensitivity of female workers is potentially attributed to the strictness in their schedule as a result of the need to balance family and professional life in the face of common views on traditional gender roles in Switzerland (Nguyen, 2018, The Economist, 2018).

Furthermore, it is observed that younger workers are more sensitive to schedule delay than older workers of the same gender during the home-to-work commute. This is expected as younger workers are more junior and typically have less flexibility (i.e. expected to report on time). However, the situation is different for the work-to-home commute. Here, it is observed that older female workers are more sensitive than young female workers. This again could be attributed to the levels of responsibility at home as older female workers are more likely to have families already. However, the lower sensitivity of older male workers in comparison with younger male workers is an interesting observation that needs to be investigated further.

For the GSM data, we also observe that the sensitivity to schedule delay is generally higher during the home-to-work commute when compared to the work-to-home commute. This is expected as late arrival on the home-to-work commute probably has more serious consequences than on the work-to-home commute. However, this was not captured in the GPS data due to differences in the sample composition resulted by the big time gaps in the GPS data.

Table 3: Model estimation results

Variable	Home-to-work commute				Work-to-home commute			
	GSM data		GPS data		GSM data		GPS data	
	Parameter	t-stat	Parameter	t-stat	Parameter	t-stat	Parameter	t-stat
Travel time (hours)								
Workers < 28 years	-2.9700	-2.37	-0.9486	-1.05	-1.0674	-2.59	-0.2340	-0.32
Workers >= 28 years	-1.3266	-1.35			-1.0314	-2.15		
Schedule delay term (hours^2)								
Female workers < 28 years	-0.7682	-6.51	-0.2560	-1.88	-0.4412	-5.70	-0.3093	-6.13
Female workers >= 28 years	-0.6441	-5.34			-0.4785	-7.12		
Male workers < 28 years	-0.6841	-3.49			-0.4148	-4.46		
Male workers >= 28 years	-0.5468	-6.20			-0.3749	-6.78		
Preferred departure time distribution parameters								
$\hat{\mu}$	7.9652	67.55	8.0105	18.76	17.7240	170.59	17.2903	112.59
μ	7.9661		8.0081		17.7244		17.2903	
$\hat{\sigma}$	1.0029	7.85	1.5077	1.99	0.7465	11.98	0.4729	4.40
σ	0.9783		1.7503		0.5562		0.2236	
Number of stops (Time period specific parameters)*								
δ_1	-2.2939	-2.24	-1.1547	-0.43	1.8821	1.97	-2.4504	-2.17
δ_2	-0.8544	-1.80	-1.3607	-0.53	2.0265	2.18	-2.9766	-2.63
δ_3	-1.3742	-2.05	-1.5777	-0.66	1.8313	2.00	-3.4635	-3.05
δ_4	-0.3398	-0.81	16.7402	6.94	1.2878	1.39	11.7430	7.91
δ_5	-1.1779	-2.68	-1.9351	-0.91	1.7281	1.87	11.5697	8.53
δ_6	-1.1063	-2.76	-2.1946	-1.13	1.3622	1.44	12.1103	8.28
δ_7	-1.2352	-3.28	-2.3215	-1.26	1.8133	1.95	-4.7535	-4.36
δ_8	-1.0663	-2.88	-2.3982	-1.43	1.5522	1.68	-4.8664	-4.55
δ_9	-1.0169	-3.17	15.1030	8.56	1.6290	1.77	11.1953	8.50
δ_{10}	-0.6124	-1.70	15.1026	11.49	1.4974	1.61	-4.7149	-4.67

Table 3 cont'd

Variable	Home-to-work commute				Work-to-home commute			
	GSM data		GPS data		GSM data		GPS data	
	Parameter	t-stat	Parameter	t-stat	Parameter	t-stat	Parameter	t-stat
δ_{11}	-1.2679	-3.30	-2.2488	-1.79	1.8324	1.99	-4.7034	-4.72
δ_{12}	-1.1857	-3.69	-2.1632	-1.87	1.3271	1.45	11.2286	8.07
δ_{13}	-0.9265	-2.95	-2.1864	-2.09	1.1121	1.19	-4.6527	-4.75
δ_{14}	-1.1458	-2.84	-2.3160	-2.67	0.8245	0.88	-4.4459	-4.61
δ_{15}	-1.1957	-2.61	-2.3777	-3.25	0.8900	0.94	-4.1544	-4.35
δ_{16}	-1.7638	-4.76	-2.4956	-4.26	0.3669	0.38	-3.7369	-4.05
δ_{17}	-1.1030	-2.80	-2.7512	-6.41	1.1553	1.26	-3.2570	-3.67
δ_{18}	-2.8825	-4.97	-2.7162	-8.04	1.2105	1.30	-2.7495	-3.25
δ_{19}	-1.0044	-3.49	-2.6429	-10.06	0.9961	1.03	-2.2151	-2.77
δ_{20}	-1.2905	-4.90	-2.5495	-13.12	0.9618	1.01	-1.7011	-2.28
δ_{21}	-1.2117	-4.28	-2.4537	-15.91	1.1151	1.21	-1.2485	-1.80
δ_{22}	-0.5033	-1.98	-2.3064	-17.13	0.7528	0.79	-0.8655	-1.35
δ_{23}	-0.6143	-2.71	-2.0963	-16.51	1.6520	1.89	-0.5606	-0.96
Measures of fit in estimation								
Number of observations	2043		69		2668		112	
Number of decision makers	78		29		78		35	
LL(C)	-6492.76		-219.29		-8479.05		-355.94	
LL(F)	-5575.19		-203.50		-7644.19		-323.95	
Number of parameters	31		27		31		27	
ρ^2_{adj} w.r.t LL(C)	0.1365		-0.0511		0.0948		0.0140	
LR w.r.t LL(C)	1835.15		31.58		1669.71		63.99	
p-value of LR	0.0000		0.2480		0.0000		0.0001	

* 1 refers to the first 15-minute interval in the period of analysis (i.e. 5:00 to 5:15 for morning home-to-work commute, and 15:00 to 15:15 for the evening work-to-home commute). The rest of the numbers refer to the subsequent 15-minute intervals in ascending order.

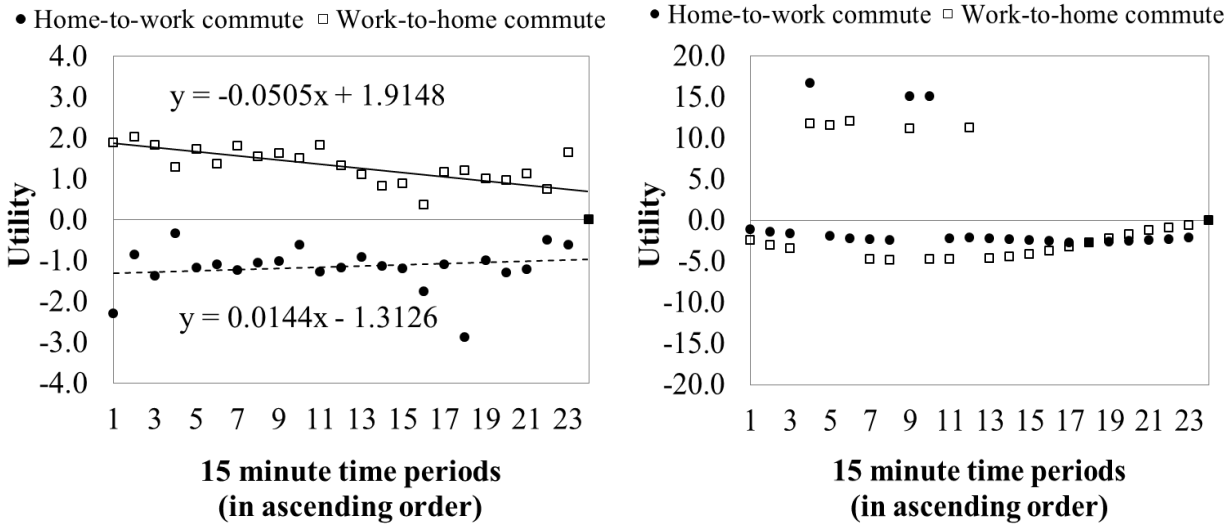
5.2.3 Sensitivity to travel time

From Table 3, the parameter signs for the travel time variable are negative in both the home-to-work and the work-to-home commute models, which is consistent with a priori expectations. In general, time periods with higher travel times are less attractive, and prompt individuals to choose earlier or later time periods at the expense of increasing the schedule delays.

Keeping all other things constant, it is observed that the sensitivity to travel time during the home-to-work commute is generally higher than that during the work-to-home commute in both the GSM and the GPS models. This implies that people are less willing to spend longer times in traffic during the home-to-work commute as opposed to the reverse direction, which could be attributed to the higher stakes attached to the home-to-work leg.

5.2.4 Time-period specific parameters related to the number of stops

Another important issue worth highlighting concerns the time-period specific parameters related to the number of stops. For easy parameter identification, we normalised to zero the effect linked to the last departure time interval of each analysis period (i.e. 10:45 am to 11:00 am for the morning home-to-work commute and 08:45 pm to 09:00 pm for the evening work-to-home commute). Thus, the reported parameters represent the differential impact on utility with respect to the reference time periods and can either be positive or negative. These are reproduced in Figure 6 for easy visualisation.



(a) GSM data (b) GPS data

Figure 6: Time period specific parameters for the number of stops

The interpretation of these parameters is however difficult given that they probably incorporate some other unobserved factors associated with the different time periods. Nevertheless, for GSM data (Figure 6a), the trend of the parameters in the work-to-home model shows that if some someone is to make a trip with more stops, they will find the earlier departure time periods more suitable, which is reasonable. However, the trend is different in the home-to-work model, where it is observed that the later departure time periods would be more suitable. This could be attributed to the less traffic during the inter-peak period and the opening times at the various stop locations. For GPS data (Figure 6b), there is no clear trend in the parameters for both commute directions, which is attributed to the weaknesses in the data as mentioned earlier which contributes to higher levels of noise.

5.2.5 Overall model performance

From Table 3, it is observed that the adjusted-rho square values of the GPS models are far less compared to those of the GSM models. While acknowledging that the models cannot be directly

compared due to differences in the sample compositions, the overall poor performance of the GPS models is largely attributed to the small GPS sample sizes.

5.3 Policy insights

Travel time is usually at its worst when the schedule delay for most individuals is at the minimum. Therefore, people are generally faced with a trade-off between travel time and schedule delay when choosing the most appropriate departure time periods. Thus, to gain better insights, it is critical to analyse the sensitivities to schedule delay versus travel time to obtain the values attached to schedule delay.

The time valuation of schedule delay ($TVSD$) represents the amount of delay an individual is willing to experience for a unit reduction in travel time by changing his/her departure schedule. This unitless metric is calculated as the ratio of the partial derivatives of the systematic utility with respect to schedule delay and travel time as follows;

- For the GSM data

$$TVSD_{age\_gender} = \frac{\partial V_{ntk}/\partial SD_{nt}}{\partial V_{ntk}/\partial T_{ntk}} = \frac{\alpha_{del\_age\_gender} * 2SD_{nt}}{\beta_{time-age}} \quad (11)$$

- For the GPS data

$$TVSD = \frac{\partial V_{ntk}/\partial SD_{nt}}{\partial V_{ntk}/\partial T_{ntk}} = \frac{\alpha_{del} * 2SD_{nt}}{\beta_{time}} \quad (12)$$

Since we used square-transformations, the time valuation of schedule delay depends on the amount of earliness/lateness of the individual as shown in Equations (11) and (12). We therefore used the estimation data (including the normal draws) to calculate the average values across individuals, which we report in Table 4.

Table 4: Time valuation of schedule delay

Commute direction	GSM data					GPS data
	Female worker < 28 years	Female worker >= 28 years	Male worker < 28 years	Male worker >= 28 years	Weighted mean	Generic
Home to work	1.2196	2.5116	0.9995	2.2031	2.0167	2.3435
Work to home	1.6729	1.8290	1.5570	1.6747	1.7025	4.6266
Weighted mean	1.4824	2.1352	1.2964	1.8893	1.8388	3.4850

To assess how realistic our estimates are, we compared them with the typical averages for Europe reported in the meta study conducted by Wardman et al. (2012), summarised in Table 5.

Table 5: Time valuations from other sources

Average values for Europe (Wardman et al., 2012)	GSM data	GPS data
0.81 to 1.71 (Schedule delay early to schedule delay late)	1.84 (From Table 4)	3.49 (From Table 4)

A comparison of the GSM and the GPS valuation estimates shows that those of the former are closer to the average range for Europe. As mentioned earlier, this is likely due to the weaknesses

in the specific GPS dataset that has been used (collected in 2010), and the findings cannot be generalised. It is expected that more recent datasets could lead to different findings as GPS technology has greatly improved over the last few years. Although we do not have any ground truth data from the study area for direct validation, the above results give some reassurance that GSM data can be used for understanding departure time choice. Indeed, these results show that mobile phone data can be feasibly used to analyse time-of-travel choices despite the lack of information on the preferred departure/arrival times. The availability of demographic data offers additional benefits in terms of explaining the differences in sensitivity across individuals.

6 Summary and conclusions

An initial comparison of the GSM and the GPS datasets collected in parallel for the same users showed that the amount of time gaps in the GPS data (in this specific case) were very substantial. This was potentially due to technical issues such as signal losses in urban environments and large public transport vehicles, as well as the users turning off their GPS apps due to battery issues. On the other hand, the amount of time gaps in the GSM data were not as pronounced. Due to these challenges, the GPS data could not capture most of the trips made, and the extracted sample size was very small compared to that extracted from the GSM data. Consequently, the models based on GPS data were not as reliable in this case as those based on GSM data. An important point to note is that advances in smartphone GPS technology have occurred since 2010, and it is likely that some of the technical issues encountered in this study have been resolved. Therefore, it is would be important to re-evaluate the feasibility of GPS data using more recent datasets. Nevertheless, this paper has successfully demonstrated the potential of GSM data as an alternative source of information for departure time choice modelling.

An important aspect we recognise is the fact that the preferred departure/arrival times of the users are not known due to the anonymous nature of mobile phone data and yet this is an important aspect of departure time choice models. We propose a modelling framework in which the unobserved preferred departure times are assumed to vary across the users following a particular distribution, and estimate the distribution parameters (i.e. the mean and standard deviation) using the mixed logit framework. This approach allows us to simultaneously estimate the distribution parameters alongside the sensitivities to schedule delay, which are found to be intuitive. Although we have applied the proposed approach in the context of anonymous mobile phone data, it can be applied to model departure time choice using traditional RP datasets where the desired times of travel are sometimes not known.

Furthermore, since mobile phone data only reports the revealed departure time preferences of the users, the modeller does not know the other alternatives that were considered while making these choices. We make a general assumption that all the possible departure time intervals in the analysis period (i.e. morning or evening peak period) were considered. However, the attributes for some of the departure time intervals may not be observed for some users if they rarely travel during those periods. Consequently, we propose a practical approach for imputing the travel times associated with different departure time intervals using time-period specific congestion factors derived from Google maps for a sample of O-D pairs. This approach is particularly beneficial in the absence of Google Distance Matrix API services for duration in traffic.

The model results reflect the generally expected behaviour. When these results were applied to estimate the time valuations of schedule delay, we obtained reasonable estimates from the models based on GSM data in comparison with those from the literature, which are largely based on stated choice data. We conclude that the results of this study serve as a proof-of-concept that mobile phone network records are a promising source of information for transport modelling and policy analysis, especially in contexts where traditional data sources are unavailable. This is the case in most developing countries with limited budgets for transport studies.

Acknowledgements

The research in this paper used the MDC Database made available by Idiap Research Institute, Switzerland and owned by Nokia. We would like to thank the Economic and Social Research Council (ESRC) of the UK and the Institute for Transport Studies, University of Leeds for funding this research. Professor Stephane Hess' time is supported by the European Research Council through the consolidator grant 615596-DECISIONS.

References

Abkowitz, M. D. 1981. An analysis of the commuter departure time decision. Transportation, 10, 283-297.

Arellana, J., Daly, A., Hess, S., de Dios Ortúzar, J. and Rizzi, L.I., 2012. Development of Surveys for Study of Departure Time Choice: Two-Stage Approach to Efficient Design. Transportation Research Record, 2303(1), pp.9-18.

Arellana, J., Ortúzar, J.D.D. and Rizzi, L.I., 2013. Survey data to model time-of-day choice: methodology and findings. In Transport Survey Methods: Best Practice for Decision Making (pp. 479-506). Emerald Group Publishing Limited.

Bates, J., Shepherd, N., Roberts, M., Van Der Hoorn, A. & Pol, H. A model of departure time choice in the presence of road pricing surcharges. 18th PTRC Summer Annual Meeting, 1990 University of Sussex, United Kingdom.

Ben-Akiva, M. & Abou-Zeid, M. 2013. Methodological issues in modelling time-of-travel preferences. Transportmetrica A: Transport Science, 9, 846-859.

Bernardi, S. & Rupi, F. 2015. An analysis of bicycle travel speed and disturbances on off-street and on-street facilities. Transportation Research Procedia, 5, 82-94.

Bhat, C. R. 1998a. Accommodating flexible substitution patterns in multi-dimensional choice modeling: formulation and application to travel mode and departure time choice. Transportation Research Part B: Methodological, 32, 455-466.

Bhat, C. R. 1998b. Analysis of travel mode and departure time choice for urban shopping trips. Transportation Research Part B: Methodological, 32, 361-371.

Bhat, C. R. 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transportation Research Part B: Methodological, 35, 677-693.

Brey, R. & Walker, J. L. 2011. Latent temporal preferences: An application to airline travel. Transportation Research Part A: Policy and Practice, 45, 880-895.

Bwambale, A., Choudhury, C. F. & Hess, S. 2017. Modelling trip generation using mobile phone data: A latent demographics approach. Journal of Transport Geography.

Calastri, C., Hess, S., Choudhury, C., Daly, A. & Gabrielli, L. 2017. Mode choice with latent availability and consideration: theory and a case study. Transportation Research Part B: Methodological.

Chen, C., Gong, H., Lawson, C. & Bialostozky, E. 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. Transportation Research Part A: Policy and Practice, 44, 830-840.

Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiretta, S. R. & González, M. C. Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities. Transportation Research Board 94th Annual Meeting, 2015.

Cosslett, S. 1977. The trip-timing decision for travel to work by automobile: demand model estimation and validation. The Urban Travel Demand Forecasting Project Phase I Final Report, Volume 5. Institute for Transportation Studies, University of California, Berkeley.

Daly, A. 2010. Cost damping in travel demand models: Report of a study for the

Department for Transport. United Kingdom: RAND Corporation.

Daly, A., Gunn, H., Hungerink, G., Kroes, E. & Mijjer, P. Peak-period proportions in large-scale modelling. 18th PTRC Summer Annual Meeting, 1990 University of Sussex, United Kingdom.

De Groote, A. 2005. GSM Positioning Control. University of Fribourg, Switzerland, 13.

De Jong, G., Daly, A., Pieters, M., Vellay, C., Bradley, M. & Hofman, F. 2003. A model for time of day and mode choice using error components logit. Transportation Research Part E: Logistics and Transportation Review, 39, 245-268.

Doyle, J., Hung, P., Kelly, D., Mcloone, S. F. & Farrell, R. 2011. Utilising mobile phone billing records for travel mode discovery.

Ettema, D., Tamminga, G., Timmermans, H. and Arentze, T., 2005. A micro-simulation model system of departure time using a perception updating model under travel time uncertainty. Transportation Research Part A: Policy and Practice, 39(4), pp.325-344.

Everitt, B. S., Landau, S., Leese, M. & Stahl, D. 2011. Hierarchical clustering. Cluster Analysis, 5th Edition, 71-110.

Gong, H., Chen, C., Bialostozky, E. & Lawson, C. T. 2012. A GPS/GIS method for travel mode detection in New York City. Computers, Environment and Urban Systems, 36, 131-139.

Google Developers. 2018. Distance Matrix Service [Online]. Google. Available: https://developers.google.com/maps/documentation/javascript/distancematrix [Accessed 29 June 2018].

Google Maps. 2018. Lausanne, Switzerland [Online]. Google. Available: https://www.google.co.uk/maps/place/Lausanne,+Switzerland/@46.5284586,6.5824556,12z/data=!4m5!3m4!1s0x478c293ecd89a7e5:0xeb173fc9cae2ee5e!8m2!3d46.5196535!4d6.6322734 [Accessed 03 May 2018].

GSM Association. 2017. The Mobile Economy 2017 [Online]. Available: https://www.gsmaintelligence.com/research/?file=9e927fd6896724e7b26f33f61db5b9d5&download [Accessed 04 November 2017].

Hess, S., Daly, A., Rohr, C. & Hyman, G. 2007a. On the development of time period and mode choice models for use in large scale modelling forecasting systems. Transportation Research Part A: Policy and Practice, 41, 802-826.

Hess, S., Polak, J. W. & Bierlaire, M. Functional approximations to alternative-specific constants in time-period choice-modelling. Transportation and Traffic Theory: Flow, Dynamics and Human Interaction, Proceedings of the 16th International Symposium on Transportation and Traffic Theory, 2005. 545-564.

Hess, S., Polak, J. W., Daly, A. & Hyman, G. 2007b. Flexible substitution patterns in models of mode and time of day choice: new evidence from the UK and the Netherlands. Transportation, 34, 213-238.

Hydén, C., Nilsson, A. & Risser, R. 1999. How to enhance WALking and CYcliNG instead of shorter car trips and to make these modes safer. Public. Deliverable D6. Walcyng Contract No: UR-96-SC. 099. Department of Traffic Planning and Engineering, University of Lund, Sweden & FACTUM Chaloupka, Praschl & Risser OHG, Vienna, Austria.

Iqbal, M. S., Choudhury, C. F., Wang, P. & González, M. C. 2014. Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies, 40, 63-74.

Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A. & Willinger, W. Human mobility modeling at metropolitan scales. Proceedings of the 10th international conference on Mobile systems, applications, and services, 2012. Acm, 239-252.

Jiang, S., Fiore, G. A., Yang, Y., Ferreira Jr, J., Frazzoli, E. & González, M. C. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, 2013. ACM, 2.

Kiukkonen, N., Blom, J., Dousse, O., Gatica-Perez, D. & Laurila, J. 2010. Towards rich mobile phone datasets: Lausanne data collection campaign. Proc. ICPS, Berlin.

Koppelman, F. S., Coldren, G. M. & Parker, R. A. 2008. Schedule delay impacts on air-travel itinerary demand. Transportation Research Part B: Methodological, 42, 263-273.

Kristoffersson, I. & Engelson, L. 2018. Estimating preferred departure times of road users in a large urban network. Transportation, 45, 767-787.

Laurila, J. K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T.-M.-T., Dousse, O., Eberle, J. & Miettinen, M. The mobile data challenge: Big data for mobile computing research. Pervasive Computing, 2012.

Marschak, J. 1960. Binary Choice Constraints on Random Utility Indications. In: ARROW, K. (ed.) Stanford Symposium on Mathematical Methods in the Social Science. Stanford, California: Stanford University Press.

Mcfadden, D. 1974. Conditional logit analysis of qualitative choice behavior. Frontiers in Econometrics, 105-142.

Murtagh, F. 1985. Multidimensional clustering algorithms. Compstat Lectures, Vienna: Physika Verlag, 1985.

NCO. 2018. Official U.S. government information about the Global Positioning System (GPS) and related topics: GPS Accuracy [Online]. National Coordination Office for Space-Based Positioning, Navigation, and Timing. Available: https://www.gps.gov/systems/gps/performance/accuracy/ [Accessed 01 June 2018].

Nguyen, D.-Q. 2018. How work has evolved for Switzerland's women and men [Online]. swissinfo.ch. Available: https://www.swissinfo.ch/eng/society/gender-roles-since-1970_how-work-has-evolved-for-switzerland-s-men-and-women/43953426 [Accessed 14 June 2018].

Nychka, D., Furrer, R., Paige, J. & Sain, S. 2018. Package 'fields', The Comprehensive R Archive Network (CRAN).

Peer, S., Knockaert, J., Koster, P., Tseng, Y.-Y. & Verhoef, E. T. 2013. Door-to-door travel times in RP departure time choice models: An approximation method using GPS data. Transportation Research Part B: Methodological, 58, 134-150.

Qu, Y., Gong, H. & Wang, P. Transportation mode split with mobile phone data. Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on, 2015. IEEE, 285-289.

Ramos, R., Cantillo, V., Arellana, J. and Sarmiento, I., 2017. From restricting the use of cars by license plate numbers to congestion charging: Analysis for Medellin, Colombia. Transport Policy, 60, pp.119-130.

Sanko, N., Hess, S., Dumont, J. & Daly, A. 2014. Contrasting imputation with a latent variable approach to dealing with missing income in choice models. Journal of choice modelling, 12, 47-57.

Schlaich, J. 2010. Analyzing route choice behavior with mobile phone trajectories. Transportation Research Record: Journal of the Transportation Research Board, 78-85.

Schulz, D., Bothe, S. & Körner, C. Human mobility from gsm data-a valid alternative to gps. Mobile data challenge 2012 workshop, June, 2012. 18-19.

Small, K. A. 1982. The scheduling of consumer activities: work trips. The American Economic Review, 72, 467-479.

Small, K. A. 1987. A discrete choice model for ordered alternatives. Econometrica: Journal of the Econometric Society, 409-424.

Switzerland Tourism. 2018. Business hours [Online]. Switzerland Tourism. Available: https://www.myswitzerland.com/en-gb/business-hours.html [Accessed 15 June 2018].

The Economist. 2018. The glass-ceiling index: Progress has been slow but steady [Online]. The Economist Group Limited. Available: https://www.economist.com/graphic-detail/2018/02/15/the-glass-ceiling-index [Accessed 15 June 2018].

Themakart 2017. Key Figures. Urban Audit portraits 2013: core cities. Neuchâtel, Switzerland: Swiss Federal Statistical Office, ThemaKart.

Thorhauge, M., Cherchi, E., Walker, J.L. and Rich, J., 2017. The role of intention as mediator between latent effects and behavior: application of a hybrid choice model to study departure time choices. Transportation, pp.1-25.

Tomtom. 2016. Tomtom Traffic Index - Measuring Congestion Worldwide [Online]. TomTom International BV. Available: https://www.tomtom.com/en_gb/trafficindex/list?citySize=ALL&continent=ALL&country=CH [Accessed 26 May 2018].

Tukey, J. W. 1977. Exploratory data analysis, Addison-Wesley.

Wardman, M., Chintakayala, P., De Jong, G. & Ferrer, D. 2012. European wide meta-analysis of values of travel time. ITS, University of Leeds, Paper prepared for EIB.

Xiong, C. and Zhang, L., 2013. Positive model of departure time choice under road pricing and uncertainty. Transportation Research Record, 2345(1), pp.117-125.

Appendix A: Cluster identification for GPS data

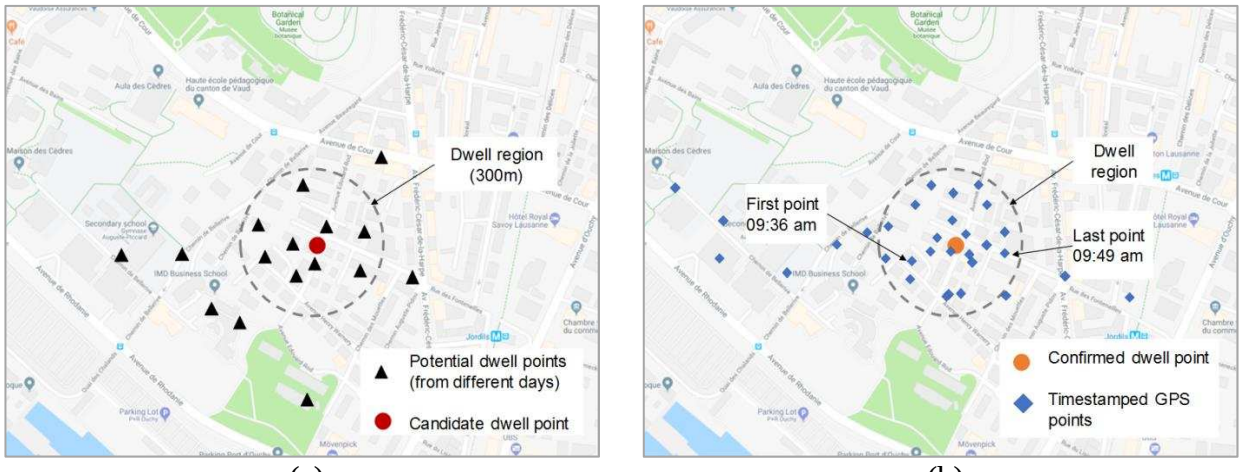
As cluster analysis on large datasets is a very challenging task because most spatial clustering algorithms require a full distance matrix, we conducted the clustering in stages.

We first split the data of each user according to the date observed. Full distance matrices comprising of all the possible pairs of GPS points observed on a particular day were then generated (Nychka et al., 2018). Thereafter, we conducted complete-linkage hierarchical based on the matrices of each day to identify groups of points that were potentially linked to the same dwell location (Everitt et al., 2011, Murtagh, 1985). Complete-linkage clustering ensures that we constrain the output cluster diameter to a specified size. This is particularly desirable when we know the accuracy range of the records. In this study, we specified a threshold distance of 300 meters as used in previous studies (Çolak et al., 2015, Jiang et al., 2013).

However, it is worth noting that spatial clustering algorithms need at least two data points to identify clusters. As a result, some of the identified clusters might have few points, and would not pass as potential dwell locations. We therefore specified a minimum duration of at least 10 minutes per day calculated using consecutive GPS points. The centroids of the points within the identified clusters were then labelled as potential dwell points.

The potential dwell points of each user from different days were combined and complete-linkage clustering conducted again with a threshold distance of 300 meters. This was aimed at clustering potential dwell points from different days in the vicinity of one another, thereby limiting the dwell region size to 300 meters. The centroids of the dwell regions of each user were then computed and labelled as candidate dwell points. At this stage, we did not impose a lower limit on the number of potential dwell points within dwell regions. Therefore, isolated potential dwell points that did not form clusters were simply re-labelled as candidate dwell points.

After establishing the candidate dwell points of each user, we identified all the GPS points within a radius of 150 meters from these locations and ordered the data according to timestamp. This was followed by applying a minimum dwell time constraint of 10 minutes each time a user was continuously observed within the vicinity of a candidate dwell point. Whenever this condition was met, the candidate dwell point was relabelled as a confirmed dwell point. This is illustrated in Figure A1.



(a) (b)

Figure A1: GPS dwell point identification (a) identification of a candidate dwell point from the potential dwell points, (b) application of a dwell time constraint to confirm the candidate dwell point

Appendix B: Cleaning the trip data to identify travel modes

B1. Setting the minimum travel time constraint

To begin with, it is important to note that the observed travel times of the users relate to the inter-boundary components of the O-D links since the trip start and end times are only captured when the users cross the home/work location dwell boundaries. However, these inter-boundary travel times need to be sufficient to enable the observation of reasonable variations in travel time across different time periods. As earlier mentioned, the morning and evening peak travel time increment factors for Lausanne are 1.44 and 1.63 respectively. Since these factors are quite low, for very close O-D pairs, the variations in travel time would not be significant enough to influence changes in departure time choices. In this study, we specify a median travel time of 10 minutes as the lower threshold for direct trips between the users' home-to-work O-D pairs and only consider those meeting this criterion. It may be noted that the exclusion of close O-D pairs also mitigates the observation of potential false trips due to signal jumps that were undetected during the data pre-processing phase (Iqbal et al., 2014).

B2. Identification of trips with unreasonably long travel times

We analyse each user's travel time for a particular trip in relation to the minimum travel time observed for the user along the same trip chain to identify trips with unreasonably long travel times. Travel times generally increase due to traffic congestion, however, when the increase is very big, we suspect other factors such as uncaptured trip segments due to switching off of phones.

To determine the most reasonable upper limits of travel time, we calculate the ratios of the observed travel times versus the minimum travel times for each of the user's trips. These ratios give an indication of the levels of congestion (i.e. the higher the ratio, the higher the level of congestion). We then combine the computed ratios for all the users and estimate the upper limit as follows; $Upper\ limit = Q3 + 1.5 * (Q3 - Q1)$, where $Q1$ and $Q3$ are the first and third quartiles respectively (Tukey, 1977). We use the GSM data for this analysis as it captures most of the trips made.

The estimated upper limits of the ratios were 2.21 and 2.12 for the home-to-work, and the work-to-home commutes respectively. It may be noted that these limits seem reasonable when compared to the congestion factors reported for Lausanne, that is, 1.44 and 1.63 for the morning and the evening peaks respectively (TomTom, 2016). We exclude trips whose travel times exceeded the estimated upper limits.

B3. Identification of potential travel modes

We first apply a minimum distance constraint of 5 kilometres as previous studies have shown that people are less likely to walk or cycle beyond this distance (Hydén et al., 1999). It may be noted that we do not use euclidean distances, rather, we calculate the minimum distances by road for each O-D pair using the Google Distance Matrix API (Google Developers, 2018).

However, another important aspect is the speed of the users. We only consider trip chains where the users' median speeds exceed 15 kilometres per hour, thereby excluding those where the users typically walk or cycle (Bernardi and Rupi, 2015). It may be noted that the calculated speeds are generally over-estimated since we use centre-to-centre O-D distances versus the inter-boundary travel times. Despite this limitation, observing median speeds above 15 kilometres per hour for trip lengths above 5 kilometres is considered a good indicator that the users generally use motorised transport for those trip chains.