---

**Article:**
Bwambale, A, Choudhury, CF orcid.org/0000-0002-8886-8976 and Hess, S orcid.org/0000-0002-3650-2518 (2019) Modelling trip generation using mobile phone data: A latent demographics approach. Journal of Transport Geography, 76. pp. 276-286. ISSN 0966-6923

https://doi.org/10.1016/j.jtrangeo.2017.08.020

---

# MODELLING TRIP GENERATION USING MOBILE PHONE DATA: A LATENT DEMOGRAPHICS APPROACH

**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, Corresponding Author**
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

Initial Submission Date 24 February 2017
Resubmission Date 26 June 2017

**ABSTRACT**

Traditional approaches to trip generation modelling rely on household travel surveys which are expensive and prone to reporting errors. On the other hand, mobile phone data, where spatio-temporal trajectories of millions of users are passively recorded has recently emerged as a promising input for transport analyses. However, such data has primarily been used for the development of human mobility models, extraction of statistics on human mobility behaviour, and origin-destination matrix estimation as opposed to the development of econometric models of travel demand. This is primarily due to the exclusion of user demographics from mobile phone data made available for research (owing to privacy reasons). In this study, we address this limitation by proposing a hybrid trip generation model framework where demographic groups are treated as latent or unobserved. The proposed model first predicts the demographic group membership probabilities of individuals based on their phone usage characteristics and then uses these probabilities as weights inside a latent class model for trip generation, with different classes representing different socio-demographic groups. The model is calibrated using the call log data of a sub-sample of users with known demographics and trip rates extracted from their GSM mobility data. The performance of the hybrid model is compared with that of a traditional trip generation model which uses observed demographic variables to validate the proposed methodology. This comparative analysis shows that the model fit and the prediction results of the hybrid model are close to those of the traditional model. The research thus serves as a proof-of-concept that the mobile phone data can be successfully used to develop econometric models of transport planning by having additional information for a subset of the users.

*Keywords*: Trip generation, Mobile phone data, Demographic prediction

## 1 INTRODUCTION

Trip generation is the first step of the four stage model (Ortúzar and Willumsen, 2011) and is critical to the accuracy of the subsequent stages. Generally, trip generation models establish mathematical relationships between trip making rates and the demographics of individuals or households (e.g. Bwambale et al., 2015 and the cited references). Traditional approaches to estimating trip generation models rely on household travel surveys which are expensive and prone to reporting errors. Furthermore, the application of traditional models is often hindered by the lack of detailed demographic information in the application context.

Consequently, there has been growing interest in the use of ubiquitous data for mobility modelling. Examples include social media data (e.g. Hawelka et al., 2014, Hasan et al., 2013, Wu et al., 2014), smart card data (e.g. Agard et al., 2006, Chakirov and Erath, 2012), and mobile phone data (e.g. Çolak et al., 2015, Song et al., 2010). However among these, mobile phone data has emerged as the most promising source due to the high penetration rate of mobile phones. Unique subscriber penetration in the developed world is currently very high, estimated at 79% in 2014, and projected to grow to 81% by the end of 2020, while that in the developing world was estimated at 44.6% in 2014, and is projected to grow to 56% by the end of the same period (GSMA Intelligence, 2015).

Mobile phone records, which can consist of Call Detail Records[1] (CDRs) or Global System for Mobile Communications[2] (GSM) data, have been widely used to develop human mobility models (e.g. Gonzalez et al., 2008, Jiang et al., 2013, Çolak et al., 2015, Song et al., 2010, Deville et al., 2016, Isaacman et al., 2012), calibrate traffic models (e.g. Bolla et al., 2000), develop origin-destination matrices (e.g. Iqbal et al., 2014, White and Wells, 2002, Pan et al., 2006, Çolak et al., 2015), and estimate trip rates (e.g. Çolak et al., 2015). However, they have not been used in econometric models of travel demand like trip generation, mode choice, and route choice due to missing demographic information.

The inclusion of demographic attributes into travel demand models improves their behavioural underpinning, policy sensitivity, and forecasting potential and the lack of information on such attributes is thus a valid reason for the lack of applications of mobile phone passive data in travel demand models. However, while privacy regulations make it difficult to make a 1-1 link between the socio-demographic details of a user and his/her CDRs, previous studies have demonstrated that characteristics like age, gender, employment status can be predicted by analysing the phone usage behaviour derived from the CDRs of a sub-sample of known user (e.g. Blumenstock et al., 2010, Dong et al., 2014, Brdar et al., 2012, Aarthi et al., 2011, Ying et al., 2012, Mo et al., 2012). Such techniques can be used to incorporate demographic information into human mobility models based on mobile phone data, however, there is a need to evaluate the feasibility of such an approach as this has not been tested before.

In this study, we propose a novel hybrid trip generation modelling framework to make mobile phone data usable for developing econometric models of travel behaviour and demonstrate it in the context of trip generation models. The proposed hybrid trip generation model first predicts the demographic group membership probabilities of individuals as a function of their observed mobile

[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 has more detailed location compared to CDRs and reports the IDs of all the GSM cells traversed by an active mobile phone at regular time intervals.

phone usage. These probabilities are then used as weights inside a latent class model for trip generation, with different classes representing different socio-demographic groups.

The proposed model needs GSM locations, CDR, and the socio-demographics from a small sub-sample for estimation/calibration. However, once calibrated, it only needs *anonymous* CDR data to predict the trip rates. Given that CDR data is routinely saved by the mobile phone companies for billing purposes, the proposed model thus provides as a low-cost, yet accurate method for predicting trip rates – especially in the context of developing countries where traditional data is not available/reliable and acquiring large-scale GSM data is difficult (due to privacy concerns and requirement of very large storages).

We use the Nokia Mobile Data Challenge (MDC) dataset (Laurila et al., 2012, Kiukkonen et al., 2010), which is described later in this paper, to investigate the feasibility of the proposed hybrid trip generation model. We compare the goodness-of-fit of the hybrid model against that of a traditional model (which directly uses the observed demographics). We then conduct multiple runs of predictions to compare the accuracy of the trip rates predicted by the two models to validate our hypotheses that the proposed hybrid model, which only uses the predicted demographics from the CDR data, has the potential to substitute the traditional trip generation model with observed demographics.

The rest of the paper is arranged as follows. We start with a review of relevant literature, followed by an overview of the framework and the detailed model structure. We then provide a description of the data used for this study and the model estimation and validation results. Finally, we present a summary of the findings and the conclusions.

## 2 LITERATURE REVIEW

We start by reviewing the literature on demographic prediction followed by that on passive inferring of dwell regions and trip rate extraction from mobile phone data. We end with a brief review of mathematical models of trip generation.

### 2.1 Demographic prediction from mobile phone data

The earliest attempt to use CDRs for demographic prediction was made in Rwanda (Blumenstock et al., 2010). In this study, a logit model was estimated to predict the gender of users based on the net number of calls per day and the net call duration. The study used a sample of 901 users whose demographic information was obtained through phone interviews. The estimated logit model gave a prediction accuracy of 74%. Since then, logistic regression has been applied in other demographic prediction studies (e.g. Blumenstock, 2015, Mo et al., 2012). However, most other studies have used supervised learning classification algorithms for demographic prediction. Typically, these studies involve the training of various supervised learning classifiers (e.g. Support Vector Machines and Random Forests) to make separate predictions of the demographic attributes of users based on phone usage variables (e.g. Aarthi et al., 2011, Frias-Martinez et al., 2010, Brdar et al., 2012, Ying et al., 2012, Mo et al., 2012).

Following the observation that most of the studies above had focused on predicting demographic attributes in isolation, Dong et al. (2014) investigated the possibility of improving accuracy through simultaneous demographic attribute predictions. This was motivated by the hypothesis that mobile phone usage is influenced by a combination of demographic attributes and that separate

prediction of individual demographics would reduce the probability of success due to excluded attributes. They estimated a Double Dependent Variable Factor Graph Model capable of making joint age and gender predictions based on phone usage variables and found that this improved prediction accuracy by up to 10%.

### 2.2 Inferring dwell regions from mobile phone data

In trip generation modelling, it is also important to know the home, work and other dwell regions of individuals in order to distinguish trips by purpose (e.g. Ortúzar and Willumsen, 2011). Previous studies have developed spatio-temporal algorithms for passively detecting and labelling an individual's dwell locations using CDRs (e.g. Çolak et al., 2015, Pan et al., 2006, Jiang et al., 2013, Toole et al., 2015, Akin and Sisiopiku, 2002). The nature of such algorithms depends on the accuracy used to record the location of the communication events in the CDRs. Locations are usually recorded either as triangulated mobile phone coordinates, coordinates of the cell tower that transmitted the call or as the ID of the cell from which the call was made.

Where the CDRs contain triangulated coordinates, dwell locations have been detected by applying an upper limit (usually 300m) on the distance between consecutive mobile phone coordinates and a lower limit (usually 10 minutes) on the time difference between the first and last points of a potential dwell location (e.g. Çolak et al., 2015, Jiang et al., 2013, Toole et al., 2015). For each user, the centroids of different dwell locations in close proximity to each other are then clustered into dwell regions using different clustering approaches (for example grid-based clustering (Zheng et al., 2010)) since these could be referring to the same actual point.
Where the CDRs contain cell tower coordinates, dwell locations have been detected by linking a series of consecutive communication events transmitted by cell towers in close proximity to each other followed by linking a series of consecutive events transmitted by the same tower in order to distinguish between tower jumps and actual mobile phone movements (Çolak et al., 2015). This is because mobile operators sometimes carry out tower-to-tower balancing to optimize network performance. Dwell regions are then detected by applying a lower limit (usually 10 minutes) on the time difference between the first and last records in a series of consecutive events transmitted by the same tower. A similar approach is appropriate for CDRs containing cell IDs.

Irrespective of the type of CDRs, the extracted dwell regions for each user are labelled as home, work, or other depending on the detected visitation frequency between particular times of the day, for example, home and work locations are usually defined as the most commonly visited dwell regions at night and during daytime respectively while the rest are defined as others (e.g. Çolak et al., 2015, Jiang et al., 2013). The success of the methods described above depends on the phone usage frequency of the individuals in the sample and requires long observation periods. Nevertheless, the methods can be applied to large anonymous CDR datasets to infer the dwell regions of users during trip generation model application. Methods for assigning the inferred dwell regions to Traffic Analysis Zones have been developed to ensure consistency with the existing transport models (e.g. Çolak et al., 2015, Pan et al., 2006).

### 2.3 Extraction of trip rates from mobile phone data

Previous studies have made attempts to directly estimate trip generation from CDRs (e.g. Çolak et al., 2015). CDRs have the advantage of being readily available in large quantities, however, they only report locations when the mobile phone is in use (e.g. when calls are made) making them unable to capture movements when the phone is not in use. Çolak et al. (2015) attempted to address

this issue by making several assumptions e.g. by assuming an arbitrary home-based trip where the first or the last reported position of the day in the CDRs is at a non-home location. While these are reasonable assumptions, they do not properly address the issue of missed trips between communication events.

Ultimately, the best way to extract trip rates from mobile phone data is when continuous locations are provided. However, network operators usually discard such information due to its large size. Nevertheless, we note that it is feasible to store continuous location data for a reasonable sub-sample of users as was done during the Lausanne Data Collection Campaign where continuous GSM cell references were stored (Laurila et al., 2012, Kiukkonen et al., 2010).

### 2.4 Mathematical models of trip generation

Discrete choice models have been the preferred approach for modelling trip generation since the ground-breaking work of McFadden (1974). This is because trip generation levels are discrete, mutually exclusive and finite. However, trip generation levels are ordered choices. An individual cannot choose to make the $n^{th}$ trip if he/she has not previously made $(n-1)$ trips. The decision to make an additional trip depends on the number of trips already made which introduces inter-trip correlations. This has previously been taken into account using either Naturally Ordered Logit Choice Models (e.g. Vickerman and Barmby, 1985) or Ordered Response Choice Models (e.g. Bwambale et al., 2015), where the latter approach is more popular and is thus also used in this study. We also note that other trip generation modelling techniques e.g. linear regression and cross-classification (Ortúzar and Willumsen, 2011) are commonly used in practice, however, these are not considered for this study.

## 3 FRAMEWORK

The hybrid trip generation model uses a demographic prediction model to replace the observed demographics with probabilistic latent classes of socio-demographic groups. The estimation and application frameworks are presented in Figures 1a and 1b respectively.

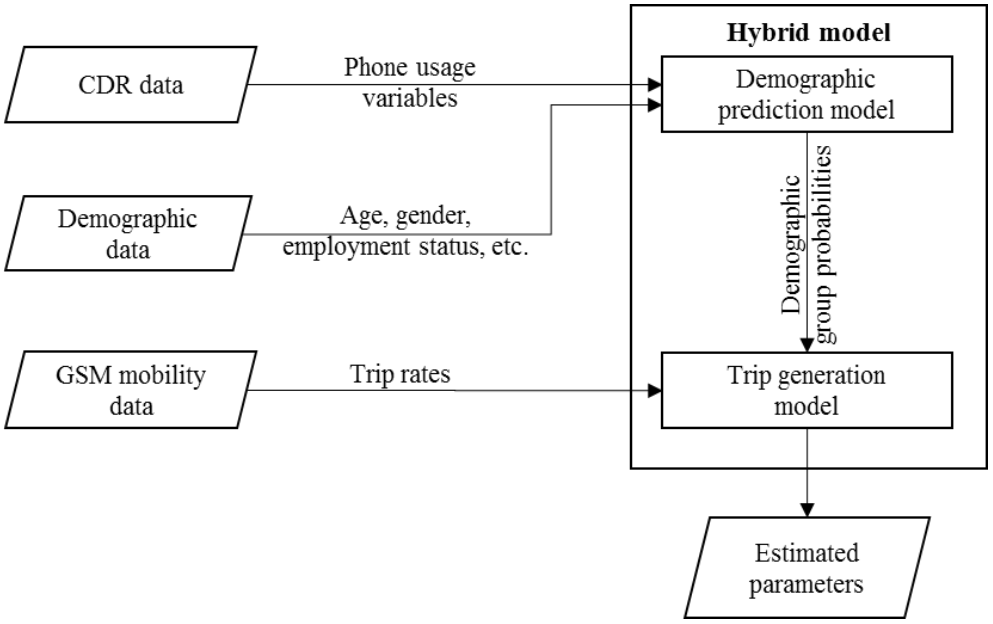
As presented in Figure 1a, the data used for estimating the hybrid trip generation model includes the GSM locations, the CDRs and the socio-demographic characteristics of a small sub-sample. The GSM data reports the IDs of all the GSM cells traversed by an active mobile phone at regular time intervals and reliably captures all the trips made by the different users, except some short trips made within the boundaries of the same GSM cell. It may be noted GSM data is commonly discarded by mobile network operators due to storage space constraints and hence though it can be stored for a small sub-sample, it is not typically available for the wider population.

On the other hand, CDR data, which is readily available, has a timestamped record of the phone usage activities and can be used to derive phone usage behaviour. It also records the ID of the tower that handles the call, but the location data has not been used in this case because of the availability of the GSM data which is more reliable.

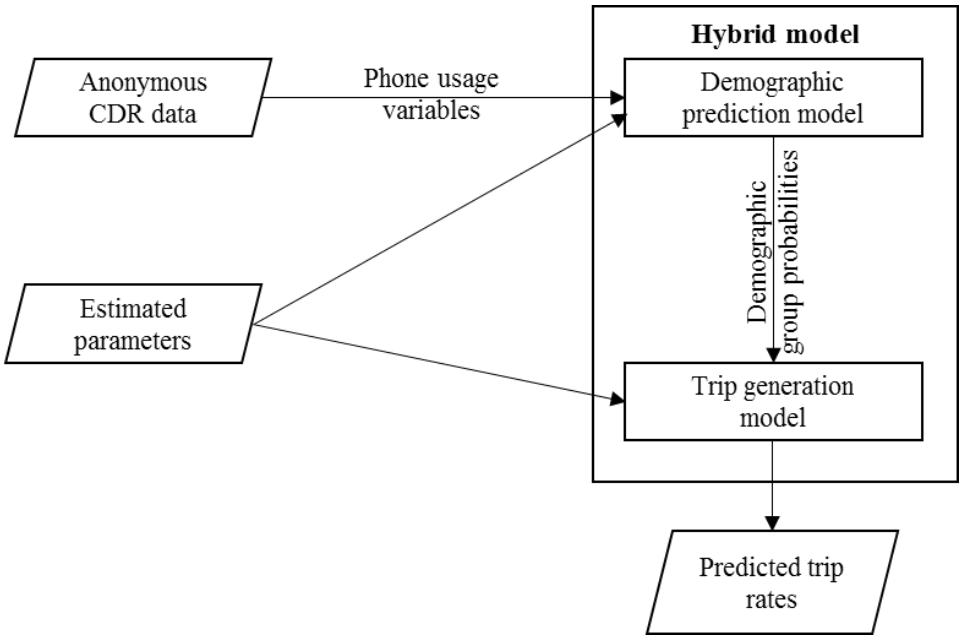
These data sources are used to calibrate the hybrid choice model which has two components:

1. Demographic prediction component
2. Latent class based trip generation component

In the demographic group prediction component, the demographic group membership probabilities of individuals are predicted as a function of their observed mobile phone usage (derived from CDR data). These probabilities are then used as weights inside a latent class model for trip generation, with different classes representing different socio-demographic groups. The trip rates used for calibrating the proposed hybrid model are extracted from the GSM mobility data.




**a) Estimation framework (sub sample of users)**




**b) Application framework (for all users)**

**FIGURE 1 Overall framework**

In the application stage (Figure 1b), the pre-estimated hybrid model uses the anonymous CDR data as the input, predicts the latent socio-demographic classes of the users using the CDR data and predicts the trip rates by feeding these latent classes to the pre-estimated trip generation model. Therefore, the socio-demographic information and the GSM data are not required in the application stage.

The detailed structure of the hybrid model is presented in Figure 2 and described in the following section.

## 4 MODEL STRUCTURE

To implement the proposed hybrid framework in Figure 1, we propose an expanded approach that integrates two types of discrete choice mechanisms, that is, the unordered-response choice mechanism (for demographic prediction) and the ordered-response choice mechanism (for trip generation) (Ben-Akiva and Lerman, 1985).

### 4.1 Demographic prediction

The proposed demographic prediction model is based on the Random Utility Theory (Marschak, 1960). We use the phone usage data to explain which socio-demographic group a given individual falls into. To do this, we assume that individuals in a particular demographic group are more likely to be associated with specific mobile phone usage behaviour. We use random utility theory by assuming that the segment that a given respondent falls into has the highest utility as a function of the observed phone usage behaviour.

Let $U_{ns}$ be the utility of individual $n$ falling in demographic group $s$ as a function of mobile phone usage behaviour. This can be expressed as;

$$\begin{aligned} \mathrm{U}_{\mathrm{ns}} &= \beta_{\mathrm{s}}' \mathrm{x}_{\mathrm{n}} + \xi_{\mathrm{ns}} \\ &= \beta_{\mathrm{s}}' \mathrm{x}_{\mathrm{n}} + \left(\eta' \mathrm{z}_{\mathrm{ng}} + \psi' \mathrm{h}_{\mathrm{na}} + \lambda' \mathrm{m}_{\mathrm{nw}} + \varepsilon_{\mathrm{ns}}\right) \end{aligned} \quad (1)$$

Where $x$ is a vector of observed phone usage variables; $\beta_s$ is a vector of group-specific parameters; and $\xi_{ns}$ is the random component of utility. As shown, the random term comprises of the error term $\varepsilon_{ns}$ and three demographic attribute specific components, one along each dimension of the demographic groups. $\eta'$ is the gender specific constant while $z_{ng}$ is a vector of dummy variables for the gender dimension. The additional terms $\psi'$ and $h_{\mathrm{na}}$; and $\lambda'$ and $\mathrm{m}_{\mathrm{nw}}$ are defined in a similar way to $\eta'$ and $z_{ng}$ but in the context of the age-group and the working status dimensions. The demographic attribute specific constants account for the unobserved phone usage dynamics that are shared across different demographic groups sharing one or more demographic attribute.

The phone usage variables are respondent specific and thus constant across the 'alternatives', which are the demographic groups. Each group has a different set of parameters associated with it, reflecting the fact that the amount of usage has a differential impact on the likelihood of falling into a given group.

We make an assumption that the error term is independently and identically distributed across the alternatives and use the Multinomial Logit (MNL) Model (McFadden, 1974) to estimate the demographic group membership probabilities as expressed below.

$$P_{ns} = \frac{\exp\left(\beta'_s x_n + \eta' z_{ng} + \psi' h_{na} + \lambda' m_{nw}\right)}{\sum_{s^*} \exp\left(\beta'_{s^*} x_n + \eta' z_{g^*} + \psi' h_{a^*} + \lambda' m_{w^*}\right)} \quad (2)$$

The model parameters are then estimated by maximising the log-likelihood function below.

$$LL(\beta'_s) = \sum_{n=1}^{N} \sum_{s=1}^{S} Z_{ns} \ln(P_{ns}) \quad (3)$$

Where $Z_{ns} = 1$ if and only if individual $n$ belongs to demographic group $s$ otherwise, $Z_{ns} = 0$. As a result, each respondent has a non-zero probability of falling into each of the different socio-demographic groups, but the more the model is able to link the socio-demographic characteristics to phone usage, the more deterministic the allocation to these groups becomes in the model.

**4.2 Trip generation**

As mentioned, the ordered response choice mechanism is used in the trip generation model component. This mechanism assumes that every individual has a latent trip making propensity which is a function of their demographics (Ben-Akiva and Lerman, 1985). This propensity is then converted to discrete trips using estimated cut-off points. We first present the traditional framework (where demographics are observed) and then present our proposed extension that addresses the issue of unobserved demographics.

The traditional trip generation model (with observed demographics)

Let $h_n^*$ be the latent trip-making propensity for individual $n$ based on his observed demographic attributes. Using the ordered-response choice mechanism, this can be expressed as;

$$h^*_n = \gamma' w_n + \varepsilon_n \quad (4)$$

$$t = \begin{cases} <10, & \text{if } h_n^* \le \delta_0 \\ 10 - 15, & \text{if } \delta_0 < h_n^* \le \delta_1 \\ 16 - 20, & \text{if } \delta_1 < h_n^* \le \delta_2 \\ 21 - 25, & \text{if } \delta_2 < h_n^* \le \delta_3 \\ >25, & \text{if } h_n^* > \delta_3 \end{cases}$$

Where; $w_n$ is a vector of the observed demographic attributes for individual $n$ ; $\varepsilon_n$ is the random error term; $\gamma'$ is a vector of the model coefficients; $t$ is the number of trips per week; and $\delta_0 < \delta_1 < \delta_2 < \delta_3$ are the cut-off points. Note that different categorisations of the weekly trip rates were tested and these were found to provide the best model fit.

We make an assumption that the random error term follows a Gumbel Distribution and use the Ordered Response Logit (ORL) Model (Ben-Akiva and Lerman, 1985) to estimate the trip generation probabilities as expressed below;

$$
\begin{aligned}
&P_{n,\ t<10} = \Lambda(\delta_0 - \gamma' w_n) \\
&P_{n,\ 10 \le t \le 15} = \Lambda(\delta_1 - \gamma' w_n) - \Lambda(\delta_0 - \gamma' w_n) \\
&P_{n,\ 16 \le t \le 20} = \Lambda(\delta_2 - \gamma' w_n) - \Lambda(\delta_1 - \gamma' w_n) \\
&P_{n,\ 21 \le t \le 25} = \Lambda(\delta_3 - \gamma' w_n) - \Lambda(\delta_2 - \gamma' w_n) \\
&P_{n,t>25} = 1 - \Lambda(\delta_3 - \gamma' w_n)
\end{aligned} \tag{5}
$$

Where $\Lambda(q) = \exp[-\exp(-q)]$ represents the standard cumulative Gumbel Distribution. The model parameters are then estimated by maximising the log-likelihood function below.

$$
LL(\gamma, \delta) = \sum_{n=1}^{N} \sum_{t=0}^{t=3+} Z_{nt} \ln(P_{nt}) \tag{6}
$$

The hybrid trip generation model (with predicted demographics)
Let $y^*_{n|s}$ be the latent trip-making propensity for individual $n$ on condition that he/she belongs to latent demographic group $s$. This latent propensity can be expressed as a function of the typical demographic attributes associated with latent demographic group $s$ as shown below;

$$
y^*_{n|s} = \alpha' w_{n|s} + \varepsilon_{n|s} \tag{7}
$$

$$
t = \begin{cases}
< 10, & \text{if } y^*_{n|s} \le \mu_0 \\
10 - 15, & \text{if } \mu_0 < y^*_{n|s} \le \mu_1 \\
16 - 20, & \text{if } \mu_1 < y^*_{n|s} \le \mu_2 \\
21 - 25, & \text{if } \mu_2 < y^*_{n|s} \le \mu_3 \\
> 25, & \text{if } y^*_{n|s} > \mu_3
\end{cases}
$$

Where; $w_{n|s}$ is a vector of the typical demographic attributes for individual $n$ given that he/she is associated with latent demographic group $s$; $\varepsilon_{n|s}$ is the random error term; $t$ is the number of trips; $\mu_0 < \mu_1 < \mu_2 < \mu_3$ are the cut-off points; and $\alpha'$ is a vector of the model coefficients.

We again assume that the random error term follows a Gumbel Distribution and estimate the conditional trip generation probabilities as expressed below;

$$
\begin{aligned}
&P_{n,\ (t<10)\,|\,s} = \Lambda(\mu_0 - \alpha' w_{n|s}) \\
&P_{n,\ (10 \le t \le 15)\,|\,s} = \Lambda(\mu_1 - \alpha' w_{n|s}) - \Lambda(\mu_0 - \alpha' w_{n|s}) \\
&P_{n,\ (16 \le t \le 20)\,|\,s} = \Lambda(\mu_2 - \alpha' w_{n|s}) - \Lambda(\mu_1 - \alpha' w_{n|s}) \\
&P_{n,\ (21 \le t \le 25)\,|\,s} = \Lambda(\mu_3 - \alpha' w_{n|s}) - \Lambda(\mu_2 - \alpha' w_{n|s}) \\
&P_{n,\ (t>25)\,|\,s} = 1 - \Lambda(\mu_3 - \alpha' w_{n|s})
\end{aligned} \tag{8}
$$

These calculations are conditional on knowing the socio-demographics of respondent *n*, reflected by that respondent falling into demographic group *s*. However in reality, we do not know which class the respondent falls into. Therefore, the unconditional trip generation probabilities are

estimated as the weighted averages of the conditional probabilities as expressed in Equation 9. The weights $P_{ns}$ are the demographic group membership probabilities estimated from Equation 2 at the maximum likelihood estimates.

$$
\begin{aligned}
P_{n,\,t<10} &= \sum_{s=1}^{S} P_{ns} \bullet \Lambda(\mu_0 - \alpha' w_{n|s}) \\
P_{n,\,10 \le t \le 15} &= \sum_{s=1}^{S} P_{ns} \bullet \left[\Lambda(\mu_1 - \alpha' w_{n|s}) - \Lambda(\mu_0 - \alpha' w_{n|s})\right] \\
P_{n,\,16 \le t \le 20} &= \sum_{s=1}^{S} P_{ns} \bullet \left[\Lambda(\mu_2 - \alpha' w_{n|s}) - \Lambda(\mu_1 - \alpha' w_{n|s})\right] \\
P_{n,\,21 \le t \le 25} &= \sum_{s=1}^{S} P_{ns} \bullet \left[\Lambda(\mu_3 - \alpha' w_{n|s}) - \Lambda(\mu_2 - \alpha' w_{n|s})\right] \\
P_{n,\,t>25} &= \sum_{s=1}^{S} P_{ns} \bullet \left[1 - \Lambda(\mu_3 - \alpha' w_{n|s})\right]
\end{aligned}
\tag{9}
$$

Figure 2 presents the full path diagram of the hybrid model structure where unobserved variables are shown in circles and observed variables in rectangles.

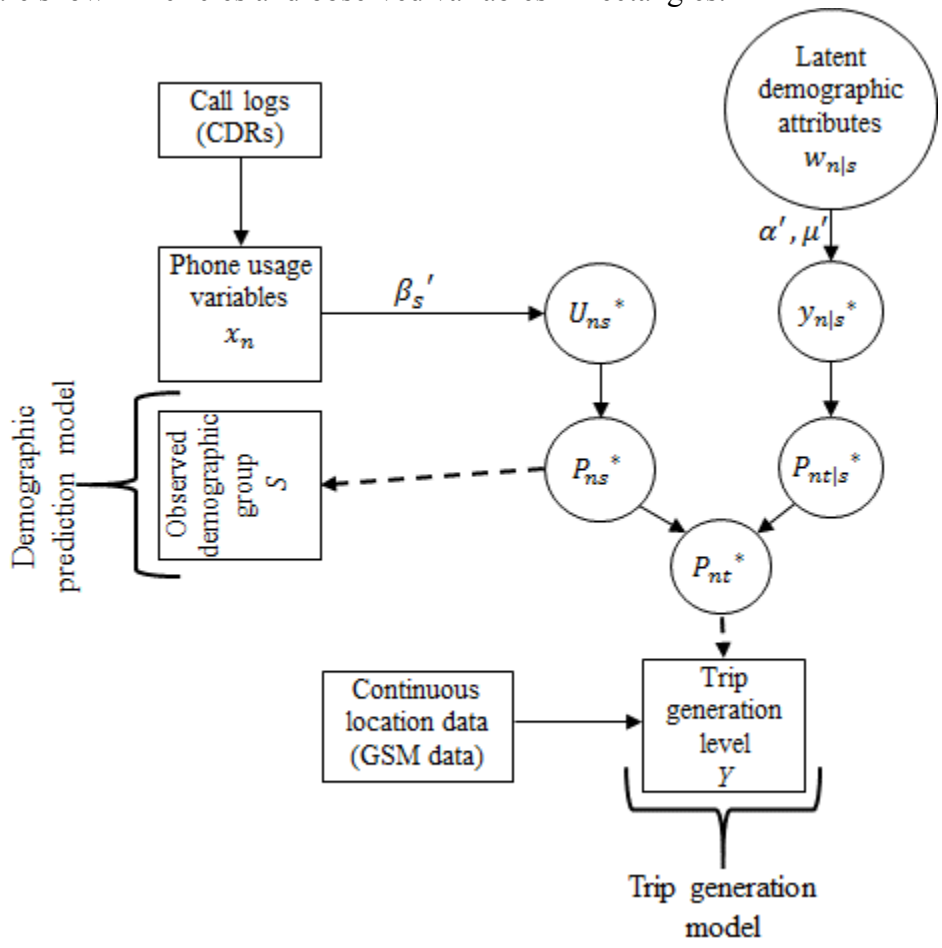



**FIGURE 2 Full path diagram of the hybrid model structure**

***Notation***

| | |
|---|---|
| $\alpha, \beta, \mu$ | Vectors of unknown parameters to be estimated |
| $X$ | Phone usage variables |
| $S$ | Demographic groups |
| $t$ | Number of trips |
| $U_{ns}^{*}$ | Utility of individual *n* falling in demographic group *s* |
| $P_{ns}^{*}$ | Membership probability to demographic group *s* for individual *n* |
| $w_{n\|s}$ | A vector of the typical demographic attributes for individual *n* given that he is associated with latent demographic group *s* |
| $y_{n\|s}^{*}$ | Latent trip-making propensity for individual *n* on condition that he belongs to latent demographic group is *s* |
| $P_{nt\|s}^{*}$ | Conditional probability for making *t* trips given that the latent demographic group for individual *n* is *s* |
| $P_{nt}^{*}$ | Unconditional probability for making *t* trips for individual *n* |
| $Y$ | Number of trips made |

It may be noted that two sequential estimators are used to estimate the hybrid framework. The first results in parameters that provide the best fit for the demographic prediction model and the second results in parameters that provide the best fit for the trip generation model. This is different from a simultaneous estimator which tries to jointly predict both the demographic groups and the trip generation levels. A simultaneous model could lead to gains in efficiency (i.e. smaller standard errors) but also opens up risk in terms of confounding between the two model components.

### 4.3 Evaluation criteria for model performance

For model evaluation, we compare the goodness-of-fit during estimation and validation. For the estimation, we use the adjusted-rho square and the likelihood ratio test (Ben-Akiva and Lerman, 1985) which are defined as follows, respectively;

$$\rho_{adj}^{2} = 1 - \frac{LL(F) - k}{LL(0)} \qquad \text{and} \qquad LR = -2[LL(0) - LL(F)] \tag{10}$$

Where; $k$ is the number of model parameters, $LL(F)$ and $LL(0)$ are the values of the log-likelihood function at convergence and at zero respectively.

For model validation, a hold-out sample (not used for model estimation) is used to confirm that the estimation results are not simply due to overfitting. In this stage, we use both aggregate and disaggregate measures of fit. At the aggregate level, we compare the predicted and actual shares and compute the Root Mean Square Error (RMSE). At the disaggregate level, we use the predictive rho-square and the average probability of correct prediction. The predictive rho-square is obtained by calculating the log-likelihood for the validation sample at the pre-estimated maximum likelihood parameters and at zero and then applying Equation 10 without the $k$. The average probability of correct prediction is obtained by computing the mean probability of success for the validation sample based on the pre-estimated maximum likelihood parameters.

## 5 DATA

We use data from the Nokia Mobile Data Challenge (MDC) for this study (Laurila et al., 2012, Kiukkonen et al., 2010). The data was generated during the Lausanne Data Collection Campaign by 158 participants with known demographics. These participated in the campaign at different time periods between 2010 and 2012, each lasting several months. This makes the data rich in terms of temporal coverage. The full database contains several types of smartphone records (e.g. Bluetooth usage data), however, we only use the call logs and the GSM cells data (mobility data) to improve the transferability of our approach. The subsequent sections briefly describe the data used including the analysis undertaken.

### 5.1 Extraction of demographic groups from the demographic data

The demographic data file contains the demographics of 158 participants. Each record in this file is described by; a user ID, the gender, the age-group, and the working status of the participant, among others (e.g. marital status). Out of these, 4 participants were disregarded because they had missing demographic information, leaving 154 participants. Demographic groups were formed by generating various possible combinations of age-group, gender, and working status. In total, seven demographic groups were observed in the data as shown in Table 1, where some of the demographic groups have very small sub-samples. This problem could have been avoided by conducting demographically stratified random sampling of the participants in the data collection phase (which was beyond our control).

### 5.2 Extraction of phone usage variables from the call log data

The call log data file contains a register of all the communication events of the participants (calls and short messages). In total, there are over 0.42 million call log events. Each call log event is described by; a user ID, the time of the call, the status of sent short messages, the direction of the call, the type of call, the other party's anonymized phone number, and the call duration. The information in this file is equivalent to what would be found in CDRs. The call log data was analysed to extract several phone usage variables based on guidance from previous literature (e.g. Aarthi et al., 2011, Blumenstock, 2015, Blumenstock et al., 2010, Frias-Martinez et al., 2010) and intuition. Table 1 presents the summary statistics of the extracted phone usage variables.

### 5.3 Extraction of trip rates from the GSM (mobility) data

The GSM data file contains a register of all the GSM cells seen by the participants' mobile phones at an interval of approximately 60 seconds. This data file contains over 50.8 million records generated by all the participants. Each GSM record is described by; a user ID, a unique internal ID for the GSM cell, and the record creation time and date. The GSM data was analysed to extract the number of trip origins from home using the following approach.

First, all the GSM cell IDs seen at night (between 8 pm and 6 am) by the different user IDs were extracted and ordered according to the record creation time and date. For each date, the GSM cell ID seen for the longest continuous time at night was established and the most common among these across the different dates determined as the home GSM cell for the user ID. The weekly trip rates from home were then estimated by analysing the GSM mobility data to determine the number of times per week the different user IDs were not seen in their respective home GSM cells for periods longer than 10 minutes. We considered 10 minutes as the appropriate threshold for distinguishing between actual trips and tower jumps. We do not classify the trips by purpose because the geographical locations of the GSM cells have been anonymised thereby making it difficult to infer activities by map matching. We acknowledge that the resolution of the GSM

mobility data only enables the capture of trips made outside the home GSM cell and misses short trips made within the boundaries of the home GSM cell. Our approach is therefore suitable for urban areas such as Lausanne where GSM cell sizes can be as small as 100m (De Groote, 2005).

At this stage, we disregarded 11 participants who had incomplete weeks of data, leaving 143 participants. Table 1 presents the summary statistics of the extracted trip rates.

**TABLE 1 Summary statistics**

*Demographic group summary statistics*

| Demographic group | Assigned code | Number of participants | Percentage, % |
|---|---|---|---|
| Female non-worker aged below 21 years | F-NW-U21 | 7 | 4.9 |
| Female worker aged above 21 years | F-WO-A21 | 29 | 20.3 |
| Female non-worker aged above 21 years | F-NW-A21 | 19 | 13.3 |
| Male non-worker aged below 21 years | M-NW-U21 | 4 | 2.8 |
| Male worker aged below 21 years | M-WO-U21 | 3 | 2.1 |
| Male non-worker aged above 21 years | M-NW-A21 | 22 | 15.4 |
| Male worker aged above 21 years | M-WO-A21 | 59 | 41.3 |
| **Total** | | **143** | **100** |

*Sample phone usage summary statistics (extracted from call log data)*

| Variable | Statistic |
|---|---|
| Average number of outgoing calls per user, per day | 3.4 |
| Average number of incoming calls per user, per day | 1.5 |
| Average number of outgoing short messages per user, per day | 1.7 |
| Average number of incoming short messages per user, per day | 2.5 |
| Average number of missed calls per user, per day | 0.7 |

*Trip generation summary statistics (extracted from GSM data)*

| Number of trips per week from home | Number of participants | Percentage, % |
|---|---|---|
| < 10 | 4 | 2.8 |
| 10 – 15 | 26 | 18.2 |
| 16 – 20 | 43 | 30.1 |
| 21 – 25 | 14 | 9.8 |
| > 25 | 56 | 39.2 |
| **Total** | **143** | **100** |

## 6 ESTIMATION RESULTS

In this section, we present the estimation results for both the demographic group prediction model and the trip generation models based on the full sample.

### 6.1 Demographic group prediction model

Table 2 presents the estimation results of the demographic prediction model. We tested various combinations of phone usage variables in terms of the statistical performance of the associated parameters and the overall model performance and settled for a set of eleven shown in Table 2. We found that differentiation of phone usage by time segment (e.g. working hours and night) was statistically important for most of the variables while differentiation by weekdays versus weekends

was not. We also found that interacting some of the variables (e.g. net = outgoing - incoming) was statistically important, however, we acknowledge that we have not exhausted all the possibilities.

The parameters of the demographic prediction model represent the effect of the variables on the utility of each demographic group relative to that of the reference group M-WO-A21 (male workers aged above 21 years). We do not have a priori expectations of the parameter signs since this is still a new area of research, moreover, mobile phone usage behaviour is likely to differ from place to place. Therefore, we analyse this particular case using our intuitive reasoning. To do this, we first analyse the demographic attribute specific constants. Among these, we find that the only statistically significant constant is that associated with individuals above 21 years. This indicates the existence of statistically strong unobserved phone usage dynamics common across different demographic groups sharing the same age-group. The rest of the constants are statistically insignificant probably because the associated phone usage dynamics have been captured by the specified explanatory variables.

We then analyse the parameters of the demographic groups having only one attribute not in the reference group M-WO-A21 so as to establish the unique effect of each attribute. These groups (and the complement attributes) are; F-WO-A21 (female), M-WO-U21 (age below 21 years), and M-NW-A21 (non-worker). See Table 1 for the group definitions.

From Table 2, it is observed that the net number of calls and outgoing short messages during working hours, the number of outgoing calls and the total call duration (outgoing and incoming) in the evening, and the social network indegree (the unique number of incoming contacts) have positive parameter signs for the F-WO-A21 group. However among these, the only statistically significant parameter is that for the net number of calls during working hours. This suggests that females in comparison to males tend to use their phones more during working hours. On the other hand, the net number of calls in the morning and at night, the number of outgoing short messages at night, the number of outgoing calls during lunch time, and the social network outdegree (the unique number of outgoing contacts) have negative parameter signs for the same group. Most of these parameters are statistically significant except those for the outgoing short messages at night, and the social network outdegree. This suggests that males in comparison to females tend to make more phone calls during non-working hours since majority of them are workers.

Similarly, it is observed that the number of outgoing calls and the total call duration (outgoing and incoming) in the evening, the net number of calls at night, and the number of outgoing short messages during working hours have positive parameter signs for the M-NW-A21 group. However, the only statistically significant parameter that for the number of outgoing calls in the evening. This points to the idea that workers in comparison to non-workers tend to call fewer people in the evenings probably because they prefer to utilize this time preparing for the next day. On the other hand, the net number of calls in the morning and during working hours, the number of outgoing calls at lunch, the number of outgoing short messages at night, and the social network indegree and outdegree have negative parameter signs for the M-NW-A21 group. However, the only statistically significant parameter (at the 90% confidence level) is that for the number of outgoing short messages at night. This implies that workers in comparison to non-workers tend to send out more short messages at night probably because they do not have time to do so during the day due to work.

**TABLE 2 Parameter estimates of the demographic prediction model (See Table 1 for the parameter definitions)**

| Variable | Parameter | t-statistic |
|---|---|---|
| **Net number of calls (outgoing – incoming) in the morning (06:00 AM – 08:00 AM)** | | |
| F-NW-U21 | -2.9394 | -0.82 |
| F-WO-A21 | -3.6428 | -1.92 |
| F-NW-A21 | -0.9796 | -0.32 |
| M-NW-U21 | -40.3761 | -3.69 |
| M-NW-A21 | -0.6865 | -0.32 |
| M-WO-U21 | -18.6114 | -3.77 |
| **Number of outgoing calls at lunch time (01:00 PM – 02:00 PM)** | | |
| F-NW-U21 | -7.0073 | -1.94 |
| F-WO-A21 | -9.6277 | -2.91 |
| F-NW-A21 | 1.2208 | 0.60 |
| M-NW-U21 | -6.0815 | -1.47 |
| M-NW-A21 | -0.2364 | -0.10 |
| M-WO-U21 | 4.5906 | 1.90 |
| **Net number of calls (outgoing – incoming) during working hours (08:00 AM – 01:00 PM and 02:00 PM – 05:00 PM)** | | |
| F-NW-U21 | 1.9960 | 3.15 |
| F-WO-A21 | 2.4266 | 3.30 |
| F-NW-A21 | -0.2648 | -0.34 |
| M-NW-U21 | 1.9878 | 1.99 |
| M-NW-A21 | -0.4622 | -0.64 |
| M-WO-U21 | 2.0136 | 3.61 |
| **Number of outgoing calls in the evening (05:00 PM – 08:00 PM)** | | |
| F-NW-U21 | -1.3946 | -0.99 |
| F-WO-A21 | 0.5150 | 0.43 |
| F-NW-A21 | 1.4971 | 1.61 |
| M-NW-U21 | 0.8424 | 0.26 |
| M-NW-A21 | 2.4243 | 2.54 |
| M-WO-U21 | 0.3342 | 0.24 |
| **Net number of calls (outgoing – incoming) at night (08:00 PM – 06:00 AM)** | | |
| F-NW-U21 | -0.2181 | -0.11 |
| F-WO-A21 | -4.8651 | -1.99 |
| F-NW-A21 | -2.3378 | -1.01 |
| M-NW-U21 | 2.4107 | 0.56 |
| M-NW-A21 | 0.7189 | 0.31 |
| M-WO-U21 | 0.9151 | 0.31 |
| **Number of outgoing short messages during working hours (08:00 AM – 01:00 PM and 02:00 PM – 05:00 PM)** | | |
| F-NW-U21 | 1.6257 | 2.64 |
| F-WO-A21 | 0.7478 | 1.21 |
| F-NW-A21 | 0.0456 | 0.08 |
| M-NW-U21 | 1.9497 | 2.7 |
| M-NW-A21 | 1.2729 | 1.47 |
| M-WO-U21 | -1.4161 | -1.97 |

**TABLE 2 continued**

| Variable | Parameter | t-statistic |
|---|---|---|
| **Number of outgoing short messages at night (08:00 PM – 06:00 AM)** | | |
| F-NW-U21 | -1.5717 | -1.65 |
| F-WO-A21 | -1.2529 | -0.99 |
| F-NW-A21 | 0.5355 | 0.61 |
| M-NW-U21 | -1.166 | -1.04 |
| M-NW-A21 | -3.3759 | -1.70 |
| M-WO-U21 | 2.1387 | 2.30 |
| **Average duration of outgoing calls in the evening (05:00 PM – 08:00 PM)** | | |
| F-NW-U21 | 0.0032 | 0.93 |
| F-WO-A21 | 0.0020 | 0.76 |
| F-NW-A21 | 0.0001 | 0.05 |
| M-NW-U21 | 0.0516 | 3.04 |
| M-NW-A21 | 0.0005 | 0.16 |
| M-WO-U21 | -0.0028 | -0.94 |
| **Average duration of incoming calls in the evening (05:00 PM – 08:00 PM)** | | |
| F-NW-U21 | 0.0017 | 1.08 |
| F-WO-A21 | 0.0008 | 0.57 |
| F-NW-A21 | 0.0013 | 1.01 |
| M-NW-U21 | -0.0897 | -2.94 |
| M-NW-A21 | 0.0002 | 0.15 |
| M-WO-U21 | 0.0018 | 0.86 |
| **Outdegree of the social network** | | |
| F-NW-U21 | -0.0360 | -1.89 |
| F-WO-A21 | -0.0063 | -0.72 |
| F-NW-A21 | 0.0016 | 0.18 |
| M-NW-U21 | 0.1579 | 2.24 |
| M-NW-A21 | -0.0076 | -0.68 |
| M-WO-U21 | 0.0083 | 0.57 |
| **Indegree of the social network** | | |
| F-NW-U21 | 0.0238 | 1.25 |
| F-WO-A21 | 0.0037 | 0.45 |
| F-NW-A21 | -0.0163 | -1.48 |
| M-NW-U21 | -0.1633 | -2.00 |
| M-NW-A21 | -0.0088 | -0.78 |
| M-WO-U21 | -0.0357 | -1.62 |
| **Demographic attribute specific constants** | | |
| Males | 0.0401 | 0.08 |
| Workers | -0.0205 | -0.03 |
| Individuals > 21 years | 2.1583 | 1.86 |
| **Measures of fit** | | |
| Number of observations | 143 | |
| Log-likelihood at zero | -278.27 | |
| Log-likelihood at convergence | -169.83 | |
| Number of parameters | 69 | |
| Adjusted-rho square | 0.14 | |
| Likelihood ratio | 216.9 | |
| Chi-square statistic (69, 0.05) | 89.39 | |

In addition, it is observed that the number of outgoing calls during lunch time and in the evening, the net number of calls during working hours and at night, the number of outgoing short messages at night, the average duration of incoming calls in the evening, and the social network outdegree have positive parameter signs for the M-WO-U21 group. However, the only statistically significant parameters among these are those for the net number of calls during working hours, the number of outgoing calls during lunch time, and the number of outgoing short messages at night. This is a reflection of the possibility that individuals aged below 21 years tend to make more phone calls while at work and during lunch breaks and also send more short messages at night. On the other hand, the net number of calls in the morning, the number of outgoing short messages during working hours, the average duration of outgoing calls in the evening, and the social network indegree have negative parameter signs for the same group. However, the only parameters that are statistically significant among these are those for the net number of calls in the morning and the number of outgoing short messages during working hours. This indicates that individuals aged below 21 years in comparison to those above 21 years tend to make fewer calls in the morning and send fewer short messages during working hours (because they already make more calls during this time as earlier noted).

The last rows of Table 2 provide the measures of fit in estimation. From these, it is noted that the model passes the likelihood ratio test at the 95% confidence level in comparison with a model giving equal probabilities to the different groups for each individual (Ben-Akiva and Lerman, 1985).

### 6.2 Trip generation models

The variables commonly used in trip generation models include; income, car ownership, working status, age, and gender (e.g. Bwambale et al., 2015). Among these, income and car ownership were not available in the MDC dataset and hence could not be considered in the demographic prediction model, therefore, we only considered gender, working status, and age.

The estimation results of the hybrid model (which uses the predicted demographics) are presented in Table 3 alongside those of a model which uses the observed demographics (referred as the traditional model in the following sections).

The parameters of these models indicate the effect of the variables on the trip making propensity of individuals. The signs of all the parameters are consistent with a priori expectations. When individuals are employed, they are always engaged at the workplace and tend to travel less frequently to and from home in comparison to non-workers, hence the negative parameter sign for workers. Similarly, females generally run more errands (e.g. shopping, taking children to school etc.) irrespective of who pays the costs. Therefore, females tend to travel more frequently to and from home in comparison to males, hence the negative parameter sign for males. On the other hand, individuals above 21 years are generally out of school and if not already employed, are usually in active search for employment opportunities which requires a lot of travel hence the positive parameter sign.

The signs of all the parameters for both models are the same and the t-statistics for the differences between the parameters are insignificant. This shows that both models capture the same trip generation behaviour and would lead to similar policy conclusions.

The last rows of Table 3 provide the measures of fit in estimation. The adjusted-rho square values and the final log-likelihoods show that the traditional model performs slightly better than the hybrid model. This is to be expected given the error free measures of the socio-demographics used in the traditional model. On the other hand, it is also worth acknowledging that part of the performance of the hybrid model could be due to allowing for heterogeneity through the probabilistic component as individuals are not assigned deterministically to classes.

**TABLE 3 Parameter estimates of the trip generation models**

| Variable | Traditional model (With observed demographics) | | Hybrid model (With predicted demographics) | | t-statistic for the difference between the parameters |
|---|---|---|---|---|---|
| | Parameter | t-statistic | Parameter | t-statistic | |
| **Dummies specific to** | | | | | |
| Males | -0.3981 | -2.33 | -0.9125 | -2.90 | -1.43 |
| Workers | -0.3962 | -2.24 | -0.7817 | -1.65 | -0.76 |
| Individuals > 21 years | 0.6020 | 1.91 | 1.0127 | 1.80 | 0.64 |
| **Cut-off points specific to** | | | | | |
| Trips per week <10 | -1.2775 | -4.19 | -1.6997 | -3.82 | -0.78 |
| Trips per week 10 - 15 | -0.4221 | -1.35 | -0.7087 | -1.69 | -0.55 |
| Trips per week 16 – 20 | 0.4647 | 1.44 | 0.3119 | 0.73 | -0.29 |
| Trips per week 21 - 25 | 0.7726 | 2.34 | 0.6542 | 1.49 | -0.22 |
| **Measures of fit in the estimation sample** | | | | | |
| No. of observations | 143 | | 143 | | |
| Zero log-likelihood | -230.2 | | -230.2 | | |
| Sample shares log-likelihood | -197.07 | | -197.07 | | |
| Final log-likelihood | -191.8 | | -192.3 | | |
| Number of parameters | 3 | | 3 | | |
| Adjusted-rho square | 0.14 | | 0.13 | | |
| Likelihood ratio w.r.t sample shares | 10.54 | | 9.54 | | |
| Chi-square stat (3, 0.05) | 7.81 | | 7.81 | | |

## 7 VALIDATION RESULTS

In order to compare the predictive power of the traditional and the proposed hybrid model, we randomly split the data into five parts at the person level and generated five rolling subsets, each comprising of 80% of the data for model estimation purposes. For each of the five estimation subsets, we generated a complementary subset comprising of 20% of the data for validation purposes.

We estimated models based on each of the five estimation subsets and the general interpretation of the model results remains the same. Table 4 presents the measures of fit of the models based on each of the subsets. As can be observed, the final log-likelihoods of the hybrid model remain close to those of the traditional model across the different subsets of the data.

We tested the predictive power of each of these models using the corresponding complementary subsets. The subsequent sections present the validation results of both the demographic group prediction model and the trip generation models.

**TABLE 4 Final log-likelihoods of the models on the estimation subsets**

| Model | Log-likelihood | Subset 1 *N=115* | Subset 2 *N=114* | Subset 3 *N=114* | Subset 4 *N=114* | Subset 5 *N=115* |
|---|---|---|---|---|---|---|
| Demographic prediction model | Initial | -223.78 | -221.83 | -221.83 | -221.83 | -223.78 |
| | Final | -121.21 | -130.74 | -135.37 | -120.48 | -122.64 |
| Hybrid trip generation model (With predicted demographics) | Initial | -185.09 | -183.48 | -183.48 | -183.48 | -185.09 |
| | Final | -153.66 | -151.90 | -155.14 | -155.32 | -154.68 |
| Traditional trip generation model (With observed demographics) | Initial | -185.09 | -183.48 | -183.48 | -183.48 | -185.09 |
| | Final | -152.34 | -151.23 | -154.50 | -154.14 | -152.30 |

### 7.1 Demographic group prediction

We start by assessing the predictive performance of the demographic prediction model using the five validation subsets. The actual and predicted demographic group shares in the validation subsets are presented in Figure 3.

As can be observed, both the actual and predicted shares tend to follow a similar trend albeit with observable differences across all the five subsets. This is probably due to weaknesses in variable specification, however, there is a possibility that more sophisticated models (e.g. the mixed logit model) could improve the performance. Nevertheless, the similarity in trends is a good starting point and motivates further research to improve our approach.

### 7.2 Trip generation

In this section, we assess the predictive performance of both the traditional and the hybrid trip generation models using the five validation subsets. The actual and predicted trip generation shares in the validation subsets are presented in Figure 4.

As can be observed, both the actual and the predicted shares tend to follow a similar trend for both models albeit with observable differences across all the five subsets. The difference between the actual and predicted shares for both models is probably due to the use of weak explanatory variables. As mentioned, previous trip generation studies have shown that income and car ownership are some of the most important explanatory variables (e.g. Bwambale et al., 2015) and yet these were not considered in this study.

The predictive measures of fit for both models were computed and are presented in Table 5. At the aggregate level, the hybrid model performs better than the traditional model in three out of the five subsets in terms of the root mean square error. At the disaggregate level, the hybrid model performs better than the traditional model in four out of the five subsets in terms of the average probability of correct prediction, and the predictive rho-square. As mentioned earlier, the relatively better performance of the hybrid model could be in part due to allowing for heterogeneity through the probabilistic component.

Nevertheless, these results prove that the proposed hybrid model is a feasible alternative to the traditional model particularly where other reliable data sources are absent, thereby supporting the use of predicted demographics.

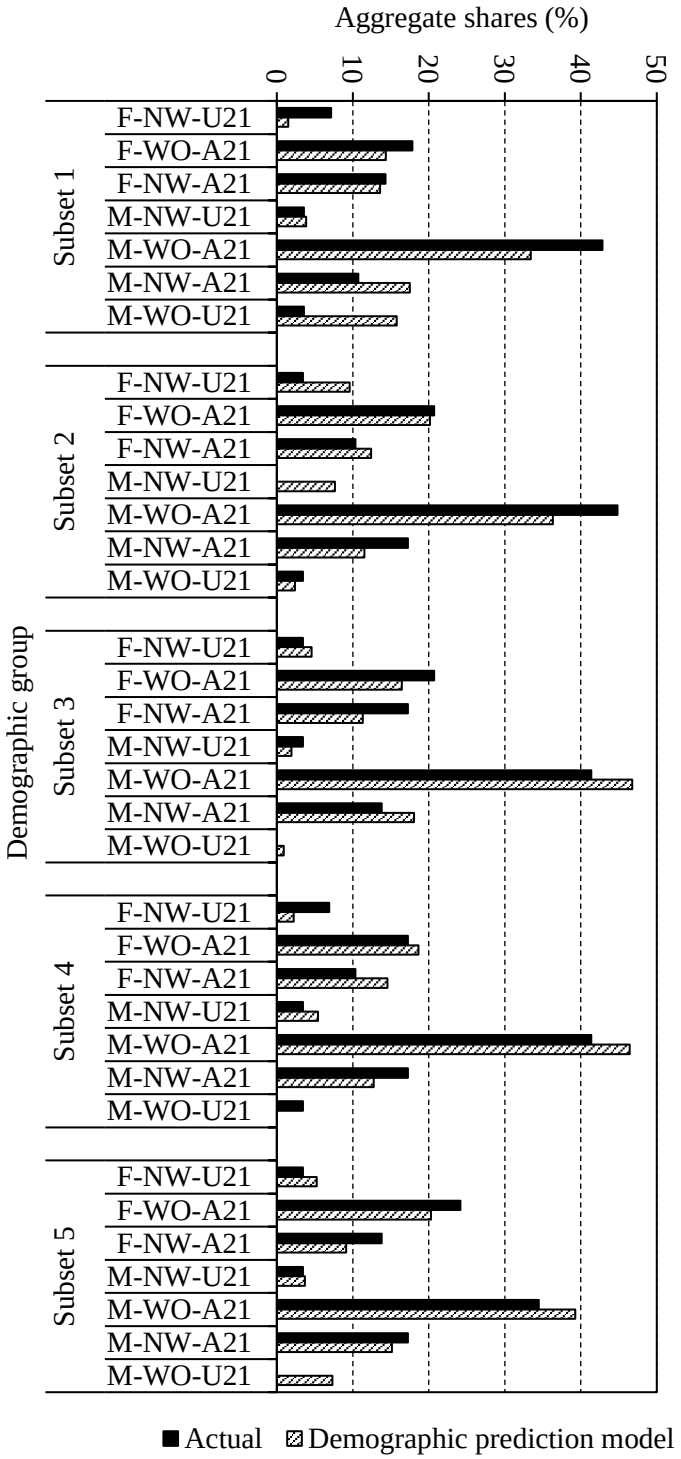



**FIGURE 3 Demographic model predictive performance in the validation subsets (See Table 1 for the demographic group definitions)**

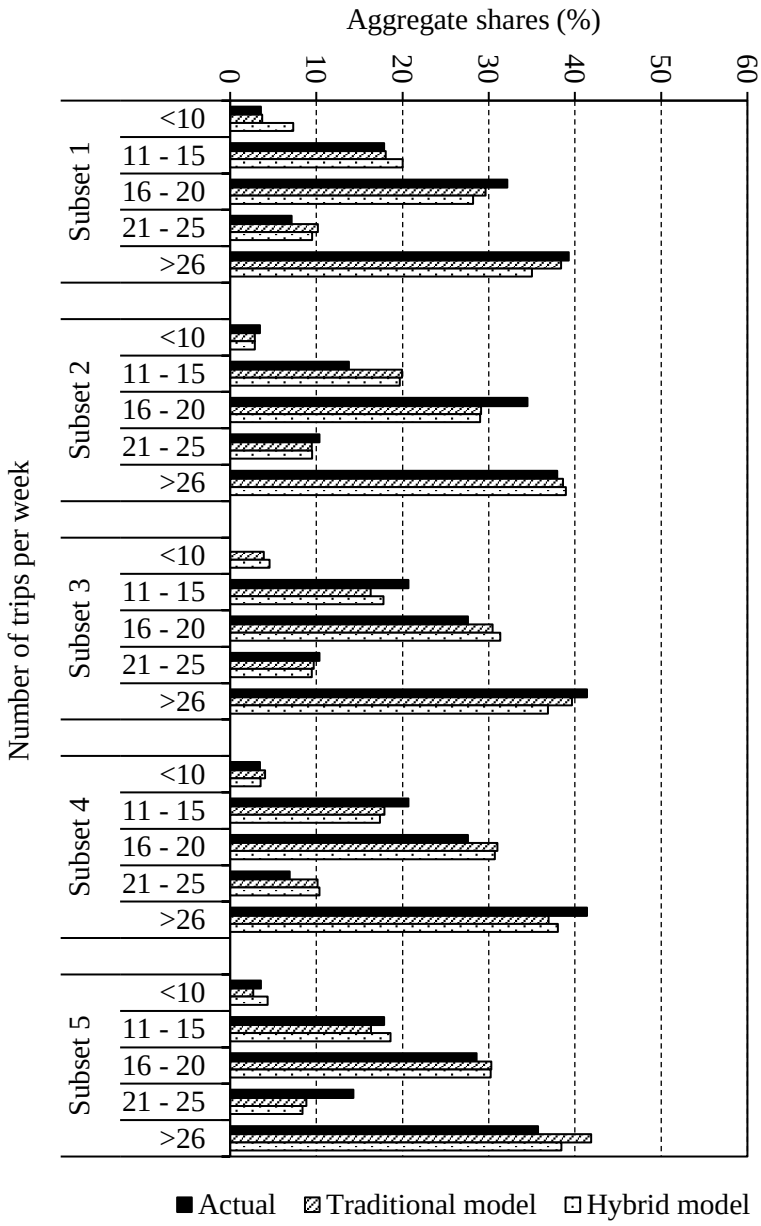



**FIGURE 4 Trip generation model predictive performance in the validation subsets**

**TABLE 5 Trip generation model measures of fit in the validation subsets**

| Validation subset | Root Mean Square Error | | Average probability of correct prediction | | Predictive rho-square | |
|---|---|---|---|---|---|---|
| | Traditional model | Hybrid model | Traditional model | Hybrid model | Traditional model | Hybrid model |
| Subset 1 | 1.81 | 3.41 | 0.273 | 0.287 | 0.102 | 0.121 |
| Subset 2 | 3.69 | 3.65 | 0.287 | 0.301 | 0.122 | 0.132 |
| Subset 3 | 3.04 | 3.59 | 0.302 | 0.304 | 0.190 | 0.202 |
| Subset 4 | 3.16 | 2.97 | 0.303 | 0.279 | 0.171 | 0.129 |
| Subset 5 | 3.84 | 3.03 | 0.280 | 0.282 | 0.107 | 0.120 |

## 8 SUMMARY AND CONCLUSIONS

The paper demonstrates the feasibility of the hybrid framework to mitigate the challenges associated with the estimation and the application of trip generation models using mobile phone data. An examination of the parameter signs and the t-statistics for the differences between the parameters of the hybrid trip generation model with predicted demographics and a traditional model (with observed demographics) shows that both models capture the same trip generation behaviour, an indication that both models would lead to similar policy conclusions.

We also assess the performance of the traditional and the hybrid trip generation models using several measures of fit in five estimation and validation samples. For the estimation samples, we compare the final log-likelihoods while for the validation samples, we compare the root mean square error values (the predicted and actual shares), the predictive rho-square values, and the average probabilities of correct prediction. We find that the traditional model performs slightly better than the hybrid model during estimation and attribute this to the error free measures of the socio-demographic variables in the traditional model with observed demographics. However, we find that the hybrid model generally performs better than the traditional model during validation in terms of the root mean square error values, the predictive rho-square values, and the average probabilities of correct prediction. We attribute this improved performance to the possibility that the hybrid model allows for heterogeneity through the probabilistic component.

For demographic prediction, we find that the performance of the model is satisfactory. However, this being a secondary data set, there are limitations in the sample size and distribution that are beyond our control. For example, we note that some demographic groups have very small sub-sample sizes which could have affected the overall model performance. We therefore recommend further research into different ways of improving the demographic prediction component of the hybrid model by dedicated data collection efforts.

In practice, the proposed hybrid framework could be used where one has the demographic information, call detail records, and GSM mobility data for just a small representative section of willing users for the purposes of model calibration and anonymous CDR data for the full population. We note that GSM mobility data is generally discarded by mobile phone operators due to storage space constraints, however, it is possible to store such data for a small sub-sample of willing users. Once calibrated, the model only needs the phone usage characteristics of the individuals to be implemented and these can be derived from the anonymous CDRs of the entire population. The model can thus be applied for planning purposes, particularly where other reliable data sources are absent.

We conclude that the validation results serve as a proof-of-concept that having the demographics of a sub-sample of willing mobile phone users can make mobile phone data feasible for econometric travel behaviour modelling and travel demand estimation. Further, the proposed hybrid framework has promise in improving the modelling of the other stages of the 4 step model (e.g. mode choice, route choice, etc.) using mobile phone data by enriching them with probabilistic latent socio-demographic classes in the absence of observed ones.

**REFERENCES**

AARTHI, S., BHARANIDHARAN, S., SARAVANAN, M. & ANAND, V. Predicting customer demographics in a mobile social network. Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on, 2011. IEEE, 553-554.

AGARD, B., MORENCY, C. & TRÉPANIER, M. 2006. Mining public transport user behaviour from smart card data. *IFAC Proceedings Volumes,* 39**,** 399-404.

AKIN, D. & SISIOPIKU, V. P. Estimating origin–destination matrices using location information from cell phones. Proc. 49th Annual North American Meetings of The Regional Science Association Int, Puerto Rico, 2002.

BEN-AKIVA, M. E. & LERMAN, S. R. 1985. *Discrete choice analysis: theory and application to travel demand*, MIT press.

BLUMENSTOCK, J. E. 2015. Calling for better measurement: Estimating an individual's wealth and well-being from mobile phone transaction records.

BLUMENSTOCK, J. E., GILLICK, D. & EAGLE, N. 2010. Who's calling? Demographics of mobile phone use in Rwanda. *Transportation,* 32**,** 2-5.

BOLLA, R., DAVOLI, F. & GIORDANO, F. Estimating road traffic parameters from mobile communications. Proceedings 7th World Congress on ITS, Turin, Italy, 2000.

BRDAR, S., CULIBRK, D. & CRNOJEVIC, V. Demographic attributes prediction on the real-world mobile data. Proc. Mobile Data Challenge by Nokia Workshop, in Conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK, 2012.

BWAMBALE, A., CHOUDHURY, C. F. & SANKO, N. Modelling Car Trip Generation in the Developing World: The Tale of Two Cities. Transportation Research Board 94th Annual Meeting, 2015.

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.

DE GROOTE, A. 2005. GSM Positioning Control. *University of Fribourg, Switzerland***,** 13.

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.

DONG, Y., YANG, Y., TANG, J., YANG, Y. & CHAWLA, N. V. Inferring user demographics and social strategies in mobile social networks. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014. ACM, 15-24.

FRIAS-MARTINEZ, V., FRIAS-MARTINEZ, E. & OLIVER, N. A gender-centric analysis of calling behavior in a developing economy. AAAI Symposium on Aritificial Intelligence and Development, 2010.

GONZALEZ, M. C., HIDALGO, C. A. & BARABASI, A.-L. 2008. Understanding individual human mobility patterns. *Nature,* 453**,** 779-782.

GSMA INTELLIGENCE. 2015. *The Mobile Economy 2015* [Online]. Available: http://www.gsmamobileeconomy.com/GSMA_Global_Mobile_Economy_Report_2015.pdf [Accessed 26 July 2016.

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.

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*.

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.

MO, K., TAN, B., ZHONG, E. & YANG, Q. Report of task 3: your phone understands you. Nokia mobile data challenge 2012 workshop, Newcastle, UK, 2012. Citeseer, 18-19.

ORTÚZAR, J. D. D. & WILLUMSEN, L. G. 2011. *Modelling transport*, John Wiley & Sons.

PAN, C., LU, J., DI, S. & RAN, B. 2006. Cellular-based data-extracting method for trip distribution. *Transportation Research Record: Journal of the Transportation Research Board***,** 33-39.

SONG, C., KOREN, T., WANG, P. & BARABÁSI, A.-L. 2010. Modelling the scaling properties of human mobility. *Nature Physics,* 6**,** 818-823.

TOOLE, J. L., COLAK, S., STURT, B., ALEXANDER, L. P., EVSUKOFF, A. & GONZÁLEZ, M. C. 2015. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*.

VICKERMAN, R. & BARMBY, T. 1985. Household trip generation choice—Alternative empirical approaches. *Transportation Research Part B: Methodological,* 19**,** 471-479.

WHITE, J. & WELLS, I. 2002. Extracting origin destination information from mobile phone data.

WU, L., ZHI, Y., SUI, Z. & LIU, Y. 2014. Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PloS one,* 9**,** e97010.

YING, J. J.-C., CHANG, Y.-J., HUANG, C.-M. & TSENG, V. S. 2012. Demographic prediction based on users mobile behaviors. *Mobile Data Challenge*.

ZHENG, V. W., ZHENG, Y., XIE, X. & YANG, Q. Collaborative location and activity recommendations with gps history data. Proceedings of the 19th international conference on World wide web, 2010. ACM, 1029-1038.