Getting the best of both worlds - a framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling

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

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

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

Md. Shahadat Iqbal
Lehman Centre for Transportation Research
Department of Civil and Environmental Engineering
Florida International University
10555 W. Flagler Street, EC 3729, Miami, FL 33174
Email: miqba005@fiu.edu

Submission Date 12 September 2019~~12 March 2019~~

Abstract

Traditional approaches to travel behaviour modelling primarily rely on household travel survey data, which is expensive to collect, resulting in small sample sizes and infrequent updates. Furthermore, such data is prone to reporting errors which can lead to biased parameter estimates and subsequently incorrect predictions. On the other hand, mobile phone call detail records (CDRs), which report the timestamped locations of mobile communication events, have been successfully used in the context of generating travel patterns. However, due to their anonymous nature, such records have not been widely used in developing mathematical models establishing the relationship between the observed travel behaviour and influencing factors such as the attributes of the alternatives and the decision makers. In this paper, we propose a joint modelling framework that utilises the advantages offered by both travel survey data and low-cost CDR data to optimise the prediction capacity of traditional trip generation models. In this regard, we develop a model that jointly explains the reported trips for each individual in the household survey data and ensures that the aggregated zonal trip productions are close to those derived from CDR data. This framework is tested using data from Dhaka. Bangladesh consisting of household survey data (65419 persons in 16750 households), mobile phone CDR data (over 600 million records generated by 6.9 million users), and aggregate census data. The model results show that the proposed framework improves the spatial and temporal transferability of the joint models over the base model which relies on household travel survey data alone. This serves as a proof-of-concept that augmenting travel survey data with mobile phone data holds significant promise for the travel behaviour modelling community, not only by saving the cost of data collection, but also improving the prediction capability of the models.

Keywords: Trip generation, CDR data, mobile phone data, household travel survey data, census data, population synthesis, transferability, Bangladesh, developing country

1 Introduction

Traditional approaches to developing travel behaviour models rely on household travel surveys to establish the mathematical relationship between the choices made by the travellers, the attributes of the network and socio-demographics characteristics of the travellers.. However, household surveys are often affected by low response rates and reporting errors (e.g. Rolstad et al., 2011, Groves, 2006). Further, the surveys are expensive to conduct which leads to small sample sizes and lower update frequencies. Consequently, transport models designed to fit household travel survey data alone can result in biased parameters capturing the noise in the data rather than the actual relationships in the population.

On the other hand, there has been growing interest in the use of mobile phone data for mobility modelling over the last few decades. Among the various transport related applications, such data has been widely used to estimate origin-destination matrices (e.g. Çolak et al., 2015, Iqbal et al., 2014, Pan et al., 2006, White and Wells, 2002) and trip generation (e.g. Çolak et al., 2015). Since mobile phone data generally covers significant proportions of the population (GSM Association, 2017), the data is able to reliably capture the aggregate travel patterns. However, due to its anonymous nature, mobile phone data is not traditionally used in developing mathematical models of travel behaviour that establish the relationship between observed travel behaviour and causal factors such as the attributes of the alternatives and the decision makers. The existing mobility models based on mobile phone data alone cannot be used to reliably test alternative or future travel demand scenarios, and yet this is one of the core roles of transport models.

We are thus in a situation where traditional survey data is small in size, potentially unrepresentative and inaccurate, but contains information on key causal variables. On the other hand, mobile phone data is larger in size, more representative and accurate but missing information on key causal variables. This situation motivates the present research where we propose a framework that brings in a third type of data, namely census information, which is representative and contains detailed socio-demographic variables but does not have travel behaviour information. We thus combine household travel survey data, aggregate census data, and mobile phone data to jointly optimise the aggregate and the disaggregate fit of travel behaviour models. The framework is calibrated and tested in the context of trip generation models.

In the context of trip generation, the traditional models based on household survey data establish the mathematical relationship between the number of trips made by an individual or household with the socio-demographics (see Bwambale et al., 2015 and the cited references). But the household survey data is prone to under-reporting of the number of trips (e.g. Zhao et al. 2015, Stopher et al. 2007, Itsubo and Hato 2006). Aggregating models based only on household survey data for estimating the zonal travel patterns can lead to errors, with serious consequences for the different steps of the four-stage model. This prompts us to investigate various ways of adjusting the parameter scales of the traditional trip generation model by using a joint optimisation process to combine it with the trip patterns derived from the mobile phone data.

The proposed joint modelling framework is both sequential and simultaneous. The approach is sequential in the sense that a base trip generation model is first estimated using household travel survey data alone to obtain the parameter priors (i.e. the sensitivities). The approach is simultaneous in the sense that when the parameter scales are being adjusted (without changing the prior parameter signs), the model jointly explains the reported trips for each individual in the household survey data and ensures that the aggregated zonal trip productions are close to those derived from CDR data. This ensures that the joint model does not lose the travel behaviour sensitivities reflected in the household survey data and is computationally tractable.

The rest of the paper is organised as follows, section 2 presents a brief review of the literature, section 3 presents the data used in this study, section 4 presents the modelling framework, section 5 presents the model results, and section 6 presents the summary and conclusions of the study.

2 Literature review

This section presents a brief review of the literature on related work in applying mobile phone data to mobility studies, as well as an overview of different population synthesis techniques.

2.1 Related studies on mobile phone data and population synthesis

The availability of large-scale mobile phone data over the last few decades has motivated a lot of research in quantifying human mobility and activity patterns using synthetic data generation methods (e.g. Chen et al., 2014).

From an epidemiology perspective, Vogel et al. (2015) combined CDR data with synthetic populations to model the spread of Ebola in West African countries and obtained promising results with respect to the Ebola predictions of the Centre for Disease Control and Prevention (CDC). Still in West Africa, Cárcamo et al. (2017) developed an intelligent epidemiology simulation software based on synthetic populations comprised of agents with realistic travel behaviour derived from CDR data. In France, Panigutti et al. (2017) compared the spread of a simulated epidemic using CDR and census survey travel patterns, finding greater similarity in areas with high population and connectivity, potentially due to the higher calling rates.

In the field of transport, Zilske and Nagel (2014) generated artificial CDR data from synthetic passengers in a simulated traffic scenario and re-used the data to approximate the amount of missed traffic at different calling rates to quantify the error introduced by CDR location discontinuities. The study found that the errors were inversely proportional to the calling rates and proposed scaling procedures based on observed data such as traffic counts. This led to a subsequent study where simulated CDR data and a synthetic population were combined with link traffic counts to generate all-day trip chains (Zilske and Nagel, 2015). An interesting outcome of this study was that even highly biased CDR data could reasonably reproduce the traffic state across different time periods. The approach of using observed traffic counts to scale CDR data has also been tested in Dhaka in the context of transient origin-destination (OD) matrix estimation (Iqbal et al., 2014).

Still in the field of transport, population synthesis has been applied on real-world mobile phone datasets. Ros and Albertos (2016) developed an improved version of MATSim (an agent-based multi-simulation software) by fusing census and CDR data from Spain to generate synthetic populations with mobility patterns observed in the CDR data. It may be noted that in this particular case, the mobile operator also provided the age and the gender of the users, which ensured a reliable dependence structure between the travel patterns and socio-demographics in the final synthetic population. However, mobile phone data is usually anonymous, which makes direct socio-demographic linkage impossible. In our earlier work (Bwambale et al., 2017), we developed a demographic group prediction model based on mobile phone usage behaviour extracted from CDR data (as part of a latent class model for trip generation), and can potentially be used for generating synthetic populations, however, this also requires a sub-sample of CDR data with known demographics, which is rarely available.

Kressner (2017) combined consumer and anonymous mobile phone data (wireless signalling and GPS data) from the United States to generate synthetic individual-level trip diaries. The socio-demographics in the disaggregate consumer data were benchmarked against the marginal census totals, while the synthetic travel was benchmarked against the mobility patterns extracted from the aggregate mobile phone data of several operators. A related study was conducted by Zhanga et al. (2017) in the context of social networks in urban simulations. Although these approaches perform quite well in terms of aggregate-level validation, the disaggregate dependency structure in the data seems arbitrary.

To maintain the underlying dependence structure, Janzen et al. (2017) combined household travel survey data, register data (national statistics) and CDR data from France to correct the under-reporting of long-distance trips in travel surveys using population synthesis techniques. The socio-demographics in the travel survey data were matched against those in the register data, while the reported long-distance trips in the travel survey data were matched against those derived from the CDR data. However, a potential issue with this approach is that it assumes uniform under-reporting for all the respondents in the travel survey data, and yet this might vary, at least across different demographic groups, with some cases of over-reporting. Furthermore, the assumed higher reliability of CDR data versus travel survey data is contentious and needs to be approached impartially. This is why we propose an optimisation approach between the two datasets.

2.2 Existing methods of population synthesis

Population synthesis is widely applied in activity-based models, and various techniques have been proposed to do this. This section presents a brief review of these methods.

The most widely applied technique is iterative proportional fitting (IPF), which works by fitting a contingency table based on disaggregate survey data to the marginal totals in aggregate census data, constrained by a set of control variables (Beckman et al., 1996). Since its development, various improvements based on the original concept have been proposed to enhance its applicability to new challenges. These improvements have mainly focussed on addressing the zero-cell problem (Guo and Bhat, 2007), simultaneous control of household and individual-level attribute distributions (Casati et al., 2015, Zhu and Ferreira Jr, 2014, Ye et al., 2009, Guo and Bhat, 2007), improving the computational speeds (Pritchard and Miller, 2012), and non-integer conversion to integers (Choupani and Mamdoohi, 2015) etc.

Another popular technique is combinatorial optimisation, which focusses on selecting a subset of households in the disaggregate sample data that closely fit the marginal distributions in the census data for the same area (Voas and Williamson, 2000). This is done by randomly selecting an initial subset of households from the sample data, and iteratively replacing these with those remaining in the sample data, if and only when this leads to improvements in the fit of the subset. Although this approach has been reported to be superior (Ryan et al., 2009), the IPF method remains the most popular due to its low data requirements, reliability, and faster optimisation (Choupani and Mamdoohi, 2015, Sun and Erath, 2015).

Besides the two methods above, other techniques have been proposed including, the sample-free method (Barthelemy and Toint, 2013), Markov chain Monte Carlo simulation (Farooq et al., 2013), and the Bayesian network framework (Sun and Erath, 2015), among others.

3 Data

This section describes the study area, the data used, and the data processing conducted prior to model estimation. The study combines different data types (i.e. household travel survey data, census data, and CDR data) collected at different times between 2009 and 2012. Despite this limitation, these periods are considered close enough to facilitate cross-comparison.

3.1 Data description

3.1.1 *Study area*

The study location is Dhaka Metropolitan Area (DMA) in Bangladesh. The area covers approximately 303 square kilometres and is one of the world's most crowded places with a population density of 30551 persons per square kilometre (BBS, 2013). Due to the high population density, the cell tower density is also very high. The area is served by 1361 towers, with most these located in the central business district. The average tower-to-tower distance is approximately 1 kilometre (Iqbal et al., 2014). The total daily trip production from DMA residents was approximately 20.8 million in 2010, with 85.46% of these being home-based (JICA, 2010).

3.1.2 *CDR data*

The CDR data used in this study was provided by Grameenphone Ltd and covers the working days (i.e. Mondays to Thursdays) between 24 June 2012 and 07 July 2012 (2 weeks). The dataset comprises of 6.9 million anonymous users, who together generated over 600 million records during this period (see an excerpt of the CDR data in Table 1).

Table 1: Excerpt of the CDR data

Unique ID	Date	Time	Duration	Tower Longitude	Tower latitude
AAH03JACKAAAgfBALW	20120624	13:41:49	15	23.9339	90.2931
AAH03JAC8AAAbZfAHB	20120624	13:41:25	73	23.7931	90.2603
AAH03JAC4AAAcvbABC	20120624	13:27:39	8	23.7761	90.4261
AAH03JAC9AAAbWFAVM	20120624	13:27:27	41	23.7097	90.4036
AAH03JABkAAHvEkAQE	20120624	13:32:38	530	23.7386	90.4494

3.1.3 *Household travel survey data*

The household travel survey data used was collected between March 2009 and March 2010 as part of the Dhaka Urban Transport Network Development Study (JICA, 2010). The sampling of households in each zone was based on the population shares at a rate of approximately 1%. The total sample comprises of 67461 individuals and 17270 households, representing an average household size of approximately four persons. The collected information includes each individual's socio-demographic details (e.g. gender, age, working status, income, household size and housing type) and a single day trip diary. Table 2 presents the summary statistics of the data.

Table 2: Summary statistics of the household survey data

Gender		Age		Working status		Trip rate shares	
Male	53%	0-9 years	15%	Employed	35%	0 trips	43%
Female	47%	10-14 years	9%	Unemployed	38%	1-2 trips	41%
		15-19 years	8%	Student	27%	3-4 trips	14%
		20-29 years	22%			5+ trips	2%
		30-49 years	32%				
		50-59 years	8%				
		60+ years	5%				

3.1.4 *Census data*

The 2011 Bangladesh Population and Housing Census data was used (BBS, 2012). The Census was conducted from 15 to 19 March 2011. The available data reports the aggregate totals of selected person and household level attributes at different geographical scales (e.g. village, ward, and zone (Thana)).

Since we could not access the detailed census data due to privacy reasons, we used population synthesis techniques (Ye et al., 2009) to generate realistic artificial populations for the different study area zones by combining the aggregate census data with the household survey data as explained later in Section 3.2.2.

It may be noted that the fusion of household survey data and census data could only be done at the zone (Thana) level due to differences in the study area delimitations at smaller geographical scales. The variables available in both datasets are summarised in Table 3.

Table 3: Variables in both the census and the household survey data

Data	Household survey data	Census data
Individual attributes	Gender	Population by gender
	Age-group	Population by age-group
	Working status *(employed, unemployed, student)*	Population by working status
	Occupation *(agriculture, industry, services)*	Population by occupation
Household attributes	Household size	Number of households by household size
	Household type *(permanent, semi-permanent, thatched etc.)*	Number of households by household type

3.2 Data processing and combination

3.2.1 General concept

Figure 1 presents a summary of the data processing framework. The subsequent sections discuss the key aspects of this framework.

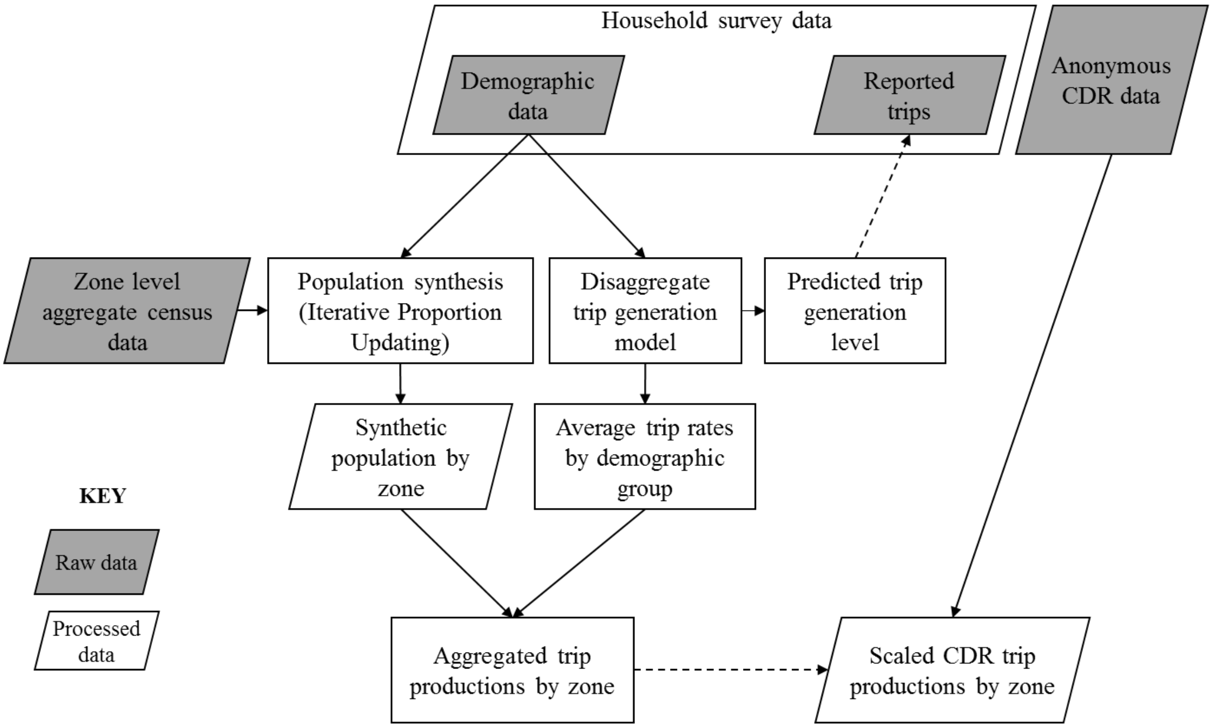
Household survey data
Demographic data
Reported trips
Anonymous CDR data
Zone level aggregate census data
Population synthesis (Iterative Proportion Updating)
Disaggregate trip generation model
Predicted trip generation level
Synthetic population by zone
Average trip rates by demographic group
KEY
Raw data
Processed data
Aggregated trip productions by zone
Scaled CDR trip productions by zone

Figure 1: Data processing framework

The overarching idea is to minimise the difference between the zonal trip productions derived from CDR data and those obtained by aggregating the disaggregate trip generation model, without compromising the behavioural sensitivities reflected in the household survey data. Model aggregation is based on a synthetic population generated using the Iterative Proportional Updating technique (Ye et al., 2009).

3.2.2 Population synthesis

Among the various software applications for population synthesis, we used PopGen (Ye et al., 2009), which is capable of conducting Iterative Proportional Updating (IPU). This algorithm simultaneously controls for both the person and the household-level attribute distributions during the fitting procedure, and has been proven to perform better than the simpler synthesis methods.

As seen in Figure 1 (top left), the algorithm relies on two raw datasets, the household survey data and the zone level aggregate census data to generate the zone-specific synthetic populations by means of IPU. The household and individual level control variables used in the IPU process are presented in Tables 4 and 5 respectively. It may be noted that we did not use the individual's occupation as there are differences in the definitions of the categories used in the household survey and the census data.

Table 4: Household-level control variables used in PopGen

HSETYP	**Housing type**	**HHLDSIZE**	**Household size**
HSETYP1	Pucka (Permanent house)	HHLDSIZE1	1
HSETYP2	Semi-pucka (Semi-permanent house)	HHLDSIZE2	2
HSETYP3	Kutcha (Thatched house)	HHLDSIZE3	3
HSETYP4	Jhupri (Slum house)	HHLDSIZE4	4
		HHLDSIZE5	5
		HHLDSIZE6	6
		HHLDSIZE7	7
		HHLDSIZE8	8+

Table 5: Individual-level control variables used in PopGen

GEND	**Gender**	**AGEP**	**Age-group**
GEND1	Male	AGEP1	0-9 years
GEND2	Female	AGEP2	10-14 years
		AGEP3	15-19 years
WRKST	**Working status**	AGEP4	20-29 years
WRKST1	Employed	AGEP5	30-49 years
WRKST2	Unemployed	AGEP6	50-59 years
WRKST3	Student	AGEP7	60+ years

Figure 2 presents the distribution of the Average Absolute Relative Differences (AARD) across the zones. This metric gives the mean deviation of the person weighted sums with respect to the household and person aggregate census totals. As observed, the AARD values for most zones are concentrated in the lower ranges of the axis, an indication that the population synthesis was successful.

Furthermore, comparisons of the synthetic versus the actual estimates for each attribute at the person and the household levels are presented in Figures 3 and 4 respectively, where the distributions are observed to have a close match.

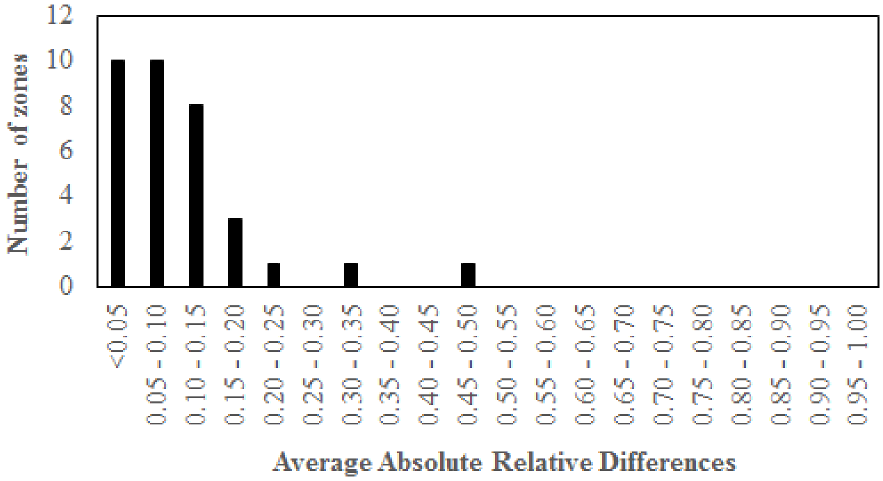

Figure 2: Distribution of the AARD values

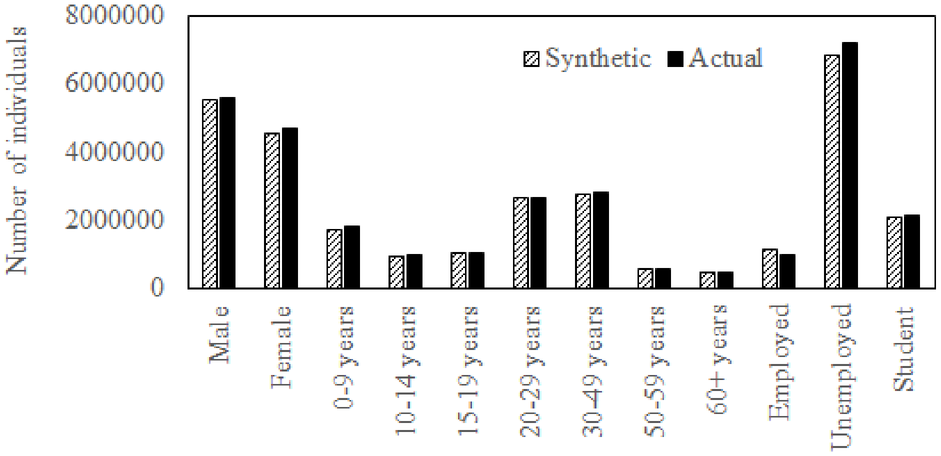

Figure 3: Distribution of the individual-level estimates

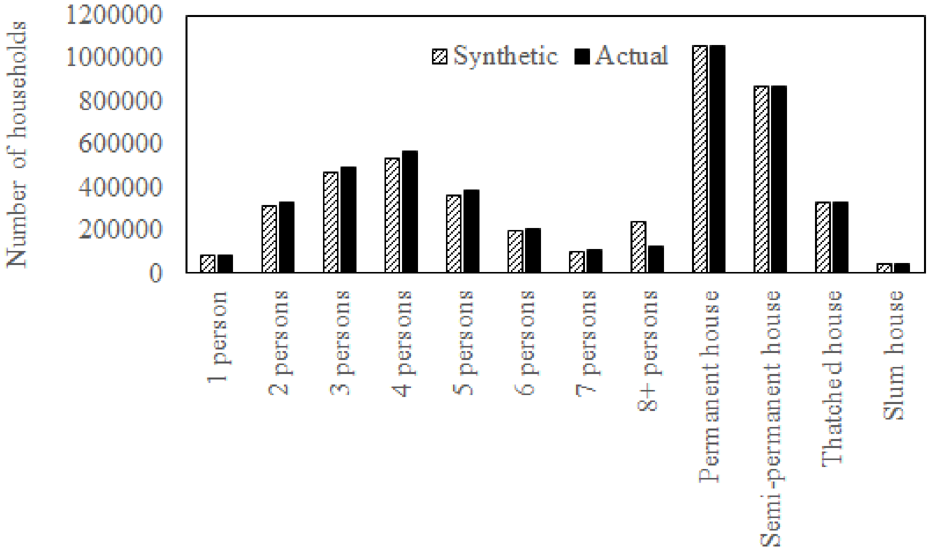

Figure 4: Distribution of the household-level estimates

3.2.3 Extraction of unscaled zonal trip productions from CDR data

The CDR data for the entire observation period was first analysed to identify each user's home location, which was defined as the most frequently observed cell tower at night (i.e. between 8 pm and 6 am). The labelled cell towers (i.e. home/others) for each user were then arranged according to the date and observation timestamp.

Home-based trips were extracted by considering any two consecutive CDR events from different cell towers, with one of those being the home cell tower. After conducting several trials, a lower distance threshold of 0.5 kilometres between subsequent towers was considered as the optimum for minimising the number of false trips due to tower jumps[1].

An upper threshold of 24 hours or midnight (whichever came first) was specified based on the assumption that a user typically travels from and back to home within the same effective day. Consequently, if the first and the last CDR events for the day were not at the home cell tower, corresponding raw trips were added (Çolak et al., 2015). This led to the unscaled zonal trip productions shown in Figure 1.

3.2.4 Scaling the CDR trip productions

The home cell towers derived from the CDR were~~are~~ mapped to the zones with the aid of a GIS software (QGIS Development Team, 2018). The total trips for each zone were then scaled using the ratio of the zonal population (from the census) to the number of users classified as residents of the zone from the CDR (see~~as used by~~ Çolak et al., 2015 for details).

4 Modelling framework

We propose an approach that combines two modelling strategies, that is, discrete choice modelling at the individual level and ordinary least squares at the aggregate level (shown in patterned textboxes in Figure 1).

4.1 Disaggregate trip generation model (Base model)

Discrete choice models have been the most preferred approach for modelling trip generation over the last few decades (e.g. Bwambale et al., 2015, Pettersson and Schmöcker, 2010, Agyemang-Duah and Hall, 1997). Although the ordered response choice mechanism has been the most preferred approach for modelling trip generation, previous findings in the context of car ownership choices (which are also ordered) have shown that the unordered response choice mechanism outperforms the former (Bhat and Pulugurta, 1998). To implement the unordered response choice mechanism, we rely on the random utility theory (Marschak, 1960). Let U_{nt} be the utility of individual n making t trips. This can be expressed as;

$$U_{nt} = \beta_t' X_n + \varepsilon_{nt} \qquad (1)$$

Where X_n is a vector of the socio-demographic attributes of individual n, β_t is a vector of the model parameters to be estimated, and ε_{nt} is the random component of utility. Since the individual socio-demographics are constant across the alternatives, we specify a different set of parameters for each trip generation level to reflect the fact that each attribute has a differential impact on the utility for each trip generation level.

Under the assumption that the error terms (ε_{nt}) are distributed independently and identically across alternatives and individuals using a type I extreme value distribution, the trip generation

[1] A false trip occurs when the user is not making a trip but there is a change in the tower as the operator reassigns the call to a different tower (due to load management purposes).

choice probabilities can be calculated using the multinomial logit (MNL) model (McFadden, 1974) as expressed below;

$$P_{nt} = \frac{\exp(\beta_t' X_n)}{\sum_{t^*} \exp(\beta_{t^*}' X_n)} \quad (2)$$

Where P_{nt} is the probability of individual n making t trips.

If we were to rely on the household travel survey data alone, the model parameters would be estimated by maximising the log-likelihood function below.

$$LL(\beta_t) = \sum_n \sum_t K_{nt} \ln(P_{nt}) \quad (3)$$

Where dummy variable $K_{nt} = 1$ if and only if individual n makes t trips, otherwise $K_{nt} = 0$.

However as mentioned earlier, fitting the model to match the trips reported in the household travel survey data alone can lead to biased parameter estimates due to reporting errors, thereby resulting in misrepresentation of the aggregate travel demand as reflected in Figure 5, where the predicted aggregate zonal trips from the base model are different from those derived from the CDR data, especially towards the right hand side of the figure. The proposed joint modelling framework (in the next section) seeks to optimise such differences.

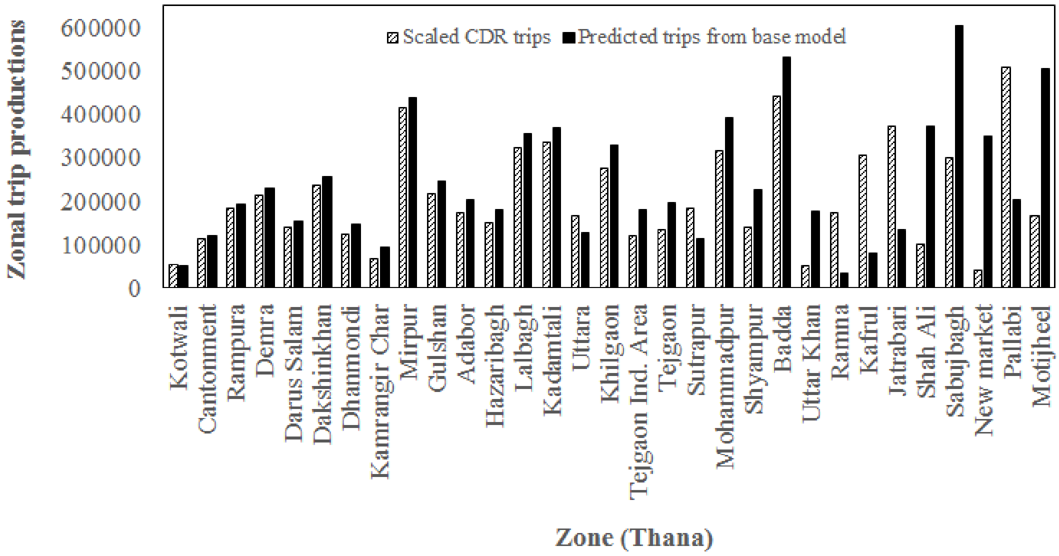

Figure 5: Distribution of the CDR trip productions

4.2 Joint trip generation model

The framework of the joint trip generation model is both sequential and simultaneous. It is sequential as it relies on the pre-estimated base model to obtain the priors of the parameter signs and relative magnitudes. However, when the parameter scales are being adjusted (without changing the prior parameter signs), the joint model simultaneously optimises performance at both the aggregate and disaggregate levels with respect to the CDR and the household travel survey data, respectively.

As mentioned earlier, this combined approach ensures that the resulting model does not lose the travel behaviour sensitivities reflected in the household travel survey data, by maintaining the sensitivities from the base model. Adjusting the parameter scales has an impact on the choice probabilities for each trip generation outcome, which influences the expected trip rates of the individuals. The framework of the joint trip generation model is described below. Let $\widehat{U}_{nt}$ be the updated utility of individual n making t trips. This can be expressed as;

$$\widehat{U}_{nt} = \alpha\beta_t^{'}X_n + \varepsilon_{nt} \tag{4}$$

Where α is a vector of the scaling factors to be estimated. The β parameters are priors derived from the base model, and are not re-estimated in the joint framework. The specification of the scaling factors is discussed later on.

The updated trip generation choice probability can be expressed as follows;

$$\widehat{P}_{nt} = \frac{\exp(\alpha\beta_t' X_n)}{\sum_{t^*}\exp(\alpha\beta_{t^*}' X_n)} \tag{5}$$

Where $\widehat{P}_{nt}$ is the updated probability of making t trips by individual n.

However, to estimate the scaling factors, we need to fulfil two objectives. The first objective is to explain the reported trips for each individual in the household survey data. The second objective is to ensure that the aggregated zonal trip productions are close to those derived from CDR data. Both outcomes have a probability attached to them and the simultaneous estimation maximises the joint probability of the two outcomes.

To estimate the aggregate zonal trip productions, we rely on the synthetic population generated in section 3.2.2. As mentioned earlier, the synthetic population was designed to match both the person and the household-level attribute distributions during the fitting procedure, thus making it more reliable. We have a synthetic population of M simulated individuals identified as m with $m = 1, \dots.., M$, and a study area comprising of Z zones identified as z with $z = 1, \dots..., Z$. Let $\widehat{P}_{mt}$ denote the updated probability of making t trips by simulated individual m. It may be noted that $\widehat{P}_{mt}$ is equivalent to $\widehat{P}_{nt}$ if both the simulated individual and the actual respondent in the household survey data have the same demographics (i.e. the values of $\widehat{P}_{mt}$ depend on the calculations of $\widehat{P}_{nt}$). Now, let $\widehat{T}_z$ denote the aggregate zonal trip production for zone z. This can be calculated by taking the weighted average trips for each simulated individual, in which the updated MNL probabilities are the weights, and summing across the zonal synthetic population as follows;

$$\widehat{T}_z = \sum_{m=1}^{M}\left[Y_{mz}\left(\sum_{t=1}^{T}(t * \widehat{P}_{mt})\right)\right] \tag{6}$$

Where dummy variable $Y_{mz} = 1$ if and only if simulated individual m belongs to zone z, otherwise, $Y_{mz} = 0$. The objective is to ensure that $\widehat{T}_z$ is as close as possible to the corrected CDR trip productions for zone z. If φ_z denotes the corrected CDR trip productions for zone z, the relationship between φ_z and $\widehat{T}_z$ can be expressed as follows;

$$\varphi_z = \widehat{T}_z + \omega_z \tag{7}$$

Where ω_z is an error term which we assume follows a normal distribution with a mean of zero, $\omega_z \sim N(0, \sigma^2)$. $P(\varphi_z)$ is then the likelihood of observing the CDR trip productions for zone z, and, from Equation 7, this can be expressed as follows;

$$P(\varphi_z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\varphi_z - \hat{T}_z)^2}{2\sigma^2}\right) \quad (8)$$

$P(\varphi_z)$ clearly depends on $\hat{P}_{nt}$ given that $\hat{T}_z$ is a function of $\hat{P}_{mt}$, which depends on the calculations of $\hat{P}_{nt}$ as explained earlier. For each survey respondent in zone z, we need to maximise the probability of the chosen alternative and ensure that the probabilities of all the alternatives maximise $P(\varphi_z)$. Let t_n^o denote the number of trips observed for individual n in the household survey data, such that $\hat{P}_{nt^o}$ gives the logit probability of the observed choice for individual n. The overall joint likelihood (L) of the observed choices and the aggregate CDR trip productions across individuals is calculated as follows;

$$L = \prod_{n=1}^{N}\left[\sum_{z=1}^{Z} H_{nz}\left(\hat{P}_{nt^o} * P(\varphi_z)\right)\right] \quad (9)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \prod_{n=1}^{N}\left[\sum_{z=1}^{Z} H_{nz}\left(\frac{\exp(\alpha\beta'_{t^o}X_n)}{\sum_{t^*}\exp(\alpha\beta'_{t^*}X_n)} * \exp\left(\frac{-(\varphi_z - \hat{T}_z)^2}{2\sigma^2}\right)\right)\right]$$

Where dummy variable $H_{nz} = 1$ if and only if survey respondent n belongs to zone z.

Since products are difficult to differentiate, we obtain the log-likelihood (LL) by applying logarithms to Equation 9 resulting in Equation 10.

$$LL = -\frac{N}{2}log(2\pi) - Nlog(\sigma) + \quad (10)$$

$$\sum_{n=1}^{N}\sum_{z=1}^{Z} H_{nz}\left(\ln\left[\frac{\exp(\alpha\beta'_{t^o}X_n)}{\sum_{t^*}\exp(\alpha\beta'_{t^*}X_n)}\right] - \frac{1}{2\sigma^2}(\varphi_z - \hat{T}_z)^2\right)$$

Three parameter scaling scenarios are tested, and these are;

- Model 1 This specification applies the same α scaling factor to the utility models of the different trip generation levels (see Equation 4), i.e. $\alpha_t = \alpha, \forall t$. The updated utility models have the same relative variable sensitivities as in the base model, albeit with different parameter scales.

- Model 2 This specification applies a different α_t scaling factor to the utility model of each trip generation level. The updated utility models maintain the base model relative variable sensitivities for each particular trip generation level, however, the variable sensitivities across the different trip generation levels are adjusted with different parameter scales, and hence the relative values across levels change from the base model.

This specification applies a different α_x scaling factor to each explanatory variable X (e.g. gender, age-group, and working status), however, α_x does not change across the different trip generation levels.

- Model 3 The updated utility models maintain the base model attribute-level relative sensitivities for a particular variable across the different trip generation levels, however, the inter-variable relative sensitivities are adjusted with different parameter scales.

4.3 Model evaluation framework

The performance of the joint models is evaluated in terms of both the temporal and the spatial transferability as presented in Figures 6 and 7, respectively.

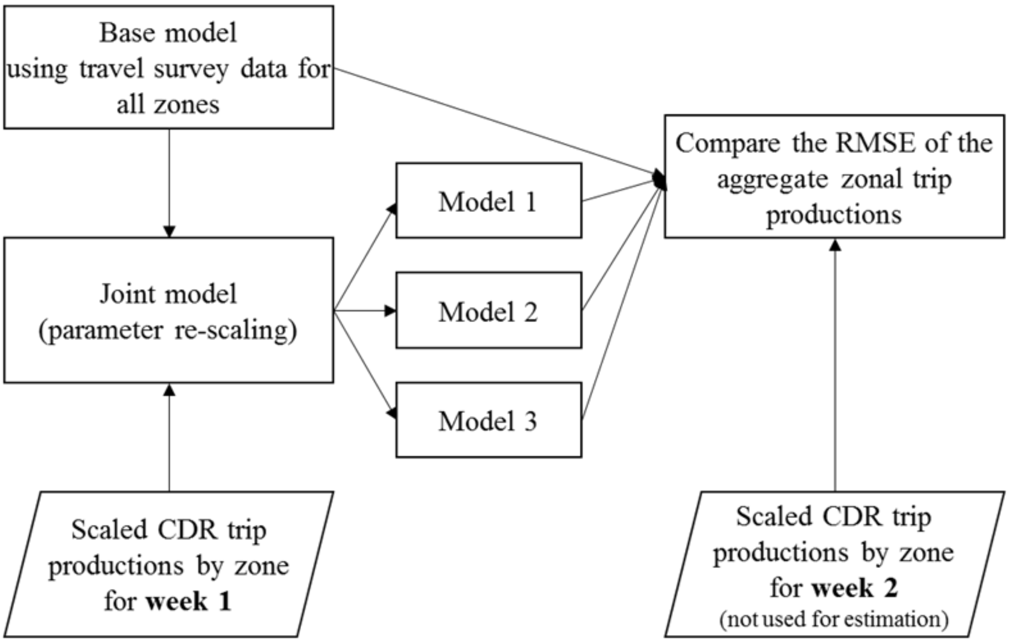

Figure 6: Temporal transferability framework

In terms of temporal transferability, the joint models associated with each parameter scaling scenario are estimated using the zonal aggregate CDR trip productions for week 1. The prediction capacities of the estimated joint models, as well as the base model are then compared in terms of the root mean square errors with respect to the zonal aggregate CDR trip productions for week 2 (see Figure 6).

In terms of spatial transferability, the study area zones are randomly divided into two groups. The base and the joint models are then estimated using the data for one group of zones and applied to the other group of zones (not used for estimation). The prediction capacities of the models are then compared in terms of the predictive joint log-likelihoods, and the root mean square errors with respect to the aggregate CDR trip productions of the application zones (see Figure 7).

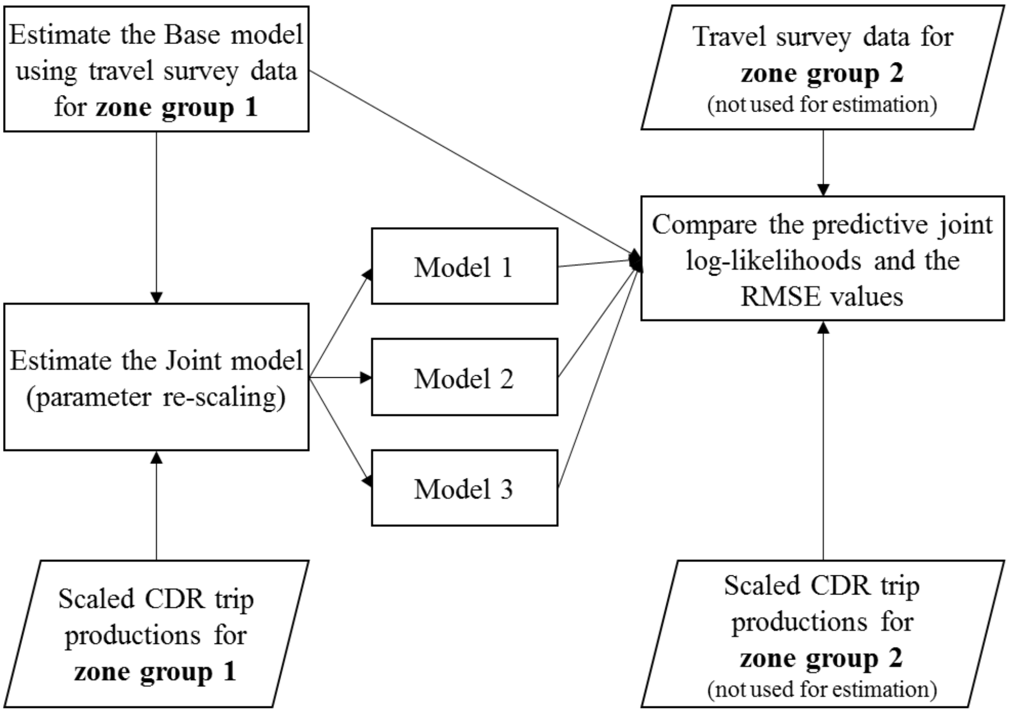
Estimate the Base model using travel survey data for **zone group 1**
Travel survey data for **zone group 2** (not used for estimation)
Model 1
Compare the predictive joint log-likelihoods and the RMSE values
Estimate the Joint model (parameter re-scaling)
Model 2
Model 3
Scaled CDR trip productions for **zone group 1**
Scaled CDR trip productions for **zone group 2** (not used for estimation)

Figure 7: Spatial transferability framework

5 Modelling results

This section presents the final model specification, as well as the model estimation and validation results.

5.1 Variable specification

The dependent variable is the number of individual home-based trips (irrespective of the trip purpose). This is because we could not reliably infer the purposes of the CDR trips. Based on distributions in the data, the trip generation levels were grouped into 0, 1-2, 3-4, and 5+ trips per day. The explanatory variables considered for possible inclusion in the model are those that were used for population synthesis. The household-level variables (i.e. household size and type) were however not included in the final model as they led to unreasonable parameter signs, potentially due to their weak influence on individual trip-making decisions. The final model specification thus contains the gender, the age-group, and the working status of the individuals, coded as dummy variables.

For model identification purposes, the parameters associated with the zero trip generation level were treated as the base (for all explanatory variables). Furthermore, male non-workers in the 30-49 age-group were treated as the base demographic group, and their preferences are entirely explained by the alternative specific constants. Thus, the model parameter estimates represent the differential impact on utility with respect to the zero trip generation level and the base demographic group.

5.2 Estimation results

5.2.1 Base model

We first estimated the base model to assess whether the parameter estimates are in line with the expected travel behaviour. The model results are presented in Table 6.

Table 6: Base model results

Variable	Parameter	t-statistic
Alternative specific constants (ASCs)		
1-2 trips	-0.2069	-7.46
3-4 trips	-1.0408	-24.56
5+ trips	-3.0859	-31.19
Dummies specific to gender (base category is males)		
Females		
1-2 trips	0.0870	3.94
3-4 trips	-0.2841	-7.95
5+ trips	-0.2654	-3.15
Dummies specific to working-status (base category is non-workers)		
Workers		
1-2 trips	0.4630	17.23
3-4 trips	0.9252	23.05
5+ trips	1.1482	12.38
Students		
1-2 trips	1.4079	46.47
3-4 trips	0.9381	17.13
5+ trips	-0.5333	-2.65
Dummies specific to age-group (base category is the 30-49 years age-group)		
Age 1-9 years		
1-2 trips	-1.6354	-50.69
3-4 trips	-3.1065	-36.73
5+ trips	-3.5549	-9.46
Age 10-14 years		
1-2 trips	-0.8143	-19.49
3-4 trips	-1.7635	-22.52
5+ trips	-1.9201	-6.00
Age 15-19 years		
1-2 trips	-0.6539	-16.22
3-4 trips	-0.9669	-15.71
5+ trips	-1.0077	-5.71
Age 20-29 years		
1-2 trips	-0.1457	-5.67
3-4 trips	-0.3249	-9.58
5+ trips	-0.3009	-4.02
Age 50-59 years		
1-2 trips	-0.1423	-4.12
3-4 trips	-0.2552	-5.92
5+ trips	-0.3721	-3.81

Table 6 cont'd

Variable	Parameter	t-statistic
Age 60+ years		
1-2 trips	-0.2494	-5.63
3-4 trips	-0.3531	-6.14
5+ trips	-0.4853	-3.47
Measures of fit		
Number of observations	65419	
Log-likelihood at zero	-90689.99	
Log-likelihood at convergence	-64859.90	
Number of parameters	30	
Adjusted rho-square	0.2845	
Likelihood ratio	51660.10	
P value of the likelihood ratio	0.0000	

The alternative specific constants capture the underlying differential impact on utility with respect to the zero trip generation level. All the estimates are negative, and their magnitude increases with respect to the trip generation level. Keeping all other factors constant, this reflects a general tendency to make fewer trips, especially by the base category (i.e. male, non-workers, aged 30-49 years).

The parameter estimates for females represent the differential impact on utility with respect to males. For 1-2 trips, we obtain a positive parameter estimate, while for the higher trip generation levels, we obtain negative parameter estimates. The proportion of women working in the garments industry, one of the leading sectors in Dhaka, is 64-90% (ADB and ILO, 2016). This probably explains the positive parameter sign for 1-2 trips. Otherwise, males are more likely to make a higher number of trips compared to females, probably due to the average higher income levels of the former (BBS, 2012) and socio-cultural factors.

The parameter estimates for the working status variables (i.e. workers and students) represent the differential impact on utility with respect to non-workers. As observed, the parameters for workers are positive, and their magnitudes increase with respect to the trip generation level, an indication that workers generally make more trips compared to non-workers. On the other hand, the parameter estimates for students are positive for 1-2 and 3-4 trips, and negative for 5+ trips. This shows that students make more trips compared to non-workers only up to a reasonable level expected for school going individuals.

Similarly, the parameter estimates for the age-group variables represent the differential impact on utility with respect to the 30-49 years age-group. As observed, the parameter estimates for all the other age-groups are negative, an indication that they generally make fewer trips compared to the base age-group (30-49 years). The active working age of white-collar workers in Bangladesh typically ranges between 29 and 60 years (i.e. the latest age for completing tertiary education and the retirement age respectively (BBS, 2012)). It is therefore reasonable that persons in the 30-49 years age-group are more active travellers due to their economic vibrancy.

Finally, it is observed that the overall model (in terms of the likelihood ratio), as well as all the parameter estimates (in terms of the t-statistics) are statistically significant at the 99% level of confidence (see Ben-Akiva and Lerman, 1985 for details).

5.2.2 Joint models

As earlier mentioned, the parameters of the base model were fixed in the joint modelling framework, and only the scaling factors were estimated. Table 7 presents the estimated scaling factors and the measures of fit for all the three models for comparison purposes. Positive scaling factors were obtained for all the three models, an indication that the resultant coefficients in the scaled joint models have the same signs as those in the base model.

A comparison of the joint convergence log-likelihoods shows that Model 3 gives the best performance, followed by Model 2, and then Model 1. This is attributed to the flexibility of the parameter scaling framework. An important point to note is that all the three joint models perform better than the base model in terms of the joint log-likelihood.

As earlier mentioned, during model optimisation, we are basically dealing with a trade-off between disaggregate and aggregate model performance. Thus, the disaggregate log-likelihood of the joint models is a little worse than that of the base model. However, if the base model parameters are directly used to estimate the joint log-likelihood, it is observed that the model yields the worst performance.

The p-values of the likelihood ratios of the joint models with respect to the base model are all less than 0.01, an indication that the improvements in performance are statistically significant at the 99% confidence level beyond the advantages offered by the additional parameters (see Ben-Akiva and Lerman, 1985 for details).

Table 7: Joint model scaling factors

Description of scaling factor	Model 1		Model 2		Model 3	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Model 1						
Uniform factor (applied to all the base model parameters)	1.3650	2280.16				
Model 2 **(Factors specific to trip generation level)**						
1-2 trips			1.2716	131.39		
3-4 trips			1.4873	247.83		
5+ trips			1.1699	158.63		
Model 3 **(Factors specific to particular variables)**						
Gender					1.5228	33.81
Working status					1.8148	105.16
Age-group					1.3262	120.70
ASCs					1.6023	171.51
Measures of fit						
Convergence LL at the disaggregate level	-66002.75		-65914.01		-67747.10	
Convergence LL at the aggregate level	-718560.40		-718377.10		-715805.30	
Joint convergence LL	-784563.20		-784291.20		-783552.40	
Base model convergence LL	-64859.90		-64859.90		-64859.90	
Base model LL at the aggregate level	-805093.10		-805093.10		-805093.10	
Base model joint convergence LL	-869953.00		-869953.00		-869953.00	
Likelihood ratio (joint model w.r.t the base model)	170780		171234		172801	
P value	0.0000		0.0000		0.0000	

5.3 Model evaluation in terms of transferability

The models based on the full sample have been presented in the previous section. To evaluate the stability and the predictive performance of the joint models as well as the base model, we compared their temporal and spatial transferability following the evaluation framework described in Section 4.3. Tables 8 and 9 present the measures of fit in terms of the temporal and the spatial transferability, respectively.

Table 8: Temporal transferability

	Measure	Base model	Model 1	Model 2	Model 3
Week 1 (Estimation)	LL (disaggregate level)	-64859.90	-66024.40	-65940.80	-67850.40
	LL(aggregate level)	-805642.50	-719566.80	-719396.20	-716695.30
	Joint LL	-870502.40	-785591.20	-785337.00	-784545.70
Week 2 (Application)	LL (disaggregate level)	-64859.90	-66024.40	-65940.80	-67850.40
	LL(aggregate level)	-804545.50	-717793.90	-717596.20	-715031.60
	Joint LL	-869405.40	-783818.30	-783537.00	-782882.00
	RMSE w.r.t CDR trips	43342.84	13547.09	13527.84	13328.49

Table 9: Spatial transferability

	Measure	Base model	Model 1	Model 2	Model 3
Zone group 1 (Estimation)	LL (disaggregate level)	-26102.10	-26712.45	-26652.76	-27724.63
	LL(aggregate level)	-321381.60	-290869.40	-290725.20	-288898.10
	Joint LL	-347483.70	-317581.85	-317377.96	-316622.73
Zone group 2 (Application)	LL (disaggregate level)	-38859.38	-39701.58	-39352.09	-41303.51
	LL(aggregate level)	-491580.30	-429017.00	-428604.80	-426638.20
	Joint LL	-530439.68	-468718.58	-467956.89	-467941.71
	RMSE w.r.t CDR trips	50626.73	13375.06	13274.68	13161.58
Zone group 2 (Estimation)	LL (disaggregate level)	-38688.76	-39227.43	-39333.92	-40185.59
	LL(aggregate level)	-482400.40	-428113.30	-427818.70	-426238.10
	Joint LL	-521089.16	-467340.73	-467152.62	-466423.69
Zone group 1 (Application)	LL (disaggregate level)	-26219.53	-26689.06	-26786.11	-27445.95
	LL(aggregate level)	-315772.10	-289862.10	-289890.20	-288799.10
	Joint LL	-341991.63	-316551.16	-316676.31	-316245.05
	RMSE w.r.t CDR trips	38776.13	13702.57	13758.49	13602.58

From Table 8, it is observed that the temporal transferability of the joint models is generally higher than that of the base model in terms of the joint log-likelihoods and the root mean square errors (RMSE) with respect to the zonal CDR trips. Among the three joint models, Model 3 offers the best transferability, however, Model 2 gives the best prediction at the disaggregate level in both the estimation and the application contexts.

For spatial transferability, we tested both directions of model transfer. It may be noted that the general interpretation of the base model parameters for each group of zones did not change. From Table 9, it is again observed that the joint models are generally more transferrable compared to the base model in terms of the joint log-likelihoods and the root mean square errors for both directions.

In this particular case, it is observed that Model 2 gave the best disaggregate prediction for the zone group 1 to 2 transfer direction, while Model 1 gave the best disaggregate prediction for the reverse transfer direction.

An important point worth mentioning is that the superior performance of the base model at the disaggregate level is expected as it was designed to fit the travel survey data alone, but as mentioned earlier, this could be prone to reporting errors and hence less dependable.

From the results, it is clear that Model 3 gives the best overall spatial and temporal transferability, however, the disaggregate performance of Models 1 and 2 as highlighted above shows that these parameter scaling approaches offer some benefits as well. These results present initial efforts to exploit the benefits of both household travel survey and mobile phone data to optimise the performance of travel behaviour models, and there is a need for further research using data from different contexts to investigate the different parameter scaling approaches in further detail.

6 Summary and conclusions

This paper started by highlighting the reporting errors and sampling bias associated with household travel survey data, and how these could lead to biased model parameters (e.g. Rolstad et al., 2011, Groves, 2006). The paper outlines the possible consequences of such issues in the context of trip generation, where the estimated models would misrepresent the distribution of the aggregate travel demand across zones.

Although traditional travel surveys are increasingly being replaced by smartphone based surveys, which alleviate the issue of misreporting of trips, issues with representativeness and sample size remain, as well as with encouraging respondents to provide a sufficiently long stream of data (cf. Calastri et al., 2019). On the other hand, while mobile phone call detail record (CDR) data is widely available, large in size and more representative, it is lacking information on core causal variables.

The paper demonstrates the feasibility of a joint modelling framework to find the best fit at both the aggregate and disaggregate levels by combining household travel survey, census, and CDR data. The census data is crucial in creating a bridge between the two other data sources. The joint modelling framework operates by adjusting the parameter scale(s) of a pre-estimated base model to jointly optimise the prediction accuracy with respect to the reported trips in travel survey data and the zonal aggregate trip productions derived from CDR data. Three different approaches of parameter scaling were investigated (i.e. uniform, alternative specific, and variable specific scaling corresponding to joint models 1, 2, and 3 respectively). All the three joint models were found to have higher temporal and spatial transferability compared to the base model which relies on household travel survey data alone, thus making them more reliable. Although variable specific scaling (Model 3) produced the best overall results, there is a need for further research using data from different contexts to investigate if this finding is universally applicable.

Although the proposed framework has been tested in the context of trip generation, it has potential benefits in improving the modelling of the other transport choices (such as mode choice, route choice, departure time choice etc.). We conclude that the results of this study serve as a proof-of-concept that mobile phone data can be fused with traditional data sources to improve the temporal and spatial transferability of models. This approach is particularly important in the context of developing countries where reliable traditional data sources are scarce, and models making use of low-cost passive data to enhance their temporal and spatial transferability are invaluable.

Acknowledgements

The research in this paper used mobile phone data made available by Grameenphone Ltd, Bangladesh, household travel survey data provided by the Japan International Cooperation Agency (JICA), and aggregate census data obtained from the Bangladesh Bureau of Statistics (BBS). We would like to thank the Economic and Social Research Council (ESRC) of the UK and the Institute for Transport Studies, University of Leeds for funding this research. Professor Stephane Hess' time is supported by the European Research Council through the consolidator grant 615596-DECISIONS.

References

ADB & ILO 2016. Bangladesh: Looking beyond garments: Employment diagnostic study. Manila, Phillipines: Asian Development Bank and International Labour Organization.

Agyemang-Duah, K. & Hall, F. L. 1997. Spatial transferability of an ordered response model of trip generation. *Transportation Research Part A: Policy and Practice,* 31**,** 389-402.

Barthelemy, J. & Toint, P. L. 2013. Synthetic population generation without a sample. *Transportation Science,* 47**,** 266-279.

BBS 2012. Community Report: Dhaka Zila: June 2012. *Population and Housing Census 2011.* Dhaka: Bangladesh Bureau of Statistics (BBS).

BBS 2013. District Statistics 2011 Dhaka. Dhaka: Bangladesh Bureau of Statistics.

Beckman, R. J., Baggerly, K. A. & Mckay, M. D. 1996. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice,* 30**,** 415-429.

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

Bhat, C. R. & Pulugurta, V. 1998. A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. *Transportation Research Part B: Methodological,* 32**,** 61-75.

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

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.

Cárcamo, J. G., Vogel, R. G., Terwilliger, A. M., Leidig, J. P. & Wolffe, G. Generative models for synthetic populations. Proceedings of the Summer Simulation Multi-Conference, 2017. Society for Computer Simulation International, 7.

Casati, D., Müller, K., Fourie, P. J., Erath, A. & Axhausen, K. W. 2015. Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record: Journal of the Transportation Research Board***,** 107-116.

Chen, C., Bian, L. & Ma, J. 2014. From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies,* 46**,** 326-337.

Choupani, A.-A. & Mamdoohi, A. R. 2015. Population Synthesis in Activity-Based Models: Tabular Rounding in Iterative Proportional Fitting. *Transportation Research Record: Journal of the Transportation Research Board***,** 1-10.

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.

Farooq, B., Bierlaire, M., Hurtubia, R. & Flötteröd, G. 2013. Simulation based population synthesis. *Transportation Research Part B: Methodological,* 58**,** 243-263.

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].

Guo, J. & Bhat, C. 2007. Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board***,** 92-101.

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.

Itsubo, S. and Hato, E., 2006. *Effectiveness of household travel survey using GPS-equipped cell phones and Web diary: Comparative study with paper-based travel survey* (No. 06-0701).

Janzen, M., Müller, K. & Axhausen, K. W. Population Synthesis for Long-Distance Travel De-mand Simulations using Mobile Phone Data. 6th Symposium of the European Association for Research in Transportation (hEART 2017), 2017.

JICA 2010. Dhaka Urban Transport Network Development Study (DHUTS) in Bangladesh, Final Report. Dhaka: Japan International Cooperation Agency.

Kressner, J. D. 2017. Synthetic Household Travel Data Using Consumer and Mobile Phone Data. *Final Report for NCHRP IDEA Project 184.* Transportation Research Board.

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.

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.

Panigutti, C., Tizzoni, M., Bajardi, P., Smoreda, Z. & Colizza, V. 2017. Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models. *Royal Society open science,* 4**,** 160950.

Pettersson, P. & Schmöcker, J.-D. 2010. Active ageing in developing countries?–trip generation and tour complexity of older people in Metro Manila. *Journal of Transport Geography,* 18**,** 613-623.

Pritchard, D. R. & Miller, E. J. 2012. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation,* 39**,** 685-704.

QGIS Development Team. 2018. *QGIS Geographic Information System* [Online]. Available: https://qgis.org/en/site/ [Accessed 14 August 2018].

Rolstad, S., Adler, J. & Rydén, A. 2011. Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value in Health,* 14**,** 1101-1108.

Ros, O. G. C. & Albertos, P. G. 2016. D5.4 Enhanced Version of MATSim: Synthetic Population Module. *Innovative Policy Modelling and Governance Tools for Sustainable Post-Crisis Urban Development (INSIGHT).* Madrid, Spain: INSIGHT Consortium.

Ryan, J., Maoh, H. & Kanaroglou, P. 2009. Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis,* 41**,** 181-203.

Stopher, P., FitzGerald, C. and Xu, M., 2007. Assessing the accuracy of the Sydney Household Travel Survey with GPS. *Transportation*, *34*(6), pp.723-741.

Sun, L. & Erath, A. 2015. A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies,* 61**,** 49-62.

Voas, D. & Williamson, P. 2000. An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography,* 6**,** 349-366.

Vogel, N., Theisen, C., Leidig, J. P., Scripps, J., Graham, D. H. & Wolffe, G. 2015. Mining Mobile Datasets to Enable the Fine-Grained Stochastic Simulation of Ebola Diffusion. *Procedia Computer Science,* 51**,** 765-774.

White, J. & Wells, I. Extracting Origin Destination Information from Mobile Phone Data. Eleventh International Conference on Road Transport Information and Control (Conf. Publ. No. 486), March 2002 London. IET, pp. 30 - 34.

Ye, X., Konduri, K., Pendyala, R. M., Sana, B. & Waddell, P. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. 88th Annual Meeting of the Transportation Research Board, Washington, DC, 2009.

Zhanga, D., Caob, J., Feygina, S., Tangc, D. & Pozdnoukhova, A. 2017. Connected Population Synthesis for Urban Simulation. *Personal Communication. Draft Available from Authors by Request*.

Zhao, F., Pereira, F.C., Ball, R., Kim, Y., Han, Y., Zegras, C. and Ben-Akiva, M., 2015. Exploratory analysis of a smartphone-based travel survey in Singapore. *Transportation Research Record: Journal of the Transportation Research Board*, *2*(2494), pp.45-56.

Zhu, Y. & Ferreira Jr, J. 2014. Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record,* 2429**,** 168-177.

Zilske, M. & Nagel, K. 2014. Studying the accuracy of demand generation from mobile phone trajectories with synthetic data. *Procedia Computer Science,* 32**,** 802-807.

Zilske, M. & Nagel, K. 2015. A simulation-based approach for constructing all-day travel chains from mobile phone data. *Procedia Computer Science,* 52**,** 468-475.