Modelling long-distance route choice using mobile phone call detail record data: A case study of Senegal

The growing mobile phone penetration rates (world-wide) have led to the emergence of large scale call detail records (CDRs) that could serve as a low-cost data source for travel behaviour modelling compared to the commonly used data sources such as stated preference (SP) data. However, to the best of our knowledge, there is no previous study specifically evaluating the potential of CDR data in the context of route choice behaviour modelling. Being event-driven, the data is discontinuous and only able to yield partial trajectories, thus presenting serious challenges for route identification, especially in highly overlapping networks. This paper proposes techniques for inferring the users' chosen routes or subsets of their likely routes from partial CDR trajectories, as well as data fusion with external sources of information such as route costs and travel times, and then adapts the broad choice framework to the current modelling scenario where the road network is highly overlapped. The model results show that CDR data can capture the expected sensitivities towards route attributes and the behaviour towards overlapping routes. The value of travel time estimates derived from the models are found to be realistic for the study area (Senegal). These findings are timely for developing countries with low budgets for transport studies.

Keywords: Route choice behaviour, Broad choice, Mobile phone data, Call detail records, Value of travel time

1. Introduction

The modelling of route choice behaviour for long journeys, inter-city and inter-regional trips is an important component of large-scale transport planning models. The emergence of technologies that enable passive collection of mobility trajectories has led to a step change in route choice modelling. Most of these studies are based on Global Positioning System (GPS) data, primarily from navigational devices (Li et al., 2018, Hess et al., 2015, Broach et al., 2012, Bierlaire and Frejinger, 2008), and more recently from smartphones (Bierlaire et al., 2010, Papinski et al., 2009). The navigational devices are still not used widely in all parts of the world (in countries of the Global South in particular). Although smartphone ownership is on the rise throughout the world, the quality of data from these devices strongly rely on the availability of the Assisted-GPS (A-GPS)[1] feature, internet connectivity and data storage capacity, which leads to small sample sizes as seen in most related studies (e.g. Nitsche et al., 2014, Bierlaire et al., 2013, Nitsche et al., 2012, Bierlaire et al., 2010). Thus both sources pose a substantial risk of sampling bias.

This problem can be overcome by taking advantage of the large scale anonymous datasets that are already being passively collected by operators for different purposes, and applying these to transport studies. Such datasets have already yielded promising results various mobility studies. Examples include; social media data (Hawelka et al., 2014, Hasan et al., 2013), smart card data (Chakirov and Erath, 2012, Agard et al., 2006), and network-generated mobile phone data such as Call Detail

Records (CDRs)[2] and Global System for Mobile communications (GSM)[3] data (Çolak et al., 2015, Jiang et al., 2013, Schlaich et al., 2010). However among these, network-generated mobile phone data is particularly a promising source due to the high mobile phone penetration rate world-wide (GSM Association, 2017).

A review of literature shows that there have been a few route identification studies using network-generated mobile phone data (e.g. Nie et al., 2015, Leontiadis et al., 2014, Hoteit et al., 2014, Schlaich et al., 2010), however, most of these studies end on route identification and do not attempt to investigate the factors affecting route choice behaviour. At the moment, only Schlaich (2010) combines GSM trajectories with traffic state information to analyse the influence of variable message signs (VMS) and other factors on route choice. However, the success of Schlaich's study could in part be attributed to the use of GSM data, which is network-driven and semi-continuous in nature as opposed to CDR data, which is event-driven and discontinuous, thus presenting serious trajectory identification challenges. Despite these challenges, the data is more readily available as it is stored longer for billing purposes as opposed to GSM data, which is discarded after each location area update operation to save computer memory. Thus, methods developed to harness the potential of such a widely available data source can have wide benefits in both the developed and developing worlds. This motivates this research where we investigate the feasibility of CDR data for modelling route choice behaviour.

[2] CDR data reports the time stamped locations of communication events (i.e. voice calls, text messages, and data calls) as well as the details of the request (i.e. the duration and direction).

[3] GSM data reports the IDs of all the GSM cells traversed by an active mobile phone at regular time intervals (irrespective of the calling or texting patterns of the users).

However, it is important to underscore the practical challenges that stem from the use of CDR data, and how this impacts our work. Since the data is event-driven, there is an increased risk of not capturing any trajectories, especially for very close O-D pairs with short travel times. For this reason, we limit our scope to long-distance inter-regional route choice as there is an increased possibility that a user will use his/her phone at different points during the trip, thus enabling the capture of his/her partial trajectory. The fact that only partial trajectories are observed means that in some cases, it is not possible to precisely infer the chosen routes. In such cases, route choice is observed at a broad sub-group level (e.g. northern, southern etc.), where each sub-group comprises of a small set of possible routes. This prompts us to adapt the 'Broad Choice Modelling Framework', developed in the context of vehicle type choice (Wong, 2015) to route choice modelling using noisy CDR data. To the best of our knowledge, the Broad Choice Modelling Framework is the only model structure that can deal with the case when the choice data (dependent variable) has multiple levels of aggregation (i.e. unique choice for some observations while broad sub-group for the other ones) as in this case. We note that this may be problematic in dense inter-urban networks, where it would be difficult to identify a small enough subset of possible routes using a few CDR locations. However, with the increasing trend of mobile internet usage (Gerpott and Thomas, 2014), the frequency of CDR locations is likely to improve significantly in the near future, which adds further to the timeliness of the present paper.

Although CDR data is only able to capture the partial trajectories of frequent phone users, the samples are usually large, thus increasing the possibility that they are representative enough to capture rational route choice behaviour. The need to investigate this assertion forms the basis of our validation exercise where we obtain stable results. The developed models are used to estimate the value of travel time (VTT) for Senegal (the study area), yielding reasonable estimates. The study is timely in the sense that it extends the application of CDR data beyond travel pattern visualisation to econometric modelling of travel behaviour. This could motivate reliable and low cost policy formulation in both developed and developing countries.

The rest of the paper is arranged as follows; section 2 presents a review of relevant literature, section 3 presents the data description, section 4 presents the data processing conducted, section 5 presents the modelling framework, section 6 discusses the model results, while section 7 presents the summary and conclusions.

2. Literature review

This section briefly reviews literature on the applications of mobile phone data in transport studies, as well as different models of route choice.

2.1. Previous applications of mobile phone data to transportation studies

The last few decades have seen significant research effort aimed at investigating the potential application of mobile phone data to various aspects of transportation studies, so far yielding promising results. To date, the data has been widely applied in human mobility modelling to understand individual's travel patterns (e.g. Ahas et al. 2010, Ahas et al. 2015, Deville et al., 2016, Diao et al., 2016, Xu et al., 2015, Calabrese et al., 2013, Isaacman et al., 2012, Shaw et al. 2016, Song et al., 2010, Yuan et al. 2012, Zhao et al. 2013) and correlating it with phone usage patterns (e.g. Yuan et al. 2012), tourism surveys to understand the attractiveness of different tourist sites (Ahas et al., 2008, Ahas et al., 2007), urban area (land use) classification to facilitate urban planning (Pei et al., 2014, Yuan and Raubal, 2012), analysing commute (Kung et al., 2014, Ahas et al., 2010) and migration patterns (e.g. Wang et al. 2019), estimating trip making rates (e.g. Çolak et al., 2015), developing origin-destination matrices (e.g. Çolak et al., 2015, Iqbal et al., 2014, Calabrese et al., 2011, White and Wells, 2002), detecting travel modes to estimate mode shares (e.g. Qu et al., 2015, Wang et al., 2010, Reddy et al., 2008), and estimating traffic parameters to assess the traffic conditions of key roads (e.g. Bolla et al., 2000). However, we place focus on studies related to route identification. A few of these studies have used GSM data, which reports the complete mobile phone location area sequences of each user, thus enabling the easy identification of routes through sequence matching (e.g. Ma et al. 2013, Tettamanti et al., 2012, Schlaich et al., 2010) and probabilistic methods such locality-sensitive hashing and graph clustering (e.g. Görnerup, 2012). Furthermore, a growing number of studies have focussed on analysing the potential of CDR data, which is more widely available compared to GSM data and yet challenging to use as mentioned in the introduction section. For example, Doyle et al. (2011) use the virtual cell paths technique to extract user trajectories from CDR data, and generate the kernel density paths for different routes to validate their findings. Saravanan et al. (2011) analyse the spatial and temporal information of CDR events over a long period of time to establish the daily routines and routes of the users. Hoteit et al. (2014) join subsequent triangulated CDR locations using linear, cubic, and nearest - neighbour interpolation to model the potential trajectories. Leontiadis et al. (2014) calculate the weights for each road segment within the cell areas linked to a user's communication events and determine the shortest weighted path for a given OD pair. Nie et al. (2015) mark each route with a subset of k optimal cell handover sequences extracted from the full set of possible handover sequences and match these with those observed in the cell phone hand over data (similar to CDR data) based on the degree of similarity. In this study, we follow a slightly similar approach, however, instead of using a similarity index, we pursue the idea of unique and shared location area sequences in the context of broad choice modelling as explained later.

2.2. Existing route choice models

The vast majority of route choice models belong to the family of discrete choice models (see Ben-Akiva and Lerman, 1985 for details), with the multinomial logit (MNL) model (McFadden, 1974) being the most widely used. However, the MNL model is affected by the irrelevance of independent alternatives (IIA) property, which can be problematic for highly overlapping routes (Ramming, 2002). This has motivated the development of more advanced route choice models to address this challenge. Examples include the nested recursive logit model (Mai et al., 2015), the c-logit model (Cascetta et al., 1996), the path size logit model (Ben-Akiva and Ramming, 1998), the link nested (cross-nested) logit model (Vovsha and Bekhor, 1998), and the multinomial probit model with logit kernel (Daganzo et al., 1977). Details of how some of these models overcome the overlapping route problem are discussed in section 5.3 of this paper.

An important point to note is that the complexities of route choice modelling go beyond the overlapping route problem. Choice set generation is a key challenge, especially in highly overlapping dense urban networks, where several alternative routes can be possible, and yet individuals do not consider all the alternatives while making choices (Prato, 2009). Several choice set generation methods have been proposed in the literature including, the k-shortest path algorithms (e.g. Shier, 1979, Bellman and Kalaba, 1960), the labelling approach (Ben-Akiva et al., 1984), link elimination approaches (e.g. Azevedo et al., 1993, Bellman and Kalaba, 1960), link penalty approaches (e.g. Rouphail et al., 1995, De La Barra et al., 1993), simulation approaches (e.g. Sheffi and Powell, 1982), doubly stochastic generation functions (e.g. Nielsen, 2000), constrained enumeration methods (e.g. Prato and Bekhor, 2006), and probabilistic methods (e.g. Cascetta and Papola, 2001, Manski, 1977). However, since the focus of this paper is long distance trips, where the alternatives are usually few in number, choice set determination is more straightforward as discussed later in section 4.2 of this paper.

As mentioned earlier, the discontinuous nature of CDR data presents serious route identification challenges, resulting in route choice observations at different levels of aggregation (i.e. unique route choices for some travellers while broad sub-groups of the possible routes for the rest). Although there is no specific route choice model dealing with such a scenario, recent advances in other fields of choice modelling have addressed this challenge using the broad choice framework (proposed by Wong, 2015 and later used by Habibi et al. 2017, Lloro and Brownstone 2018, Wong et al. 2018, Yip et al.2018 in the context of vehicle choice and by Tran et al. 2019 in the context of hotel choice), which can be adapted to deal with the current route choice modelling scenario, discussed later in this paper. It may be noted that according to the choice modelling literature, this framework is the only one that can deal with the case where the choices are uniquely observed only for some respondents and for the rest, only the sub-group of choices are observed.

3. Data

This study uses CDR data collected from Senegal as part of the Orange Data for Development (D4D) challenge (de Montjoye et al., 2014).

3.1. Study area

Senegal is located in West Africa with a population of approximately 13.5 million according to the 2013 population census (ANSD, 2016). Road transport accounts for over 99% of all passenger travel (World Bank, 2004). The only long-distance train service (the Dakar-Niger line) was discontinued in May 2010 (Imedia and Calao

Production, 2013). The country has a sparse national road network (see Figure 1), and for some O-D pairs, there is only one feasible alternative, making them unsuitable for route choice modelling. This study ignores such O-D pairs and only considers those where alternative routes exist. In total, twelve distant O-D pairs are considered, and these are; Dakar-Bakel, Dakar-Matam, Thies-Bakel, Thies-Matam, Diourbel-Bakel, Diourbel-Matam, and the corresponding O-D pairs for the reverse directions as shown in Figure 1. The long travel times between these regions increase the possibility of capturing the users' partial trajectories as explained earlier.

Figure 1. Study area

3.2. CDR data

The CDR data was collected between January and December 2013 and aggregated to the arrondissement (district) level by the data provider. The geographical location of the arrondissements is presented in Figure 1 where we also show the tower locations for illustration purposes.

The original CDR data comprised of 9 million unique users (67% of the study area population). This was pre-processed to retain frequent phone users (i.e. those with interactions on 75% of the days in a year) and randomly split into smaller monthly rolling sub-samples made available for research (see de Montjoye et al., 2014 for details). The user IDs in each sub-sample are anonymised to prevent possible re-identification across the different months.

The data for each month comprises of about 150,000 users. Assuming that the most commonly observed arrondissement for each user during the month is their home district, the monthly population sampling rate ranged from 2.4% in Dakar (the capital) to 0.4% in the rural regions. On average, these users together generated over 40 million records per month (see excerpt of the CDR data in Table 1a). The data is reduced to remove duplicate records resulting in the processed arrondissement visitation data presented in Table 1b.

Table 1a. Excerpt of the raw CDR data

Table 1b. Excerpt of the processed arrondissement visitation data

The overall level of user mobility is illustrated in Figure 2. As shown, most users visited less than three unique arrondissements per month. The low levels of inter-arrondissement mobility led to the capture of few trajectories as reflected in the final sample size (see Section 4.2).

Figure 2. Average monthly arrondissement observation frequency distribution

Although it is difficult to detect false tower jump movements in aggregate CDR data (Çolak et al., 2015, Iqbal et al., 2014), this is not a big factor for distant O-D pairs where the origin and destination arrondissements have been grouped into regions.

4. Data preparation for analysis

This section describes the analysis carried out on the processed arrondissement visitation data in Table 1b to identify the routes followed, as well as the processes of estimating the route attributes.

4.1. Route identification

The route identification process is summarised in Figure 3. This is divided into two main stages as described in the subsequent sections.

Figure 3. Summary of the route identification process

4.1.1. Generation of unique and shared CDR arrondissement sequences

Route arrondissement sequences (which are extracted from maps) show the order of all the arrondissements traversed by a particular route between a given O-D pair. On the other hand, CDR arrondissement sequences (which are extracted from the CDR data) show the order of the arrondissements in which a user used his/her phone during the trip, and are subsets of the route arrondissement sequences.

For any given trip along a particular route, several possible CDR arrondissement sequences can be observed depending on the number and the location of the CDR events. These can be obtained by generating permutations of different sizes based on the route arrondissement sequence (in which order matters and no repetitions are allowed). However, since most of the O-D pairs have overlapping routes, some of the CDR arrondissement sequences can be linked to more than one route if all the intermediate CDR events occurred along the shared sections. In this case, it would only be possible to observe the subset of routes that were potentially followed (see illustration in Figure 4 where the blue areas indicate the shared arrondissements, while the grey and green areas indicate the unique arrondissements for the Northern 1 and 2 routes respectively).

Figure 4. Arrondissement paths (Dakar-Bakel O-D pair as an example)

The generated CDR arrondissement sequences linked to each route were cross-referenced to identify the permutations linked to unique routes (i.e. unique CDR arrondissement sequences) and those shared across multiple routes (i.e. shared CDR arrondissement sequences). The outcome of this analysis was a list of all the possible

CDR arrondissement sequences labelled with the associated routes or sub-groups of the possible routes (see example in Table 2).

4.1.2. Extraction the O-D pair trip trajectories from CDR data

The processed arrondissement visitation data for each user (see excerpt in Table 1b) was analysed to extract sub-sequences linked with possible trips between the regions of interest following the criteria below;

- The first and the last arrondissements in the sub-sequence must be located within different regions of interest, and the user must not be observed in an upstream or a downstream region of interest within the same day for origins and destinations respectively;
- The dwell time in the origin and the destination regions of interest must be longer than that required to directly traverse each of them to increase the possibility that these are the trip start and end locations. A user needs to use his/her phone at least twice in each of these regions to calculate the dwell time;
- The intermediate arrondissements in the sub-sequence must all be associated with one of the defined arrondissement/corridors paths (see Figure 4) to ensure only direct trips are retained; and
- The timestamp difference between the origin and the destination must not exceed 24 hours, which is used as an upper limit to distinguish between users with direct trips but delay to use their phones on arrival, and those with intermediate destinations, thereby arriving late.

The extracted sub-sequences meeting all the above criteria were either assigned to unique routes or sub-groups of the possible routes by cross-referencing with the labelled lists generated in Section 4.1.1. Table 2 presents an excerpt of the route assignment data.

Table 2. Excerpt of the route assignment data

In this data, 70% comprised of unique assignments, while 30% comprised of broad assignments. Since this is a scenario where for some users and/or trips, we know the chosen route at a more disaggregate level than for others, we use the Broad Choice Modelling Framework (Wong, 2015) to analyse route choice behaviour.

4.2. Estimation of route attributes

For model estimation, it is critical to determine the choice set and the attributes of the alternatives. We assumed that the choice set is comprised of the routes that have ever been chosen by the different users. Routes not chosen by any user for the whole year were excluded from the choice set. On average, each origin-destination pair had four alternatives. Given that the total dataset is comprised of 9,453 records from 6,497 users, this is not a very restrictive assumption. Given the choice sets, we reviewed previous studies to identify the attributes typically used in route choice models and their availability status for Senegal as summarised in Table 3. Although data on six explanatory variables was available, the final model specification contains three explanatory variables only as the inclusion of the other variables led to correlation problems and/or illogical model results. A detailed explanation of the variable specification process is presented in Section 6.1. The subsequent sections summarise the processes of estimating some of these attributes.

Table 3. Attributes typically used in route choice models

4.2.1. Link length and surface attributes

Data on the link lengths and surface attributes (i.e. paved or unpaved) was derived from the Senegal roads GIS layer (ArcGIS, 2013). This was updated to reflect the situation in 2013 relying on road condition reports sourced from government and other relevant websites (Ageroute Senegal, 2017, ANSD, 2017, Logistics Cluster, 2013).

4.2.2. Travel time

Travel time cannot be reliably estimated from the CDR data as users do not necessarily use their phones at the moment of departure or arrival. The typical travel times for most links in 2013 were obtained from the website of Logistics Capacity Assessment (Logistics Cluster, 2013). For links not covered by this website, we relied on Google Maps (Google Maps, 2017a).

4.2.3. Travel cost

Travel cost was estimated in terms of the vehicle operating costs (VOCs) per user (i.e. fuel and non-fuel costs). After a review of several VOC estimation techniques, we settled for the HDM-III model (Watanatada et al., 1987) due to its applicability to developing countries and input data availability. The HDM-III model is an earlier version of the more advanced HDM-4 (Kerali, 2000), which we could not use due to input data constraints.

The HDM-III model relies on vehicle calibration data (where we used default values) and other basic input data (see Table 4) to estimate the VOCs for each vehicle type. The model works by defining link-specific relationships between the International Roughness Index - IRI (Sayers et al., 1986) and speed, and using the IRI values at the respective average link speeds (derived from Sections 4.2.1 and 4.2.2) to estimate the link-specific VOCs, which are summed to estimate the route VOCs.

Table 4. HDM-III basic input data

The estimated route VOCs need to be converted to person costs. Given the anonymous nature of CDR data, we use information on the typical occupancy rates and mode shares (see Table 5) to estimate the weighted average VOCs per user for each route.

In this research, we have used the same average travel cost for all the users along a particular route between a given O-D pair. An improved approach would have been to estimate user-specific travel costs based on the corresponding travel speeds, however, this was not possible due to difficulties in obtaining the user-specific travel times as discussed in the previous section.

Table 5. Typical occupancy rates and mode shares in Senegal (World Bank, 2004)Other attributes

The other attributes considered were the scenic characteristics and the urban developments along each route. Scenic variables were estimated in terms of route lengths traversed through nature reserves (Google Maps, 2017b), while urban developments were reflected as the number of towns along each route as shown in Figure 1 (Worldatlas, 2017).

5. Modelling framework

We use discrete choice models in this study since route choices are discrete and mutually exclusive. To develop these models, we apply the random utility theory (Marschak, 1960), a well-established approach for estimating discrete choice models.

5.1. Basic model

Suppose U_{nr} is the utility of choosing route r by individual n. This can be expressed as;

$$\mathrm{U}_{\mathrm{nr}} = \mathrm{V}_{\mathrm{nr}} + \varepsilon_{\mathrm{nr}} \tag{1}$$

Where V_{nr} and ε_{nr} are the systematic and the random parts utility respectively. The systematic utility is a function of the observed route attributes, and may be expressed as $V_{nr} = \beta' X_{nr}$, where X_{nr} is a vector of the attributes of route r for individual n and β is a vector the model parameters. We assume that the random term ε_{nr} is independently and identically distributed across the alternatives following a type I extreme value distribution, and use the Multinomial Logit (MNL) model to estimate the route choice probabilities as follows (see McFadden, 1974 for details);

$$\mathrm{P}_{\mathrm{n}}(\mathrm{r}) = \frac{\exp(\mathrm{V}_{\mathrm{nr}})}{\sum_{\mathrm{r}^* \in \mathrm{C}_{\mathrm{n}}} \exp(\mathrm{V}_{\mathrm{nr}^*})} \tag{2}$$

Where, $P_n(r)$ is the probability of individual n choosing route r, and C_n is the choice set. Given the route choice probabilities, the model parameters can be estimated by maximising the log-likelihood function below;

$$\mathrm{LL} = \sum_{\mathrm{n}} \sum_{\mathrm{r}} \left[\mathrm{Z}_{\mathrm{nr}} \bullet \ln\left(\mathrm{P}_{\mathrm{n}}(\mathrm{r}) \right) \right] \tag{3}$$

Where Z_{nr} is a dummy variable, which is equal to 1 if and only if user n chooses route r.

5.2. Accounting for broad choices

The log-likelihood function in Equation (3) assumes that all the route choices are uniquely observed, and is inadequate for the current scenario, where we also have broad sub-group choices. Therefore, we use the broad choice modelling framework proposed by Wong (2015) to account for this situation.

In the broad choice framework, the choice probabilities of the broad sub-groups are expressed as a sum of the choice probabilities of the member alternatives. For example, the choice probability of the 'Northern' broad sub-group is the sum of the 'Northern 1' and the 'Northern 2' route choice probabilities (see Figure 4 and Table 2). The joint probabilities of the broad sub-groups capture the aggregate shares at the unique route choice level using the relative probabilities of the constituent routes.

The goal of model estimation is to maximise the probabilities of both the observed routes and the broad sub-groups for users with unique and broad choices respectively. The log-likelihood function is specified as follows (Wong, 2015);

$$LL = \sum_{n}\sum_{b}\left[Z_{nb} \bullet \ln\left(\sum_{r \in S_b} P_n(r)\right)\right] \quad (4)$$

Where S_b is a set comprising of the routes in broad category b. For uniquely assigned trips, set S_b comprises of only one alternative. Z_{nb} is a dummy variable, which is equal to 1 if and only if user n is associated with category b.

5.3. Accounting for overlap

A major weakness of the MNL model (Equation 2) is the IIA property, which could lead to illogical route choice probabilities for highly overlapping routes as is the case in this study. This is illustrated using the overlapping route problem (Ramming, 2002, Cascetta et al., 1996) in Figure 5.

Figure 5. Overlapping route problem (Cascetta et al., 1996, Ramming, 2002)

Here, all three routes have the same total length L, however, routes 1 and 2 follow the same alignment for length l followed by distinct sections, each of length $L - l$. The MNL model predicts equal shares for each of the routes irrespective of the overlap length. However, as the overlap increases (as l tends to L), it becomes difficult to distinguish between routes 1 and 2, and it is expected that route 3 will take a share of 50%, while routes 1 and 2 will each take a share of 25%.

Various models accounting for overlap were presented in Section 2.2, however, this study only uses the c-logit (Ramming, 2002, Cascetta et al., 1996) and the path size logit (Ben-Akiva and Ramming, 1998) models for illustration purposes as the modelling scenario did not require complex formulations. In general, these models are modifications of the MNL model, where the systematic route utilities are adjusted using certain correction factors as follows;

$$P_n(r) = \frac{\exp(V_{nr} + \tau_{nr})}{\sum_{r^* \in C_n} \exp(V_{nr^*} + \tau_{nr^*})} \tag{5}$$

Where τ_{nr} is the systematic utility correction factor for route r.

5.3.1. C-logit model

For the c-logit model, the correction factor τ_{nr} is a commonality factor (Cascetta et al., 1996). Different possible specifications have been proposed, however, this study uses the following common specification (Ramming, 2002, Cascetta et al., 1996);

$$\tau_{nr} = \beta_{CF} \ln\left[\sum_{r^* \in C_n} \left(\frac{L_{rr^*}}{\sqrt{L_r L_{r^*}}}\right)^{\gamma_{CF}}\right] \tag{6}$$

Where $L_{rr*}$is the overlap length between routes r and r^*, L_r and L_{r*} are the total lengths of routes r and r^* respectively, β_{CF} and γ_{CF} are the unknown parameters to be estimated. From Equation 6, the ratio in the brackets is proportional to the degree of overlap, while the corresponding logarithm has an inverse negative relationship. Thus

β_{CF} and γ_{CF} are expected to have negative and positive signs respectively to allow for the positive adjustment of route utility with decreasing overlap (Ramming, 2002).

5.3.2. Path size logit model

For the path size logit model, the correction factor τ_{nr} is a path size term, which is computed as the weighted average of the constituent link sizes. The specification adopted for this study is as follows (Ben-Akiva and Ramming, 1998);

$$\tau_{nr} = \beta_{ps} \ln\left[\sum_{a \in \Gamma_r}\left(\frac{l_a}{L_r} * \frac{1}{N_{ar}}\right)\right] \quad (7)$$

Where $1/N_{ar}$ is the inverse of the number of routes sharing the link a (the link size), and l_a/L_r is a weight representing the proportion contributed by link a to the overall route size. From Equation 7, it is observed that route size is inversely proportional to the degree of overlap, while the corresponding logarithm has a negative proportional relationship. Thus, the path size parameter is expected to be positive to allow for negative adjustment of route utility with increasing overlap.

The models accounting for overlap are generally expected to have better fit than the MNL model. Model fit is evaluated using the adjusted-rho square and the likelihood ratio tests (see Ben-Akiva and Lerman, 1985 for details).

6. Model results

This section presents the modelling results. We start by discussing the variable specification, followed by the model estimation and validation results.

6.1. Variable specification

The attributes available for possible inclusion in the model are summarised in Table 3.

However, these could not all be specified together or in the same way for various reasons as explained below.

For travel time and cost, after initial tests using the linear specification, we used the log-transforms of the variables to allow for utility damping with respect to increasing time and cost (see Daly, 2010 for details). Various interactions of these variables with others (such as surface type) were tested, however, this led to correlation problems, and hence generic variables were specified.

The urban developments along the alternative routes were incorporated in terms of the average distance between towns rather than the number of towns to avoid situations where longer routes also have more towns. Again, this was specified using the log-transform of the variable for similar reasons as the travel time and cost. This being a largely rural road network with no traffic signals, the average distance between towns is the only variable we could use to capture traffic flow interruptions.

An attempt was made to incorporate scenic beauty into the model using either the length or the proportion of route length traversed through nature reserves, however, we could not achieve intuitive model results potentially due to our lack of detailed knowledge about the characteristics of these reserves, and the security levels of the corresponding routes. The final systematic utility specification is as follows;

$$\mathrm{V_{nr}} = \beta_{\mathrm{l-cost}} \ln(\mathrm{C_{nr}}) + \beta_{\mathrm{l-time}} \ln(\mathrm{T_{nr}}) + \beta_{\mathrm{l-dtown}} \ln(\mathrm{Dt_{nr}}) \tag{8}$$

Where C_{nr}, T_{nr}, and Dt_{nr} give the travel cost, the travel time, and the average distance between towns respectively of route r for individual n, and the βs are the corresponding model parameters to be estimated.

6.2. Estimation results

We present the estimation results of the MNL model, the c-logit model, and the path size logit based on the full sample in Table 6 for comparison purposes.

Table 6. Estimation results

6.2.1. Statistical performance of the models

The estimated parameters in each of the three models are statically significant at the 95% level of confidence. The only exceptions are the parameters associated with the average distance between towns and the beta coefficient of the commonality term in the c-logit model, however, these were retained in the model as they are important.

To evaluate the collective statistical significance of the models, we used the likelihood ratio test (see Ben-Akiva and Lerman, 1985 for details), whose values are reported in the last two rows of Table 6. From these, it is noted that the p-values for all the models are less than 0.01. Therefore, the hypothesis that the respective model parameters are collectively equal to zero is rejected at the 99% level of confidence.

6.2.2. Route variables

The parameter signs for the travel cost variable are consistent with a priori expectations in each of the three models. In general, an increase in the cost of an alternative is expected to have a negative impact on its utility, hence the negative parameter sign. The same explanation holds for the travel time parameters as individuals generally prefer shorter travel times.

On the other hand, the average distance between towns has a positive parameter sign. As earlier mentioned, this variable gives an indication of the amount of uninterrupted flow. Although no traffic congestion problems have been reported in these towns, traffic generally slows down due to speed control measures leading to delays. An increase in the average distance between towns therefore indicates more uninterrupted flow, hence the positive parameter sign.

6.2.3. Overlap correction parameters

For the c-logit model, it is observed that the beta parameter of the commonality term is negative while the gamma parameter is positive. Similarly, the path size parameter in the path size logit model has a positive parameter sign. These results are in line with behavioural expectations as discussed earlier under Equations 6 and 7, an indication that CDR data is able to capture the behaviour towards overlapping routes.

6.2.4. Model comparison

A comparison of the adjusted rho-square values in Table 6 shows that the models accounting for overlap (i.e. the c-logit and the path size logit models) perform better than the MNL model. This is as expected given that the national road network of Senegal is highly overlapping (see Figure 1 and discussion under Figure 5). It is also worth noting that the path size logit model outperforms the c-logit model because the behavioural underpinning of the systematic utility adjustment process in the path size logit model is stronger than that in the c-logit model (Ramming, 2002).

The statistical significance of the improvements associated with accounting for overlap are evaluated using the likelihood ratios of the c-logit and the path size logit models with respect to the MNL model (see Table 7).

Table 7. Statistical comparison of the models

From Table 7, it is noted that the p-values for the c-logit and the path size logit models are all less than 0.01, an indication that accounting for overlap has a statistically significant effect (at the 99% confidence level) beyond the improvements contributed by the additional degrees of freedom resulting from the extra parameters (see Ben-Akiva and Lerman, 1985 for details).

6.2.5. Policy insights

This section highlights the policy implications of the reported results in terms of the value of travel time (VTT). This metric quantifies the benefits derived from reduced travel time in monetary terms, and is useful in transportation cost-benefit analysis (Mackie et al., 2001). The value of travel time is calculated by taking the ratio of the partial derivatives of the systematic utility function (V) with respect to the travel time (T_{nr}) and cost (C_{nr}) as follows;

$$\mathrm{VTT} = \frac{\partial \mathrm{V_{nr}} / \partial \mathrm{T_{nr}}}{\partial \mathrm{V_{nr}} / \partial \mathrm{C_{nr}}} = \frac{\beta_{\mathrm{l-time}}}{\beta_{\mathrm{l-cost}}} \frac{\mathrm{C_{nr}}}{\mathrm{T_{nr}}} \tag{9}$$

Figure 6 shows the variations of VTT with respect to the cost per hour of the alternatives for the range in the estimation data, and it is observed that the values increase as the alternatives become more expensive, which is expected. It is noted that the path size logit model gives the lowest values, which are deemed to be more accurate due to the superior performance (see Tables 6 and 7).

Figure 6. Variations in VTT across the alternatives and models

However, to assess how realistic these estimates are, we computed the average values for each model using the estimation data, and compared these values with those derived from other studies and other relevant statistics as summarised in Table 8.

Table 8. Comparison of the VTT estimates with other sources

In the Africa-wide meta-analysis by Teye et al. (2017), VTT is estimated as a function of the GDP per capita. However, the reported mean value (4.3213 USD/hr) seems high when compared to the toll being charged on the new Dakar–Diamniadio toll highway for a time saving of one hour (2.3411 USD/hr), a value that was highly criticised by the Senegalese media as being extremely high (Gainer and Chan, 2016).

Although the median hourly wages do not necessarily translate into the value of travel time, they give a good indication of the range in which these values should fall, and as observed in Table 8, the average VTT estimate for the path size logit model is very close to the Senegalese median hourly wage. We consider this VTT estimate to be more reasonable for Senegal.

6.3. Validation results

The models based on the full sample provide intuitive results in terms of the parameter signs and the relative model performance. To assess the stability and the predictive performance of the models, the dataset was randomly split into five parts at the individual level. Five rolling subsets, each comprising of 80% of the users were generated for model estimation purposes. For each of these, a complementary subset comprising of 20% of the users was generated for validation purposes. The models were re-estimated on each of the 80% subsets, and the parameter estimates applied to the corresponding 20% hold-out subsets to estimate the predictive measures of fit. Table 9 presents the summary outputs from this process.

Table 9. Validation results

The general interpretation of the parameter signs and the relative model performance in each of the 80% estimation subsets remained the same as in the full sample, an indication that the data is representative (detailed results available on request). A comparison of the measures of fit in estimation and validation shows that there is no significant loss in model fit, an indication that the performance of the models during estimation is not due to overfitting, rather it is due to the strong explanatory power of the variables.

A comparison of the predictive measures of fit shows that the relative model performance during estimation is mirrored on the holdout samples, with the path size logit model still giving the best model performance due to its behavioural superiority. It would have been interesting to further validate the above results with outputs of route choice models based on traditional data or GSM data, but this was not possible due to lack of data in Senegal.

7. Summary and conclusions

This paper has successfully demonstrated the potential of CDR data to capture rational route choice behaviour for long-distance inter-regional O-D pairs. The broad choice framework was used to leverage the limitations of CDR data where unique route choices could not be observed for some users, and only the broad sub-groups of the possible routes were identifiable. This study is unique in the sense that it adapts the broad choice framework to the context of route choice modelling using noisy CDR data.

An examination of the parameter signs shows that CDR data is able to capture the expected sensitivities towards particular route attributes. A review of different models accounting for overlap was conducted, and among these, the c-logit and the path size logit models were considered. A comparison of these models against the multinomial logit model (which does not account for overlap) showed significant improvements in model fit, with the path size logit model giving the best performance. The validation runs based on the 20% holdout samples largely showed the same advantages in prediction, especially for the path size logit model. These results show that CDR is able to capture the expected behaviour towards overlapping routes.

This study is timely as it extends the application of CDR data beyond travel pattern analyses to econometric modelling of route choice that can be used for forecasting. The proposed framework can help in the assessment of different policy implications at a low cost compared to traditional approaches, which involve expensive data collection. For example, the models developed in this study can be used to reliably estimate the value of travel time (VTT) as we have demonstrated. This study is thus beneficial to developing countries where budget constraints on transport studies are common and traditional data for transport studies is scarce.

We conclude that the study findings serve as a proof-of-concept that CDR data can be successfully used to model route choice behaviour for long-distance inter-regional trips, where there is a strong possibility that a user will use his/her phone during the trip, thereby enabling the capture of their partial trajectories needed for route identification. It may be noted that with the increasing trend of mobile internet usage (Gerpott and Thomas, 2014), the temporal resolution of CDR locations is likely to improve significantly in the near future, and this could make CDR data suitable for evaluating route choice behaviour for short trips.

A comparison of the study findings with those based on traditional data from Senegal would have been insightful, however, this was not possible due to data unavailability. Investigating the performance of the proposed approach in urban or intra-city scenarios would be an interesting direction for future research.

References

Agard, B., Morency, C. & Trépanier, M. 2006. Mining public transport user behaviour from smart card data. IFAC Proceedings Volumes, 39, 399-404.

Ageroute Senegal. 2017. Programmes & Projets [Online]. Ageroute Senegal. Available: http://www.ageroute.sn/index.php/rapports-d-activités-annuels/cat_view.html [Accessed 12 April 2017].

Ahas, R., Aasa, A., Mark, Ü., Pae, T. & Kull, A. 2007. Seasonal tourism spaces in Estonia: Case study with mobile positioning data. Tourism management, 28**,** 898-910.

Ahas, R., Aasa, A., Roose, A., Mark, Ü. & Silm, S. 2008. Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. Tourism Management, 29**,** 469-486.

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

Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, Z., Liu, Y., Ziemlicki, C., Tiru, M. and Zook, M., 2015. Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. International Journal of Geographical Information Science, 29(11), pp.2017-2039.

Ansd 2016. Rapport projection de la population du Senegal. Dakar, Senegal: Agence Nationale de la Statistique et de la Démographie.

Ansd. 2017. Situation Economique et Sociale [Online]. Agence Nationale de la Statistique et de la Démographie. Available: http://www.ansd.sn/index.php?option=com_sess&view=sess&Itemid=398 [Accessed 11 April 2017].

Arcgis. 2013. Senegal Roads [Online]. Available: https://services1.arcgis.com/4AWkjqgSzd8pqxQA/arcgis/rest/services/Senegal_Roads/FeatureServer/0 [Accessed 01 March 2017].

Azevedo, J., Costa, M. E. O. S., Madeira, J. J. E. S. & Martins, E. Q. V. 1993. An algorithm for the ranking of shortest paths. European Journal of Operational Research, 69**,** 97-106.

Bellman, R. & Kalaba, R. 1960. On k th best policies. Journal of the Society for Industrial and Applied Mathematics, 8**,** 582-588.

Ben-Akiva, M., Bergman, M., Daly, A. J. & Ramaswamy, R. Modeling inter-urban route choice behaviour. Proceedings of the 9th International Symposium on Transportation and Traffic Theory, 1984. VNU Science Press Utrecht, The Netherlands, 299-330.

Ben-Akiva, M. & Ramming, S. 1998. Lecture notes: Discrete choice models of traveler behavior in networks. Prepared for Advanced Methods for Planning and Management of Transportation Networks. Capri, Italy, 25.

Ben-Akiva, M. E. & Lerman, S. R. 1985. Discrete choice analysis: theory and application to travel demand, MIT press.

Bierlaire, M., Chen, J. & Newman, J. 2010. Modeling route choice behavior from smartphone GPS data.

Bierlaire, M., Chen, J. & Newman, J. 2013. A probabilistic map matching method for smartphone GPS data. Transportation Research Part C: Emerging Technologies, 26**,** 78-98.

Bierlaire, M. & Frejinger, E. 2008. Route choice modeling with network-free data. Transportation Research Part C: Emerging Technologies, 16**,** 187-198.

Bolla, R., Davoli, F. & Giordano, F. Estimating road traffic parameters from mobile communications. Proceedings 7th World Congress on ITS, Turin, Italy, 2000.

Broach, J., Dill, J. & Gliebe, J. 2012. Where do cyclists ride? A route choice model developed with revealed preference GPS data. Transportation Research Part A: Policy and Practice, 46, 1730-1740.

Calabrese, F., Di Lorenzo, G., Liu, L. & Ratti, C. 2011. Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area. IEEE Pervasive Computing, 10, 36-44.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J. & Ratti, C. 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation research part C: emerging technologies, 26, 301-313.

Cascetta, E., Nuzzolo, A., Russo, F. & Vitetta, A. A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. Proceedings of the 13th International Symposium on Transportation and Traffic Theory, 1996. Pergamon Lyon, France, 697-711.

Cascetta, E. & Papola, A. 2001. Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand. Transportation Research Part C: Emerging Technologies, 9, 249-263.

Chakirov, A. & Erath, A. 2012. Activity identification and primary location modelling based on smart card payment data for public transport.

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.

Daganzo, C. F., Bouthelier, F. & Sheffi, Y. 1977. Multinomial probit and qualitative choice: A computationally efficient algorithm. Transportation Science, 11, 338-358.

Daly, A. 2010. Cost damping in travel demand models: Report of a study for the Department for Transport. United Kingdom: RAND Corporation.

De La Barra, T., Perez, B. & Anez, J. Multidimensional path search and assignment. PTRC Summer Annual Meeting, 21st, 1993, University of Manchester, United Kingdom, 1993.

De Montjoye, Y.-A., Smoreda, Z., Trinquart, R., Ziemlicki, C. & Blondel, V. D. 2014. D4D-Senegal: the second mobile phone data for development challenge. arXiv preprint arXiv:1407.4885.

Deville, P., Song, C., Eagle, N., Blondel, V. D., Barabási, A.-L. & Wang, D. 2016. Scaling identity connects human mobility and social interactions. Proceedings of the National Academy of Sciences, 201525443.

Diao, M., Zhu, Y., Ferreira Jr, J. & Ratti, C. 2016. Inferring individual daily activities from mobile phone traces: A Boston example. Environment and Planning B: Planning and Design, 43, 920-940.

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

Gainer, M. & Chan, S. 2016. A NEW ROUTE TO DEVELOPMENT: SENEGAL'S TOLL HIGHWAY PUBLIC-PRIVATE PARTNERSHIP, 2003 – 2013. New Jersey, USA: Innovations for Successful Socities, Priceton University.

Gerpott, T. J. & Thomas, S. 2014. Empirical research on mobile Internet usage: A meta-analysis of the literature. Telecommunications Policy, 38, 291-310.

Google Maps. 2017a. Google map directions [Online]. Google. Available: https://www.google.co.uk/maps/dir [Accessed 13 April 2017].

Google Maps. 2017b. Senegal nature reserves [Online]. Google. Available: https://www.google.co.uk/maps/@15.1978209,-15.0824015,8.67z [Accessed 28 December 2017].

Görnerup, O. Scalable Mining of Common Routes in Mobile Communication Network Traffic Data. Pervasive, 2012. Springer, 99-106.

Groves, R. M. 2006. Nonresponse rates and nonresponse bias in household surveys. Public opinion quarterly, 646-675.

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

Habibi, S., Frejinger, E. and Sundberg, M., 2017. An empirical study on aggregation of alternatives and its influence on prediction in car type choice models. Transportation, pp.1-20.

Hasan, S., Zhan, X. & Ukkusuri, S. V. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. Proceedings of the 2nd ACM SIGKDD international workshop on urban computing, 2013. ACM, 6.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P. & Ratti, C. 2014. Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science, 41, 260-271.

Hess, S., Quddus, M., Rieser-Schüssler, N. & Daly, A. 2015. Developing advanced route choice models for heavy goods vehicles using GPS data. Transportation Research Part E: Logistics and Transportation Review, 77, 29-44.

Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C. & Pujolle, G. 2014. Estimating human trajectories and hotspots through mobile phone data. Computer Networks, 64, 296-307.

Imedia & Calao Production. 2013. Le chemin de fer sénégalais [Online]. AU-SENEGAL.COM. Available: http://www.au-senegal.com/le-chemin-de-fer,345?lang=fr [Accessed 06 August 2017].

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.

Kerali, H. G. R. 2000. Overview of HDM-4, Paris, The World Road Association (PIARC), Paris and The World Bank, Washington, DC.

Kung, K. S., Greco, K., Sobolevsky, S. & Ratti, C. 2014. Exploring universal patterns in human home-work commuting from mobile phone data. PloS one, 9, e96180.

Leontiadis, I., Lima, A., Kwak, H., Stanojevic, R., Wetherall, D. & Papagiannaki, K. From cells to streets: Estimating mobile paths with cellular-side data. Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies, 2014. ACM, 121-132.

Li, L., Wang, S. & Wang, F.-Y. 2018. An Analysis of Taxi Driver's Route Choice Behavior Using the Trace Records. IEEE Transactions on Computational Social Systems, 5, 576-582.
Lloro, A. and Brownstone, D., 2018. Vehicle choice and utilization: Improving estimation with partially observed choices and hybrid pairs. Journal of choice modelling, 28, pp.137-152.
Logistics Cluster. 2013. 2.3 Senegal Road Assessment [Online]. Logistics Cluster and World Food Programme. Available: http://dlca.logcluster.org/display/public/DLCA/2.3+Senegal+Road+Assessment [Accessed 11 May 2017].
Mackie, P., Jara-Dıaz, S. & Fowkes, A. 2001. The value of travel time savings in evaluation. Transportation Research Part E: Logistics and Transportation Review, 37, 91-106.
Ma, J., Li, H., Yuan, F. and Bauer, T., 2013. Deriving operational origin-destination matrices from large scale mobile phone data. International Journal of Transportation Science and Technology, 2(3), pp.183-204.
Mai, T., Fosgerau, M. & Frejinger, E. 2015. A nested recursive logit model for route choice analysis. Transportation Research Part B: Methodological, 75, 100-112.
Manski, C. F. 1977. The structure of random utility models. Theory and decision, 8, 229-254.
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.
Nie, J., Zhang, J., Zhong, G. & Hu, Y. 2015. A Novel Approach to Road Matching Based on Cell Phone Handover. CICTP 2015.
Nielsen, O. A. 2000. A stochastic transit assignment model considering differences in passengers utility functions. Transportation Research Part B: Methodological, 34, 377-402.
Nitsche, P., Widhalm, P., Breuss, S., Brändle, N. & Maurer, P. 2014. Supporting large-scale travel surveys with smartphones–a practical approach. Transportation Research Part C: Emerging Technologies, 43, 212-221.
Nitsche, P., Widhalm, P., Breuss, S. & Maurer, P. 2012. A strategy on how to utilize smartphones for automatically reconstructing trips in travel surveys. Procedia-Social and Behavioral Sciences, 48, 1033-1046.
Papinski, D., Scott, D. M. & Doherty, S. T. 2009. Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. Transportation research part F: traffic psychology and behaviour, 12, 347-358.
Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T. & Zhou, C. 2014. A new insight into land use classification based on aggregated mobile phone data. International Journal of Geographical Information Science, 28, 1988-2007.
Prato, C. & Bekhor, S. 2006. Applying branch-and-bound technique to route choice set generation. Transportation Research Record: Journal of the Transportation Research Board, 19-28.
Prato, C. G. 2009. Route choice modeling: past, present and future research directions. Journal of choice modelling, 2, 65-100.

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.

Ramming, M. S. 2002. Network knowledge and route choice. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, USA.

Reddy, S., Burke, J., Estrin, D., Hansen, M. & Srivastava, M. Determining transportation mode on mobile phones. Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on, 2008. IEEE, 25-28.

Rouphail, N. M., Ranjithan, S. R., El Dessouki, W., Smith, T. & Brill, E. D. A decision support system for dynamic pre-trip route planning. Applications of Advanced Technologies in Transportation Engineering, 1995. ASCE, 325-329.

Saravanan, M., Pravinth, S. V. & Holla, P. Route detection and mobility based clustering. Internet Multimedia Systems Architecture and Application (IMSAA), IEEE 5th International Conference, 2011. IEEE, 1-7.

Sayers, M. W., Gillespie, T. D. & Queiroz, C. a. V. 1986. The international road roughness experiment: establishing correlation and a calibration standard for measurements. World Bank Technical Paper No. 45.

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

Schlaich, J., Otterstätter, T. & Friedrich, M. Generating trajectories from mobile phone data. Proceedings of the 89th annual meeting compendium of papers, transportation research board of the national academies, 2010.

Shaw, S.L., Tsou, M.H. and Ye, X., 2016. Human dynamics in the mobile and big data era. International Journal of Geographical Information Science, 30(9), pp.1687-1693.

Sheffi, Y. & Powell, W. B. 1982. An algorithm for the equilibrium assignment problem with random link times. Networks, 12**,** 191-207.

Shier, D. R. 1979. On algorithms for finding the k shortest paths in a network. Networks, 9**,** 195-214.

Song, C., Koren, T., Wang, P. & Barabási, A.-L. 2010. Modelling the scaling properties of human mobility. Nature Physics, 6**,** 818-823.

Tettamanti, T., Demeter, H. & Varga, I. 2012. Route choice estimation based on cellular signaling data. Acta Polytechnica Hungarica, 9**,** 207-220.

Tran, L.T.T., Ly, P.T.M. and Le, L.T., 2019. Hotel choice: A closer look at demographics and online ratings. International Journal of Hospitality Management, 82, pp.13-21.

Teye, C., Davidson, P., Porter, H. & Bell, M. G. H. 2017. Meta-analysis on the value of travel time savings in Africa. International Choice Modelling Conference 2017. Cape Town, South Africa.

Vovsha, P. & Bekhor, S. 1998. Link-nested logit model of route choice: overcoming route overlapping problem. Transportation Research Record: Journal of the Transportation Research Board**,** 133-142.

Vrtic, M., Schuessler, N., Erath, A., Axhausen, K., Frejinger, E., Stojanovic, J., Bierlaire, M., Rudel, R. & Maggi, R. 2006. Including travelling costs in the modelling of mobility behaviour. Final report for SVI research program Mobility Pricing: Project B1, on behalf of the Swiss Federal Department of the Environment, Transport, Energy and Communications. IVT ETH Zurich, ROSO EPF Lausanne and USI Lugano.

Wang, H., Calabrese, F., Di Lorenzo, G. & Ratti, C. Transportation mode inference from anonymized and aggregated mobile phone call detail records. Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on, 2010. IEEE, 318-323.

Watanatada, T., Harral, C., Paterson, W., Dhareshwar, A., Bhandari, A. & Tsunokawa, K. 1987. The Highway Design and Maintenance Standards Model. Volume 1 Description of the HDM-III Model. Baltimore and London: The John Hopkins University Press.

White, J. & Wells, I. 2002. Extracting origin destination information from mobile phone data.

Wong, T. C. J. 2015. Econometric Models in Transportation. Ph.D. Thesis, University of California, Irvine.

Wong, T., Brownstone, D. and Bunch, D.S., 2018. Aggregation biases in discrete choice models. Journal of Choice Modelling.

World Bank 2004. Performance and impact indicators for transport in Senegal: Detailed statistics - June 2004. World Bank.

Worldatlas. 2017. Senegal Facts [Online]. Worldatlas. Available: https://www.worldatlas.com/webimage/countrys/africa/senegal/snfacts.htm [Accessed 22 November 2017].

Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z. & Li, Q. 2015. Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. Transportation, 42**,** 625-646.

Yip, A.H., Michalek, J.J. and Whitefoot, K.S., 2018. On the implications of using composite vehicles in choice model prediction. Transportation Research Part B: Methodological, 116, pp.163-188.

Yuan, Y. & Raubal, M. Extracting dynamic urban mobility patterns from mobile phone data. International Conference on Geographic Information Science, 2012. Springer, 354-367.

Yuan, Y., Raubal, M. and Liu, Y., 2012. Correlating mobile phone usage and travel behavior–A case study of Harbin, China. Computers, Environment and Urban Systems, 36(2), pp.118-130.

Wang, Y., Dong, L., Liu, Y., Huang, Z. and Liu, Y., 2019. Migration patterns in China extracted from mobile positioning data. Habitat International, 86, pp.71-80.

Zhao, Z., Shaw, S.L., Xu, Y., Lu, F., Chen, J. and Yin, L., 2016. Understanding the bias of call detail records in human mobility research. International Journal of Geographical Information Science, 30(9), pp.1738-1762.