A Statistical Synthetic Population Calibration for Activity-Based Model with Incomplete Census Data

Treerapot SIRIPIROTE[a], Agachai SUMALEE[*,b], H.W. Ho[c]

[a,b,c] *Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China*
[a] *E-mail: treerapot_eng@yahoo.com*
[b] *E-mail: ceasumal@polyu.edu.hk*
[c] *E-mail: cehwho@polyu.edu.hk*

Abstract: Synthetic population generator is the core component of the microsimulation in activity-based travel demand model. Typically, synthetic population is used in the way that their decisions on activity-travel pattern are simulated. Traditionally, household sample survey data is used to synthesize the population. The estimated results can be biased due to such as low-sampling size and inaccurate household sample data. To deal with this issue, a statistical maximum-likelihood method to calibrate synthetic population using the roadside observations (link counts) is proposed. Statistical performances of the proposed method are evaluated on the illustrative network and real network with census and household sample survey data. Multiday link counts are simulated from (true) activity-based model parameters and synthetic population. Tests are carried out assuming different number of observations and observation variations. The results illustrate the efficiency of the model calibration based on link counts and its potential for large and complex applications.

Keywords: Maximum-likelihood estimation, Link flow, Statistical synthetic population calibration

1. INTRODUCTION

In recent years, there is growing importance for the development of activity-based models that is able to describe travel behavior more realistically than the traditional four-step model. The activity-based model (ABM) has been developed by which travel demand is derived from activity participation and activity behavior sequences or patterns (Bhat *et al.*, 1999). The ABM framework is generally based on agent-based microsimulation by which agents in the microsimulation model often represent the individuals, grouped by households, living in the study area (e.g. UrbanSim, 2011; MATSim-T, 2011; Bradley *et al.*, 2010; Bhat *et al.,* 2004; Bowman *et al.*, 2001, 2006, 2008; Albatross (Arentze and Timmermans, 2004)).

Alternatively, *the agent* can also be represented in terms of a group of people who has the same person and household demographics (e.g. age, gender, household size, and household income). On the demand side of ABM, the microsimulation involves two major steps (Guo and Bhat, 2007): (1) construction of a microdata set representing the characteristics of the decision agents of interest (called *population synthesis*), and (2) simulation of the decision agent’s behaviour of interest such as activity-travel patterns, based on utility-based or rule-based models. On the supply side, all activity-travel patterns of each agent are simultaneously assigned on the networks.

Due to the limitation of budget and time, all characteristics of the whole agents (or all individual persons) in the study area can not be observed. Consequently, the methodology to synthesize the synthetic population from both aggregate dataset (e.g. complete census data) and disaggregate dataset (e.g. a sample of households and individuals) was conducted by many research studies (e.g. Ye *et al.*, 2009; Axhausen and Müller, 2011; Guo and Bhat, 2007; Beckman *et al.*, 1996). For instance, the public-use microdata samples (PUMS) of U.S. consist of information on all socio-demographic variables of interest, but only a sample of households and individuals is available (called *the reference sample*). In contrast, the aggregate dataset comprises some socio-demographic variables of disaggregate data (called *control variables*) from the census block, which can represent a full aggregate dataset or a larger sampling proportion of PUMS data. This is exemplified by the case that the marginal distributions/totals of population categorized by age and gender at census-block level are known (e.g. census standard tape file 3A (see Census, 1992)), but the individual data (microdata) of age and gender is still unknown at this level.

Typically, the marginal distribution of control variables is available at census block level, the joint distribution of these control variables is, however, still unknown. In addition, the area of PUMS (PUMA) is larger than the area of the census block and indeed usually contains many census blocks. Consequently, the basic population synthesis for this case study is to estimate the joint distribution of one or multi-dimensions of control variables. In the way, the marginal totals of selected control variables and the correlation structure observed in the disaggregate dataset are preserved. This procedure is generally carried out by using the iterative proportion fitting (IPF) technique. Drawing or copying households from the reference samples into the targeted traffic analysis zone is then conducted by Monte Carlo simulation (e.g. Guo and Bhat, 2007; Axhausen and Müller, 2011; Bowman, 2009).

The information available for processing population synthesis, however, can vary from one country to another. For instance, the marginal distribution of control variables is not available at the desired geographical level (traffic analysis zone), but only for a larger area. This is exemplified by the resolution of census block in some countries containing many traffic analysis zones (e.g. Swiss federal statistical office, 2000; Thai national statistical office, 2010). Consequently, the procedure to synthesize the population into a smaller area (traffic zone) is needed. According to the problem of population allocation, exogenous land use variables (e.g. number of buildings and average household size per living quarter type) are generally used to pre-estimate the population allocation at the traffic zone level. However, the estimated results can be uncertain, in particular, the case of inconsistent household size distribution by traffic zone. In addition, synthetic population is simulated from the household sample survey data. The estimated results can be biased due to such as low-sampling size and inaccurate household data (Bricka and Bhat, 2006). To deal with such problems, Cool *et al.* (2010) presented the meta-heuristic approach to calibrate synthetic population using OD trips.

As the previous research that we focused on the calibration of model parameters in the activity-based model (Siripirote *et al.*, 2013), in this study, the statistical framework for the synthetic population calibration from link counts is conducted. The maximum likelihood estimation problem based on link counts is defined in this paper. This model is solved by the proposed maximum-likelihood method. The remainder of the paper is organized as follows. Basic components of the network and activity choice dimensions are described in the following section. The formulation of the maximum-likelihood estimation problem is presented in the third section. Included also is the process to calibrate synthetic population. The numerical example is then conducted in the fourth section. Also included is the analysis of results. The proposed calibration method applied to a medium-sized network in Thailand is reported in the fifth section. Finally, conclusions are drawn in sixth section.

2. NETWORK AND ACTIVITY CHOICE REPRESENTATION

For a traffic network $(\mathbf{N},\mathbf{L})$, $\mathbf{N}$ is the set of nodes and $\mathbf{L}$ is the set of links. Activity location lo is located in each traffic zone where $\mathbf{Z}$ is the set of zone centroids and $\mathbf{L}_z$ is the set of links in traffic zone z ($\mathbf{Z} \subset \mathbf{N}$ and $\mathbf{L}_z \subset \mathbf{L}$). In addition, the activity location (*lo*) is assumed to be virtually located at the zone centroid which is one of the nodes in a network, representative of all real activity locations in the zone ($lo \in \mathbf{Z}$).

As we consider on daily activity-travel participations, user *i* makes a plan to perform *activity pattern y*. Let $\mathbf{A}_y$ denote the activity pattern consisting of an ordered set of the activities which are daily scheduled to be carried out:

$$\mathbf{A}_y = \{a_1,...,a_q,...,a_{Q_y}\} \quad \text{for } y \in \{1,...,Y\} \tag{2.1}$$

where,

Y : the total daily activity patterns,

a_q : an activity performed in sequence *q* of activity pattern *y*, $q \in \{1,...,Q_y\}$, and

Q_y : the total number of activities included in activity pattern *y*.

For instance, if the activity pattern (*y=1*) is Stay-at-home(H)-Working(W)-Stay-at-home(H), $\mathbf{A}_1 = \{\text{H},\text{W},\text{H}\}$.

Individuals can then select a *trip chain* which consists of an ordered set of activity location/mode/time of day. Given the list of activities in the specified activity pattern performed by individuals, $\mathbf{A}_y$, the trip chain *h* (the combined set of activity location/mode/time of the day performed by trip makers starting at origin zone *o*) is denoted as $\text{B}^o_{y,h}$. This is expressed by:

$$\mathbf{B}^o_{y,h} = \begin{matrix} \{(lo_1,...,lo_q,...,lo_{Q_{y,h}}),(mo_{1,2},...,mo_{q,q+1},...,mo_{Q_{y,h}-1,Q_{y,h}}) \\ ,(m_{1,2},...,m_{q,q+1},...,m_{Q_{y,h}-1,Q_{y,h}})\} \end{matrix} \tag{2.2}$$

$$\text{for } h \in \{1,...,H_y\},\ lo_1 = o,\ lo_q \in \mathbf{Z},\ q \in \{1,...,Q_{y,h}\}$$

where,

$mo_{q,q+1}$: travel mode from activity location lo_q to lo_{q+1}.

$m_{q,q+1}$: travel period of individuals traveling from lo_q to lo_{q+1}.

lo_q : activity location *q* where individuals perform an activity.

$Q_{y,h}$: the total number of visits at activity locations of individuals who make trip chain *h* associated with activity pattern *y*.

H_y : the total number of trip chains associated with activity pattern *y*.

Note that a trip chain that begins and ends at the same activity location ($lo_1 = lo_Q$) is called a *tour*. A tour of trip chain that begins at home is called *home-based tour*. In addition, for simple illustrations throughout the paper, individuals, who make trip chain *h* associated with activity pattern *y* originating from zone *o*, perform *activity chain j*. In other words, activity chain *j* is the combined decision of travelers on activity pattern *y* and trip chain *h*.

3. ESTIMATION PROBLEM FORMULATION

3.1 Demand Side

Consider ABM based on the utility-based approach, a demand function in this model is assumed to explicitly relate to exogenous variables via a *N*-dimensional parameter vector $\boldsymbol{\alpha}$. Let $\boldsymbol{\Gamma}$ be the choice set of activity chains of travellers and *p* be the type of individuals having the same person/household demographic in synthetic population (e.g. same type of gender, career, family (with/without child), and household size/income)). In addition, type *p* synthetic population (simply called type *p* population) represents a group of simulated people including anyone belonging to type *p*. Given the probability that type *p* population in origin zone *o* selecting the activity chain *j*, $\mathrm{Pr}_j^{o,p}(\boldsymbol{\alpha})$ for $j \in \boldsymbol{\Gamma}$, the demand of activity chain *j*, $u_j(\boldsymbol{\alpha})$, can be formulated as:

$$u_j(\boldsymbol{\alpha}) = \sum_p \mathrm{N}_o^p \mathrm{Pr}_j^{o,p}(\boldsymbol{\alpha}) \qquad \forall \; j \in \boldsymbol{\Gamma} \tag{3.1}$$

where,

N_o^p : the number of type *p* synthetic population in origin zone *o*.

In general, $\mathrm{Pr}_j^{o,p}(\boldsymbol{\alpha})$ is estimated by the choice logit model, in which the input data can be obtained from household and travel diary survey. For illustration of the method to calibrate synthetic population, choice set of activity chains ($\boldsymbol{\Gamma}$) and probability $\mathrm{Pr}_j^{o,p}(\boldsymbol{\alpha})$ are assumed to be given for this study. The formulation of $\mathrm{Pr}_j^{o,p}(\boldsymbol{\alpha})$ is described in section 4.1.1.

If a stochastic phenomenon is considered, activity chain demand can also be expressed as follows:

$$\mathbf{u} = \mathbf{u}(\boldsymbol{\alpha}) + \tau \tag{3.2}$$

where,

$\mathbf{u}$, $\mathbf{u}(\alpha)$: the true and expected activity chain demand, and

τ : the random error representing the unknown discrepancy between true demand $\mathbf{u}$ and expected demand $\mathbf{u}(\boldsymbol{\alpha})$ with zero mean and variance-covariance of demand **u**, $E(\tau\tau') = \Sigma_\tau$.

In addition, let OD pair *g* represent a trip (by specific mode (*mo*) and travel period (*m*)) from origin zone *o* to destination zone *d*. The OD demand matrix, **t**, can then be mapped by the activity chain demand matrix, **u**, as the following equation.

$$\mathbf{t} = \boldsymbol{\delta} \cdot \mathbf{u} \tag{3.3}$$

where,

$\boldsymbol{\delta}$: OD demand-activity chain demand proportion matrix, and

OD demand pair *g*: $t_g = \sum_j \delta_g^j \cdot u_j \quad \forall \, g$ (3.4a)

where,

δ_g^j : the number of OD pair *g* that contains in activity chain *j*.

After putting the definition (3.2) into (3.3) and taking expectation, the expected OD demand, $\mathbf{t}(\boldsymbol{\alpha})$, is obtained as follows :

$$\mathbf{t}(\boldsymbol{\alpha}) = \boldsymbol{\delta} \cdot \mathbf{u}(\boldsymbol{\alpha}) \tag{3.4b}$$

3.2 Supply Side

To consider *the stochastic supply side* of the activity-based model, the basic characteristics of link flow and path flow can be described as follows:

Let $\mathbf{R}_g$ be the non-empty path set of the OD pair *g*, and f_r^g be the traffic flow on path *r*, $r \in \mathbf{R}_g$, where the traffic flow can generally be obtained from *traffic assignment* (e.g. stochastic traffic assignment). In addition, let denote p_r^g as the proportion of OD pair *g* on path *r*. Following flow conservation' rule, traffic flow can then, also be expressed by:

$$f_r^g = p_r^g t_g \tag{3.5}$$

Let $\Delta_l^{r,g} = 1$ if path *r* of OD pair *g* uses link l, and 0 otherwise; and denote v_l as the traffic flow on link l. The link flow is then, the summation of traffic flows of all paths using the link:

$$v_l = \sum_g \sum_{r \in \mathbf{R}_g} \Delta_l^{r,g} f_r^g \quad \forall\ l \tag{3.6}$$

By replacing f_r^g from (3.5) to (3.6), the following can be obtained:

$$v_l = \sum_g \sum_{r \in \mathbf{R}_g} \Delta_l^{r,g} p_r^g t_g \tag{3.7}$$

Let $\mathbf{c}$ denote the link flow measurement vector. Since the error on link counts can be represented as the random variation in travel demand and route choices over time, link counts, $\mathbf{c}$, are assumed to be observations of the random variables as follows:

$$\mathbf{c} = \hat{\mathbf{v}} + \eta = B \cdot \mathbf{t}(\boldsymbol{\alpha}) + \eta \tag{3.8}$$

where,

B : link-OD proportion matrix (i.e. $b_l^g = \sum_{r \in \mathbf{R}_g} \Delta_l^{r,g} p_r^g$; $b_l^g = B[l, g]$), and

η : the random error with zero mean ($E(\eta) = 0$) and variance-covariance $E(\eta\eta') = \Sigma_\eta$.

3.3 Population Synthesis

The objective of population synthesis is to synthesize the whole agents from reference samples reproducing marginal distributions/totals of demographical variables (usually from census data) and joint distributions of these variables in the reference samples from household survey (Bowman, 2009). The iterative proportional fitting method (IPF), initially developed by Deming and Stephan (1940), is generally used to calculate the weights of each population type. The example of marginal totals and joint distributions in two-way tables is exemplified in Table 1.

Regarding the minimum discrimination information theorem (Ireland and Kullback, 1968) based on the IPF method, given marginal totals of attributes v1 and v2 (N_{v1} and N_{v2}), the number of synthetic populations with attributes v1 and v2, n_{v2}^{v1}, after *k* iterations can be estimated as:

$$\begin{aligned} \mathrm{n}_{v2}^{v1} &= \phi_{(v1,v2)} \cdot (\prod_{k'=1}^{k} w_{v1}^{k'}) \cdot (\prod_{k'=1}^{k} w_{v2}^{k'}) = \phi_{(v1,v2)} \cdot (w_{v1}) \cdot (w_{v2}) \\ &= \phi_{(v1,v2)} \cdot w_{(v1,v2)} \end{aligned} \tag{3.9}$$

where, $w_{v1}^0 = 1$ for k' =0; $w_{v1}^1 = \mathrm{N}_{v1} / \sum_{v2'} \phi_{(v1,v2')}$ for $k'=1$; $w_{v1}^{k'} = \mathrm{N}_{v1} / \hat{\mathrm{N}}_{v1}^{k'}$ for k' >1,

$w_{v2}^{0} = 1 \text{for}\, k' = 0$; $w_{v2}^{k'} = \mathrm{N}_{v2} / \hat{\mathrm{N}}_{v2}^{k'}$ for $k' \geq 1$,

$\phi_{(v1,v2)}$: the number of reference samples with attributes v1and v2,

$\hat{\mathrm{N}}_{v1}^{k'}, \hat{\mathrm{N}}_{v2}^{k'}$: the estimated marginal totals of attributes v1and v2 at iterative *k'*, i.e.,

$$\hat{\mathrm{N}}_{v1}^{k'} = \sum_{v2'} (\phi_{(v1,v2')} \cdot (\prod_{i=0}^{k'} w_{v1}^{i}) \cdot (\prod_{i=0}^{k'} w_{v2'}^{i})) ; \quad \hat{\mathrm{N}}_{v2}^{k'} = \sum_{v1'} (\phi_{(v1',v2)} \cdot (\prod_{i=0}^{k'} w_{v1'}^{i}) \cdot (\prod_{i=0}^{k'-1} w_{v2}^{i})) \cdot$$

w_{v1}, w_{v2} : the weight of attributes v1 and v2, respectively (after *k* iterations), and

$w_{(v1,v2)}$: the final weight of attributes v1 and v2, i.e. $w_{(v1,v2)} = w_{v1} \cdot w_{v2}$.

As Mosteller (1968) proofed that, based on the IPF method, a correlation structure of the synthetic population (e.g. n_{v2}^{v1}) is similar to that of observations in the reference sample. Regarding to the IPF example in Table 1, the correlation of synthetic population can be measured by cross-product ratio, i.e., $\mathrm{n}_{v2=1}^{v1=1} \cdot \mathrm{n}_{v2=2}^{v1=2} / \mathrm{n}_{v2=2}^{v1=1} \cdot \mathrm{n}_{v2=1}^{v1=2}$. After processing three iterations in the IPF process, the cross-product ratio of both synthetic population and the sample is equal to 0.823.

Table 1. Subtables for IPF example with two control variables (v1 and v2)

	Marginal totals					Reference samples ($\phi_{(v1,v2)}$)		
	v2 = 1	v2 = 2	Total(N_{v1})			v2 = 1	v2 = 2	Total
v1 = 1	?	?	1,000	=	v1 = 1	47	25	72
v1 = 2	?	?	2,500		v1 = 2	80	35	115
Total(N_{v2})	2,000	1,500	3,500		Total	127	60	187

x

	Marginal distribution				Final weights (after 3 iterations)		
	v2 = 1	v2 = 2	Total		v2 = 1	v2 = 2	Total
v1 = 1	?	?	0.286	v1 = 1	11.43	18.52	29.94
v1 = 2	?	?	0.714	v1 = 2	18.29	29.63	47.92
Total	0.571	0.426	1.000	Total	29.71	48.15	77.86

To calculate the marginal distributions of demographical control types normally obtaining from census data, let $\mathrm{N}_{\tilde{p}}$ be the number of population with demographical control type $\tilde{p}$, $\tilde{p} \in$ set of person control types, **P**, and N_{z} be the number of population in traffic zone *z*, z $\in$ **Z**. Note that types of person control are generally consisted of some demographical variables available from census data, but not all variables in the reference sample (e.g. only gender and household size, Table 5) are controlled. If $\mathrm{N}_{\tilde{p}}$ and N_{z} follow *multinomial distributions*, it can be expressed as:

Person control: $$\mathrm{N}_{\tilde{p}} \sim \text{Multinomial}\left(\mathrm{N}_{pop}, \mathrm{Pr}_{\tilde{p}}\right) \tag{3.10}$$

Zonal control: $$\mathrm{N}_{z} \sim \text{Multinomial}\left(\mathrm{N}_{pop}, \mathrm{Pr}_{z}\right) \tag{3.11}$$

Or, these control variables can also be presented by the fractions (marginal distributions) in each control level (person control and zonal control) as follows.

$$\mathrm{Pr}_{\tilde{p}} = \frac{\mathrm{n}_{\tilde{p}}}{\sum_{\tilde{p}' \in \mathbf{P}} \mathrm{n}_{\tilde{p}'}} \tag{3.12}$$

where,

N_{pop} : the total number of population,

$\mathrm{Pr}_{\tilde{p}}$: the success probability of selecting population with person control type $\tilde{p}$,

$\mathrm{n}_{\tilde{p}}$: the expected number of population with person control type $\tilde{p}$, and

$$\mathrm{Pr}_z = \frac{\mathrm{n}_z}{\sum_{z' \in \mathbf{Z}} \mathrm{n}_{z'}} \tag{3.13}$$

where,

Pr_z : the success probability of selecting population in traffic zone *z*, and

n_z : the expected number of population in traffic zone *z*.

In addition, the number of control type $\tilde{p}$ population, $\mathrm{n}_{\tilde{p}}$, is generally available from the census block level including the marginal totals of population of each control types (e.g. gender and household size in the numerical example). However, in this study, the population data (n_z) is assumed to be unknown at the traffic zone level (sub area of census block). To deal with this problem, the number of population at the traffic zone level can be pre-estimated by using land use data. For instance, the type of living quarters and related household size can be used to estimate the expected number of population at traffic zone as follows.

$$\mathrm{n}_z = \theta_z \sum_{s \in C_z} \mathrm{n}_{s,z} \gamma_s \tag{3.14}$$

where,

$\mathrm{n}_{s,z}$: the number of building of living quarters *s* in traffic zone *z*,

θ_z : the bias term of estimated number of population in traffic zone *z*, and

γ_s : the average household size of living quarters *s*.

Note that if θ is assumed to be constant for every traffic zones, all bias terms in (3.14) can then be cancelled. Consequently, the probability of drawing person in traffic zone *z*, Pr_z, can be formulated as follows.

$$\mathrm{Pr}_z = \frac{\mathrm{n}_z}{\sum_{z' \in \mathrm{Z}} \mathrm{n}_{z'}} \sim \frac{\sum_{s \in C_z} \mathrm{n}_{s,z} \gamma_s}{\sum_{z'} \sum_{s' \in C_{z'}} \mathrm{n}_{s',z'} \gamma_{s'}} \qquad \text{for} \quad z, z' \in \mathbf{Z} \tag{3.15}$$

where,

C*z* : the set of living quarter types in traffic zone *z*.

According to the multinomial properties, the split fractions of population in control type $\tilde{p}$, $y_{\tilde{p}}$, and traffic zone *z*, y_z, also follow multinomial distribution as shown in (3.16) and (3.17).

$$y_{\tilde{p}} \sim \mathrm{Multinomial}\left(1, \mathrm{Pr}_{\tilde{p}}\right) \tag{3.16}$$

$$y_z \sim \mathrm{Multinomial}\left(1, \mathrm{Pr}_z\right) \tag{3.17}$$

where the variances of split fraction (y_z and $y_{\tilde{p}}$) can be expressed by:

$$\left(\sigma_z^y\right)^2 = \frac{\mathrm{Pr}_z\left(1 - \mathrm{Pr}_z\right)}{\mathrm{N}_{pop}} \tag{3.18}$$

$$\left(\sigma_{\tilde{p}}^y\right)^2 = \frac{\mathrm{Pr}_{\tilde{p}}\left(1 - \mathrm{Pr}_{\tilde{p}}\right)}{\mathrm{N}_{pop}} \tag{3.19}$$

Note that for a large population size, these split fractions can be assumed to follow *the normal distribution*, according to the central limit theorem.

3.4 Maximum-likelihood Estimation Problem (MLP)

The optimization problem of generating the weight/expansion factor, *w,* for any synthetic population located in the traffic zone *z* simultaneously calibrating with link count *c* can be considered as follows:

$$\text{Min: } Z(w) = Z_1(w|\mathrm{N}_{s,z},\gamma_s) + Z_2(w|\Pr_{\tilde p}) + Z_3(w|c) \tag{3.20}$$

$$\text{subject to: } \mathrm{w}_{\tilde p} \geq 1,\ \mathrm{w}_z \geq 1 \ \text{ for } \ \forall\ \tilde p, z \tag{3.21}$$

$$\text{Marginal zonal control: } Z_1(w|\mathrm{N}_{s,z},\gamma_s) = 0.5\sum_z \frac{\left(\Pr_z - \dfrac{\mathrm{n}_z}{\mathrm{N}_{pop}}\right)^2}{\left(\sigma_z^y\right)^2} + \text{const.} \tag{3.22}$$

$$\text{where,}\quad \mathrm{n}_z = \sum_p (\phi_{(p,z)}(\sum_{\tilde p} w_{\tilde p}\Delta_p^{\tilde p})w_z) \tag{3.22a}$$

$$\text{Marginal person control: } Z_2(w|\Pr_{\tilde p}) = 0.5\sum_{\tilde p} \frac{\left(\Pr_{\tilde p} - \dfrac{\mathrm{n}_{\tilde p}}{\mathrm{N}_{pop}}\right)^2}{\left(\sigma_{\tilde p}^y\right)^2} + \text{const.} \tag{3.23}$$

$$\text{where,}\quad \mathrm{n}_{\tilde p} = \sum_z \sum_p (\phi_{(p,z)}(\sum_{\tilde p} w_{\tilde p}\Delta_p^{\tilde p})w_z) \tag{3.23a}$$

For a simple illustration of the calibration method, if errors in traffic counts are assumed to have a joint multi variate normal (MVN) distribution with zero means and ignoring the covariance terms in dispersion matrix, Σ_η , (e.g. Cascetta and Russo, 1997). It follows that:

$$\text{Link count control: } Z_3(w|c) = 0.5\sum_{\bar a}\sum_l \frac{\left(c_l^{\bar a} - \sum_{z,p} \mathrm{n}_z^p \sum_j \Pr_j^{z,p}(\boldsymbol{\alpha})\hat p_l^j\right)^2}{\sigma_c^2} + \text{const.} \tag{3.24}$$

$$\text{where,}\quad \mathrm{n}_z^p = \phi_{(p,z)}(\sum_{\tilde p} w_{\tilde p}\Delta_p^{\tilde p})w_z\,, \tag{3.24a}$$

the link-activity chain proportion, $\hat p_l^j = \sum_{r,g} \delta_g^j \Delta_l^{r,g} p_r^g$; $\mathrm{N}_{pop} = \sum_{\tilde p} \mathrm{n}_{\tilde p} = \sum_z \mathrm{n}_z$.

p : the type of individuals having the same personal and household demographic,

$\Delta_p^{\tilde p}$: indicator variable, $\Delta_p^{\tilde p}$ = 1: type *p* population is in control type $\tilde p$, and $\Delta_p^{\tilde p}$ = 0, otherwise.

n_p : the expected number of population belonging to type *p*,

N_{pop} : total number of population of all population types,

n_z : the expected number of population in traffic zone *z*,

n_z^p : the expected number of population belonging to type *p* in traffic zone *z*,

$c_l^{\bar a}$: link flow in link *l* at day $\bar a$,

$w_{\tilde p}$: weight of population in control type $\tilde p$,

w_z : weight of any person living in traffic zone *z*, and

$\phi_{(p,z)}$: the number of type *p* samples living in traffic zone *z* (in the reference sample).

To solve problem (3.20) associated with the constraint (3.21), the Lagrangian is formulated as follows:

$$L(w) = Z_1(w|\mathrm{N}_{s,z},\gamma_s) + Z_2(w|\mathrm{Pr}_{\tilde{p}}) + Z_3(w|c) + \mu_1 \cdot (w_{\tilde{p}} - 1) + \mu_2 \cdot (w_z - 1) \qquad (3.25)$$

where, μ_1, μ_2 : the Lagrange multipliers.

As the matrices of variance $(\sigma_z^y)^2, (\sigma_{\tilde{p}}^y)^2$, and σ_c^2 are positive-definite, the problem (3.25) is convex and the minimum has unique solution of final weight w_k (i.e. $w_k = w_{\tilde{p}} w_z$) as follows:

$$\frac{\partial^2 Z_1(w)}{\partial\, w_k w_k} > 0, \quad \frac{\partial^2 Z_2(w)}{\partial\, w_k w_k} > 0, \quad \frac{\partial^2 Z_3(w)}{\partial\, w_k w_k} > 0, \ \forall\, k \qquad (3.26)$$

$$\text{and,} \quad \frac{\partial^2 Z_1(w)}{\partial\, w_k w_{k'}} = 0, \quad \frac{\partial^2 Z_2(w)}{\partial\, w_k w_{k'}} = 0, \quad \frac{\partial^2 Z_3(w)}{\partial\, w_k w_{k'}} = 0, \ \forall\, k' \neq k \qquad (3.27)$$

In this study, the sequential quadratic programming in MATLAB software was adopted for solving this optimization problem. Then, after calibrating the final weights (w_k) by solving the optimization problem (3.25), the number of simulated type p population in particular traffic zone z, $\hat{\mathrm{N}}_z^p$, can be obtained from multiplying the final weights by the reference sample of the same demographic and traffic zone. This is expressed by:

$$\hat{\mathrm{N}}_z^p = \sum_i \phi_{(i,p,z)} \Big(\sum_{\tilde{p}} w_{\tilde{p}} \Delta_p^{\tilde{p}}\Big) w_z \qquad (3.28)$$

where,

$\phi_{(i,p,z)}$: the number of person ID i (in reference sample) with demographic p in traffic zone z.

It is worth noting that, to avoid the common problem that the estimated marginal totals of demographic p, $\hat{\mathrm{N}}_z^p$, create the zero entry due to no samples of this demographic from the reference sample record (e.g. Beckman *et al.*, 1996; Guo and Bhat, 2007), at least one sample of such a demographic type should be collected by the household sample survey.

4. NUMERICAL EXAMPLE

4.1 Test Activity-based Model and Network

A random utility maximization-based approach was used to construct an activity-based model (e.g. Bradley *et al.*, 2010; Bifulco *et al.*, 2010). The choice dimensions for this application are:

- activity pattern choice;
- tour choices (trip chain model), consisting of:
  (a) first tour:
    (i) time-of-day choice;
    (ii) destination and mode choice
  (b) second tour (optional)
    (i) time-of-day choice;
    (ii) destination and mode choice.

Note that time-of-day choice is a combined choice of start tour and end tour travel period.

4.1.1 Model specification

In this study, the numerical example is considered as illustrative, and the model was applied to a single category: workers associated with travel pattern of *home-based tour*. The choice alternation of out-of-home activities includes work (W) and maintenance (O) purposes. In addition, at-home activities include stay-at-home (H) purposes.

Daily activity pattern choice model (Ap): This model reproduces the choice of daily activity pattern, *y*, for each origin zone *o*. In this numerical example, four different activity patterns (i.e. $y \in \{1,2,3,4\}$) are considered: H-W-H ($y=1$), H-O-W-H ($y=2$), H-W-O-H ($y=3$), and H-W-H-O-H, (y = 4).

First tour time-of-day choice model (Ftod), reproduces the choice of the time-of-day t_1 in the first tour (with $t_1 \in \{1, 2, 3\}$, see Table 2).

Second tour time-of-day choice model (Stod), reproduces the choice of the time-of-day t_2 in the second tour (only for H-W-H-O-H). The choice set of this second tour dimension is considered as a function of the time constraints of the first tour (if the first tour has not ended, the second tour cannot start, see Table 2).

Table 2. Time-of-day alternatives (first and second tour)

	First tour			Second tour	
t_1	Start	End	t_2	Start	End
1	AM[a]	MD[a]	1	PM[a]	PM
2	AM	PM	2	PM	OP
3	AM	OP[a]	3	OP	OP

[a] AM = 7:00-9:00, MD = 12:30-14:30, PM = 17:30-19:30, and OP = 20:00 – 22:00.

Destination and mode choice model for the first tour (Fdm) *and second tour* (Sdm), reproduces the choice of the destination zone *lo* for work purpose in the first tour and for maintenance purpose in the second tour (with *lo* ∈ {1: for 1st nearest zone, 2: for 2nd nearest zone, 3: for 3rd nearest zone to origin (but not including it)}) and travel mode *mo* (with *mo* ∈ {Car, Bus}). Consequently, the combination of mode and destination for the first tour (b_1) and the second tour (b_2) can be up to 6 alternatives. Note that the destination for maintenance purposes before/after work in activity pattern (H-O-W-H and H-W-O-H) is assumed to be located at the same work zone for a simple illustrative purpose. Also, due to spatial independency of the destination for performing work in first tour and maintenance in second tour, the MNL models are used to generate the trip maker's decisions on destination and mode in first and second tour.

The mathematical formulation of the probability that type *p* population originating from zone *o* performs activity chain *j* is shown as follows:

$$\Pr_j^{o,p}(\boldsymbol{\alpha}) = \Pr(b_2 | y, t_1, b_1, t_2) \Pr(t_2 | y, t_1, b_1) \Pr(b_1 | y, t_1) \Pr(t_1 | y) \Pr(y) \tag{4.1}$$

where the probability of selecting an activity pattern *y*:

$$\Pr(y) = \frac{\exp(V_y + V'_y)}{\sum_{y'} \exp(V_{y'} + V'_{y'})} ; V'_y = \frac{1}{\mu^{t_1}} \ln \sum_{t'_1} \exp[(V_{t'_1} + V'_{y,t'_1}) \mu^{t_1}] ; V_y = \mathrm{Asc}_1 + \alpha_6 \mathrm{X}_6 + \alpha_7 \mathrm{X}_7 + \alpha_8 \mathrm{X}_8 \tag{4.2}$$

, the probability of selecting time-of-day t_1 in the first tour:

$$\Pr(t_1 | y) = \frac{\exp(V_{t_1} + V'_{y,t_1})}{\sum_{t'_1} \exp(V_{t'_1} + V'_{y,t'_1})} ; \quad V'_{y,t'_1} = \frac{1}{\mu^{t_2}} \ln \sum_{t'_2} \exp[(V_{t'_2}) \mu^{t_2}] ; \quad V_{t_1} = \mathrm{Asc}_2 + \alpha_{11} \mathrm{X}_{11} + \alpha_{12} \mathrm{X}_{12} \tag{4.3}$$

, the probability of selecting destination/mode choice model b_1 in the first tour:

$$\Pr(b_1|y,t_1) = \frac{\exp(V_{b_1})}{\sum_{b_1'}\exp(V_{b_1'})} \quad ; V_{b_1} = \mathrm{Asc}_3+\alpha_{14}\mathrm{X}_{14}+\alpha_{15}\mathrm{X}_{15}+\alpha_{16}\mathrm{X}_{16} \tag{4.4}$$

, the probability of selecting time-of-day t_2 in the second tour:

$$\Pr(t_2|y,t_1,b_1) = \frac{\exp(V_{t_2})}{\sum_{t_2'}\exp(V_{t_2'})} \quad ; V_{t_2} = \mathrm{Asc}_4+\alpha_{19}\mathrm{X}_{19}+\alpha_{20}\mathrm{X}_{20} \tag{4.5}$$

, the probability of selecting destination/mode choice model b_2 in the second tour:

$$\Pr(b_2|y,t_1,b_1,t_2) = \frac{\exp(V_{b_2})}{\sum_{b_2'}\exp(V_{b_2'})} \quad ; V_{b_2} = \mathrm{Asc}_5+\alpha_{14}\mathrm{X}_{14}+\alpha_{23}\mathrm{X}_{23}+\alpha_{24}\mathrm{X}_{24} \tag{4.6}$$

, and μ^{t_1}, μ^{t_2} : scaled parameters at time of day: level t_1 and t_2 respectively.

The test network has 5 traffic zones (including activity H, W, and O in each zone), 38 links, and 13 nodes (see Figure 1). There are 2 available modes in the network (car and bus). Feasible paths travelled by both car and bus are assumed to be generated by the k-shortest path method (Yen, 1971). In accordance with method to assign travel demand on the test network, a logit stochastic user equilibrium (SUE) was adopted in this study. Given route choice parameters, true traffic flows were then obtained by assigning the travel demand derived from the synthetic population associated with true model parameters (see Table 3) to the test network.

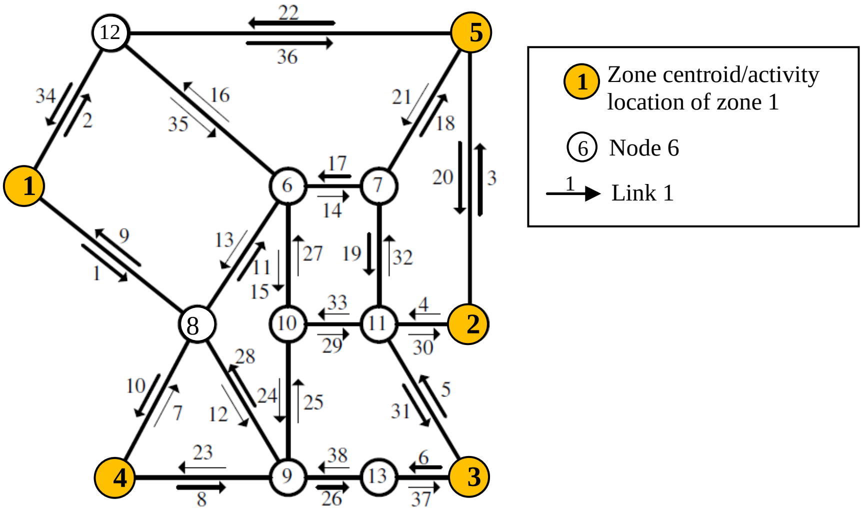


Figure 1. Test network

4.2 Evaluation Method and Test Case

To set the test cases, true population allocation in traffic zone (Pr_z) is given (column 10 in Table 4). The initial population allocation (Pr_z) set to deviate from the true population allocation representing the biases in the population synthesis is calculated by using exogenous land use data (column 9 in Table 4). In addition, true marginal totals ($\mathrm{n}_{\tilde{p}}$) and marginal distributions of person control variables ($\mathrm{Pr}_{\tilde{p}}$) categorized by 6 types are given (Table 5). Since the population size is high, the variances $(\sigma_z)^2$ and $(\sigma_{\tilde{p}})^2$ can be calculated by using true values of Pr_z and $\mathrm{Pr}_{\tilde{p}}$ described in (3.18) and (3.19).

Regarding the simulations of roadside observations, a Monte-Carlo method was used to draw link flow observations from the activity-based models described in section 4.1, given true parameters α^{true} (Table 3) and true population data. Link counts, *c,* are then drawn from a normal variate with a mean, $\hat{\mathbf{v}}$, equal to $B \cdot \mathbf{t}(\boldsymbol{\alpha})$ in (3.8) and a variance, σ_c^2, assumed to be a function of mean, $(\mathrm{vc} \cdot \hat{\mathbf{v}})^2$, where vc is the coefficient of link count variations. In addition, to representing the variations in an observation, link counts are randomly drawn 10 times (days). This describes a variation from multiday observations.

To evaluate the performance of calibration method, the true final weights (w^{true}) are calculated by solving the problems (3.20) to (3.23) associated with true population allocation (Pr_Z) in the traffic zone (column 10 in Table 4). In the case of no calibration with link counts, the initial final weights (w^0) in population synthesis process are obtained by solving the problems (3.20) to (3.23) associated with the initial population allocation in the traffic zone (column 9 in Table 4). To calibrate the synthetic population from link counts, the calibrated final weights ($\hat{w}$) are obtained by solving the problems (3.20) to (3.24) associated with the initial population allocation in the traffic zone and simulated link counts. The reference sample (see Table 6) is then expanded by multiplying the final weight estimated from the previous step to reproduce the synthetic population data as described in (3.28).

In order to see how the proposed method based on link flow information can improve the population synthesis efficiency as mentioned, the final weights were calculated, which is summarized as the following simulation scenarios:

(i) *Based case: no calibration*. This implies solving the problem (3.20)–(3.23).

(ii) *Calibrated by link counts*. This implies solving the problem (3.20)–(3.23) associated with the link count control (3.24).

In general, the efficiency of model calibration depends on quantity (number of links to be observed) of observations. The number of links to be counted in scenario (ii) is typically set to be 40%. However, lack of observation stations can cause low efficiency in the population calibration results. Consequently, the sensitivity analysis by various settings of numbers of link counts (i.e. approximately 40%, 25%, and 15% of all links to be observed) with fixed coefficient of link count variations, (i.e. vc = 0.01) was also adopted for this study.

Let $\boldsymbol{\Lambda}^0$ be the vector of the initial value of the final weight of scenario (i), $\hat{\boldsymbol{\Lambda}}$ be the vector of the calibrated value of final weight of scenario (ii), and $\boldsymbol{\Lambda}^{true}$ be the vector of the true final weight. The statistical performance of the estimation of the final weight, w, can be measured by the percentage reduction of the mean square error from the initial value of the final weight, w^0, in vector $\boldsymbol{\Lambda}^0$ and defined as follows (Cascetta and Russo, 1997).

$$\mathrm{MSE\%}(w_k) = [\mathrm{MSE}(w_k^0) - \mathrm{MSE}(w_k)] / \mathrm{MSE}(w_k^0) \cdot 100\% \tag{4.7}$$

$$\text{,and MMSE} = \sum_k \frac{\mathrm{MSE\%}(w_k)}{K} \tag{4.8}$$

where,

N : total number of trials of a dataset ($\boldsymbol{\Lambda}^0$ and $\hat{\boldsymbol{\Lambda}}$),

K : number of final weights in vector $\boldsymbol{\Lambda}$, and

$\mathrm{MSE\%}(w_k^0)$: percentage reduction of the mean square error of final weight k^{th} in vector $\boldsymbol{\Lambda}$.

$$\mathrm{MSE}(w_k^0) = \sum_{n=1}^{N} (w_{n,k}^0 - w_k^{true})^2 / N \tag{4.9}$$

$$\mathrm{MSE}(w_k) = \sum_{n=1}^{N} (\hat{w}_{n,k} - w_k^{true})^2 / N \tag{4.10}$$

where,

w_k^{true} : final weight k^{th} in true vector $\mathbf{\Lambda}^{true}$,

MSE(w_j^0): mean square error of initial value of weight k^{th} in $\mathbf{\Lambda}^0$ from N datasets; $w_{n,k}^0$ is initial value of final weight k^{th} in $\mathbf{\Lambda}^0$ of trial n^{th}, and

MSE(w_k): mean square error of calibrated value of final weight k^{th} in $\hat{\mathbf{\Lambda}}$ from N datasets; $\hat{w}_{n,k}$ is calibrated value of final weight k^{th} in $\hat{\mathbf{\Lambda}}$ of trial n^{th}.

Table 3. The setting of activity-based model parameters (α)

No.	Model	Variable name	Type of variable	Coefficient(α)
1	Ap	H-O-W-H specific const. (Asc_1 of H-W-H = 0)	Asc_1	-4.42
2		H-W-O-H specific const.	Asc_1	-3.12
3		H-W-H-O-H specific const.	Asc_1	-0.84
4		Scaled parameter (μ^{t1})	μ^{t1}	1.28
5		Scaled parameter (μ^{t2})	μ^{t2}	1.00
6		Dummy[a]: Female + H-O-W-H or H-W-O-H	X_6 = 1 or 0	1.56
7		Dummy: Family with at least one child+ H-O-W-H	X_7 = 1 or 0	3.16
8		Dummy: Family without child + H-W-O-H	X_8 = 1 or 0	0.70
9	Ftod	AM to PM specific const. (Asc_2 of AM-MD = 0)	Asc_2	-0.54
10		AM to OP specific const.	Asc_2	-2.24
11		Dummy: Full time worker + tour time: AM to MD	X_{11} = 1 or 0	-4.00
12		Dummy: Part-time worker + tour time: AM to OP	X_{12} = 1 or 0	1.92
13	Fdm	Bus specific const. (Asc_3 of car = 0)	Asc_3	0.24
14		Generalised travel time	X_{14}	-0.08
15		Number of employments (log scale)	X_{15}	0.80
16		Dummy: High household income + car mode	X_{16} = 1 or 0	5.36
17	Stod	PM to OP specific const. (Asc_4 of PM-PM = 0)	Asc_4	-1.06
18		OP to OP specific const.	Asc_4	-2.82
19		Dummy: Full time worker + tour time: PM to PM	X_{19} = 1 or 0	-4.28
20		Dummy: Part-time worker + tour time: OP to OP	X_{20} = 1 or 0	2.70
21	Sdm	Bus specific const. (Asc_5 of car = 0)	Asc_5	-0.68
22		Generalised travel time	X_{22}	-0.98
23		Number of retails (log scale)	X_{23}	0.84
24		Dummy: High household income + car mode	X_{24} = 1 or 0	6.74

[a] Dummy variable = 1 if specific demographic/choice is selected, 0 otherwise.

Table 4. Initial estimation of population by traffic zone (Marginal zonal control)

Zone	living quarters						Total	Initial Pr_z	True Pr_z
	Detached house		Town house		Row house				
	No. of building	Average household size	No. of building	Average household size	No. of building	Average household size			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)[a]	(9)	(10)
1	61	2.00	55	3.50	85	4.50	699	0.200	0.158
2	55	2.00	52	3.50	92	4.50	703	0.201	0.159
3	88	2.00	31	3.50	85	4.50	668	0.191	0.152
4	52	2.00	61	3.50	88	4.50	716	0.204	0.266
5	34	2.00	122	3.50	49	4.50	714	0.204	0.265
Total	290		321		399		3,500	1.000	1.000

[a] (8) = (2)*(3)+(4)*(5)+(6)*(7).

Table 5. The population by types (Marginal person control)

Index	Person control type						Total
	Household size = 1		Household size = 2		Household size = 3		
	Male	Female	Male	Female	Male	Female	
Number of population	525	525	700	700	525	525	3,500
True $\Pr_{\tilde{p}}$	0.15	0.15	0.20	0.20	0.15	0.15	1.00

Table 6. The example of personal records from the reference sample

Person ID	Zone	Household size[c]	Gender[c]	Family with childs	Household income	Career
1	1	3	Male	Yes	High	FT[a]
2	1	1	Female	Yes	Low	PT[b]
3	1	3	Male	Yes	Low	FT
4	1	2	Female	No	High	FT
5	1	2	Male	No	High	FT

[a] FT: Full-time worker. [b] PT: Part-time worker.

[c] two of five demographical variables (gender and household size) are assumed to be the control variables.

4.3 Analysis of the Results

Two simulation scenarios described above in the numerical test were conducted. The estimation of the final weights calibrated by link counts is generally more satisfactory than the estimation results in scenario (i). The numerical details of the calibrated results in scenario (ii), reported in Table 7 and Table 8, show the smaller level of error than the estimated results from the case without calibration in scenario (i). This error is measured by the mean square error (MSE) as described in section 4.2. To make a comparison of the error between scenarios (i) and (ii), the statistical performance of the calibration with link counts can also be interpreted by the high percentage error reduction shown in Table 9 (MSE% is close to 100%). After assigning the final weights to reproduce the synthetic population (3.28), the calibrated results are more closely related to the true population than the initial population (Figure 2).

Effect of the number of observations

As with the quantity of observation experiment, the effects of the decrease in the number of links to be observed, covering from 40% to 15% of all links in the test network, were also studied. The results from Table 10 show that, the lower the number of links to counts, the higher the estimation error (measured by MMSE between scenario (i) and (ii)) obtained. However, with 15% of links to be observed, the mean error reduction from the initial values (MMSE) is still higher than 80%. In other words, the calibrated model error is still significantly smaller than the model error without calibration, under a case that the sufficient number of links is observed.

Effect of the variation in observations

In general, it was observed (see Table 11) that the coefficient of variation, vc, exert opposite influences on the direct indicators: the mean percentage reduction of the mean square error (MMSE). Increasing the coefficient of variation, vc, values, and hence deviations of measured and true flows, cause a decrease in the estimation precision.

Table 7. The estimation results of final weights after 5 trials (N = 5)

Person control type	Household size[a]	Gender	Traffic zone				
			1	2	3	4	5
Initial final weight (Scenario (i): without link count calibration)							
1	1	Male	4.460	4.428	4.198	4.411	4.416
2	1	Female	5.587	5.548	5.260	5.526	5.533
3	2	Male	5.302	5.265	4.991	5.244	5.251
4	2	Female	5.069	5.034	4.772	5.014	5.020
5	3	Male	4.577	4.545	4.309	4.527	4.533
6	3	Female	5.530	5.491	5.206	5.470	5.476
True final weight							
1	1	Male	3.528	3.533	3.343	5.899	5.880
2	1	Female	4.413	4.419	4.181	7.379	7.354
3	2	Male	4.221	4.227	3.999	7.058	7.035
4	2	Female	3.993	3.998	3.783	6.676	6.654
5	3	Male	3.619	3.624	3.429	6.052	6.032
6	3	Female	4.143	4.149	3.926	6.928	6.905
Calibrated final weight (Scenario (ii): with link count calibration)							
1	1	Male	3.514	3.515	3.326	5.874	5.851
2	1	Female	4.406	4.407	4.170	7.365	7.336
3	2	Male	4.249	4.250	4.021	7.102	7.074
4	2	Female	4.038	4.039	3.822	6.749	6.723
5	3	Male	3.581	3.581	3.389	5.985	5.961
6	3	Female	4.159	4.160	3.936	6.951	6.924

Table 8. The statistical performance (measured by MSE) after 5 trials (N = 5)

Person control type	Household size[a]	Gender	Traffic zone				
			1	2	3	4	5
MSE of initial final weight (Scenario (i): without link count calibration)							
1	1	Male	0.868	0.801	0.732	2.216	2.142
2	1	Female	1.379	1.274	1.163	3.434	3.318
3	2	Male	1.169	1.077	0.984	3.291	3.183
4	2	Female	1.159	1.072	0.979	2.764	2.670
5	3	Male	0.918	0.848	0.774	2.324	2.246
6	3	Female	1.923	1.801	1.639	2.127	2.041
MSE of calibrated final weight (Scenario (ii): with link count calibration)							
1	1	Male	0.026	0.029	0.034	0.081	0.088
2	1	Female	0.009	0.011	0.015	0.029	0.035
3	2	Male	0.158	0.150	0.112	0.421	0.394
4	2	Female	0.012	0.010	0.005	0.027	0.024
5	3	Male	0.059	0.063	0.068	0.175	0.186
6	3	Female	0.011	0.012	0.015	0.036	0.036

Table 9. The statistical performance (measured by MSE%) after 5 trials (N = 5)

Person control type	Household size	Gender	Traffic zone				
			1	2	3	4	5
1	1	Male	99.7	99.6	99.6	99.7	99.6
2	1	Female	99.9	99.9	99.9	99.9	99.9
3	2	Male	99.3	99.3	99.3	99.3	99.3
4	2	Female	99.5	99.6	99.6	99.5	99.6
5	3	Male	98.8	98.5	98.5	98.6	98.3
6	3	Female	98.8	98.8	98.8	97.1	97.2

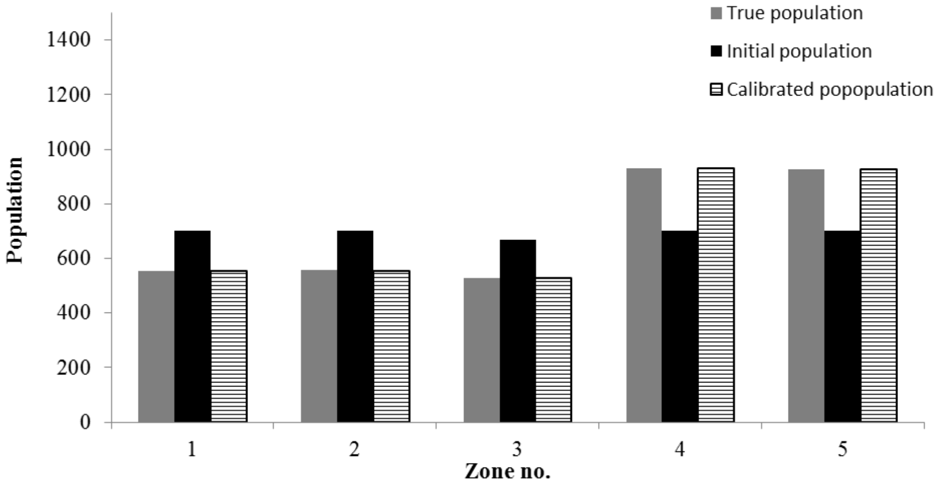


Figure 2. The comparison of true population and calibrated population by traffic zone

Table 10. Results of the various settings of number of observations

Number of links observed =	40%	25%	15%
MMSE.	99.2	95.5	83.5

Table 11. Results of the various settings of the variation in observations

Coefficient of link count variations, vc, =	0.005	0.01	0.10	0.20	0.40
MMSE.	99.9	99.2	60.4	33.4	12.2

5. AN APPLICATION TO THE MEDIUM-SIZED CITY

Phisanulok city located in northern part of Thailand has 78 traffic zones, 846 links, and 329 nodes (Figure 3). The population in this city is approximately 200,000. 10% of links is assumed to be counted (by four travel periods (AM, MD, PM, and OP period)) and used to calibrate synthetic population. Given route choice proportions, these link counts are simulated from demand model (3.4b) derived from given true population data and true activity-based model (ABM) parameters. In this test, the true ABM parameters are assumed to be equal to the estimated ABM parameters obtaining from another ABM parameter estimation problem based on household and travel sample survey (HTS) data in the previous study (Siripirote *et al.,* 2012). In addition, the choice set of activity patterns is also derived from HTS data. The study area includes two districts (i.e. CBD district: zone 1-35 and sub-urban district: zone 36-78). Based on the complete census data in this city (Thai national statistical office, 2010), the population can be categorized by 18 population types of gender and age (see Table 12). The population of each person control type obtained from the complete census data is presented in Figure 4. The reference sample, which is approximately 3% of the total population, was collected from household and travel sample survey as mentioned. The initial population allocation was pre-estimated by using the land use data as illustrated in the previous numerical example. The land use data in this city includes the number of building categorized by 8 building types such as detached house, town house, and row house. The average household size per building type is obtained from the census data (Thai national statistical office, 2010).

To consider the performance of the proposed model calibration, Figure 4 shows that the estimation of synthetic population calibrated by link flows (calibrated population) is generally closer to the true population than synthetic population without calibration (initial population). Figure 5 also shows that, after assigning the synthetic population to the network, the link flow estimation error with model calibration (measured by root mean square error (RMSE)) is significantly smaller than that of link flows without calibration.

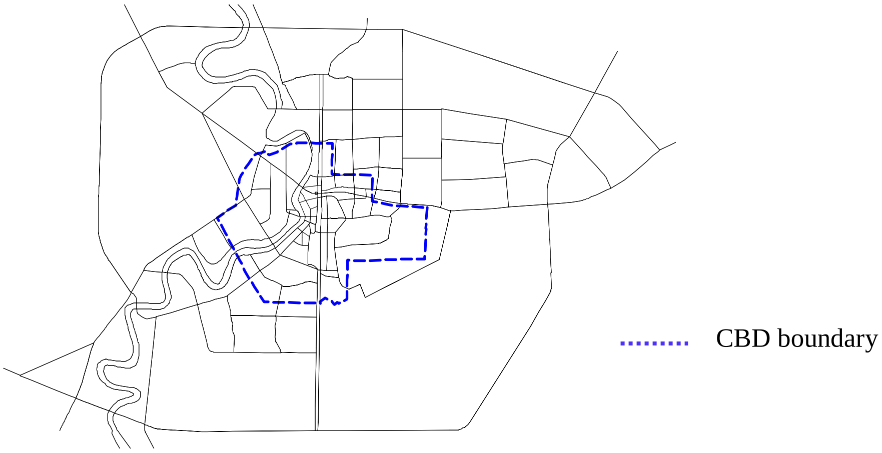


Figure 3. The traffic network representation of Phitsanulok city (Thailand)

Table 12. The population types

Population type	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Gender[a]	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F
Age (Yrs.)	0-9		10-19		20-29		30-39		40-49		50-59		60-69		70-79		≥80	

[a] Gender: M=male and F=female.

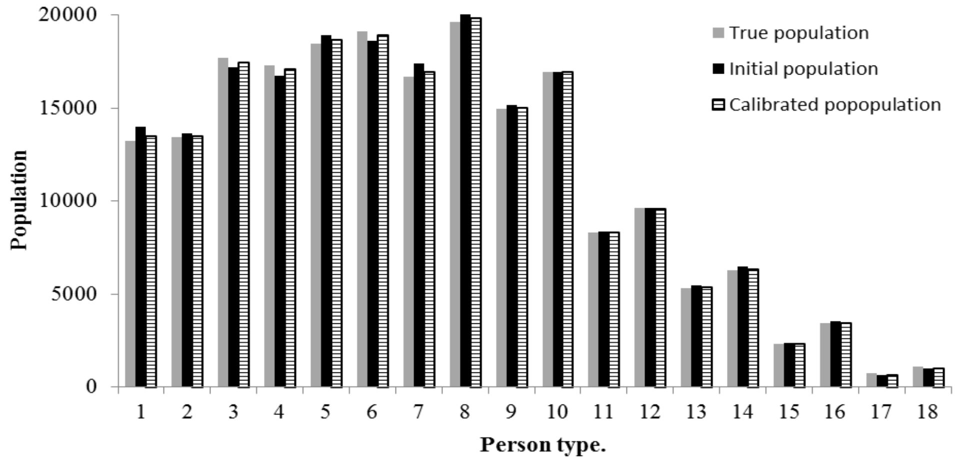
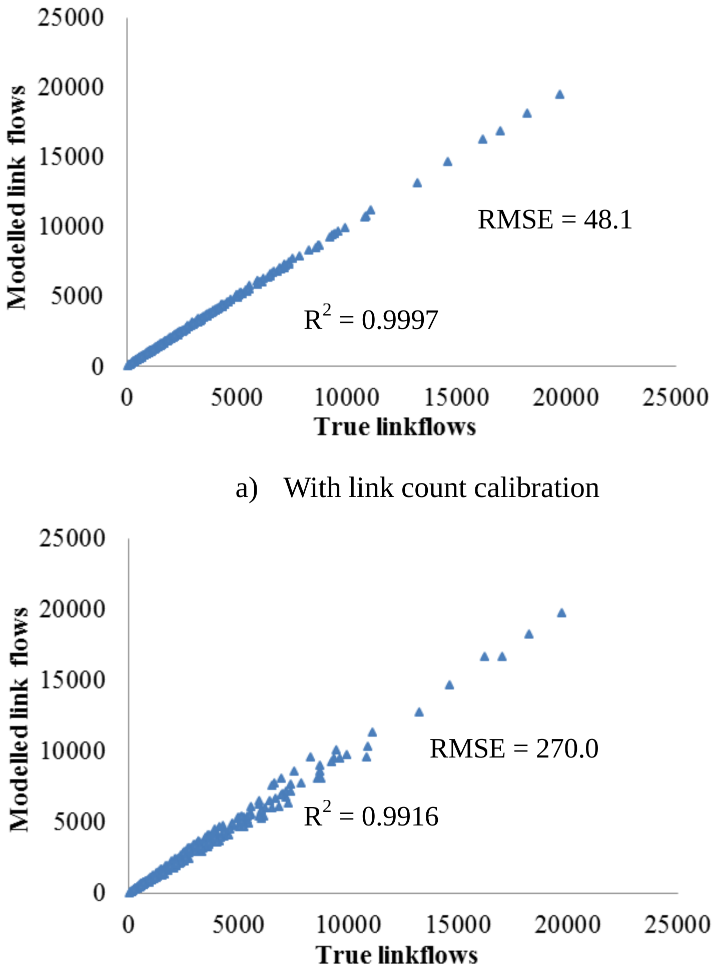


Figure 4. The comparison of true population and calibrated synthetic population



a) With link count calibration

b) Without link count calibration

Figure 5. The comparison of true and calibrated link flows

6. CONCLUSIONS

A statistical method for calibrating a synthetic population from link counts on the basis of the aggregate dataset only available in the person control level was presented in this paper. Approaches based on the maximum likelihood estimation method were examined and the statistical performances of this method were evaluated on a test network with various numbers of observations and medium-sized network application.

The calibration results are, in general satisfactory, showing the capability of the proposed method to significantly reduce a bias in synthetic population estimated from household sample surveys (the reference sample) and land use data (population allocation into traffic zone). Based on the synthetic datasets used in the test network, low numbers of observation stations and high link count variations can reduce the performance of the proposed calibration method. However, the calibrated final weights from link flows, with a sufficient number of links to be observed, have significantly less errors than the model without calibration. Also, the accuracy of model calibrations depends on link counting locations, the future research may find an optimal link counting location (e.g. Siripirote *et al.*, 2013) giving the reliable roadside observation data to reproduce the good calibration results. Another future research may consider to calibrate a synthetic population associated with the combined aggregate datasets of both person and household control levels.

Acknowledgements

The authors are indebted to the General Research Fund of the Hong Kong Research Council (RGC) (Project: PolyU 5290/09E) for support of this research.

REFERENCES

Arentze, T.A., Timmermans, H.J.P. (2004) A learning-based transportation oriented simulation system. *Transportation Research Part B*, 38(7), 613–633.

Arentze, T.A., Timmermans, H.J.P., Hofman, F. (2007) Population synthesis for micro simulating travel behavior. *Transportation Research Record*, 2014(11), 85–91.

Axhausen, K.W., Müller, K. (2011) Population Synthesis for Microsimulation: State of the Art, Paper presented at the 90th Annual Meeting of the Transportation Research Board, Washington, D.C., January 23-27.

Beckman, R.J., Baggerly, K.A., McKay, M.D. (1996) Creating synthetic baseline populations. *Transportation Research Part A*, 30 (6), 415–429.

Bhat, C.R., Koppelmen, F.S. (1999) *Handbook of Transportation Science*. Kluwer academic publishers, New York.

Bhat, C.R., Guo, J.Y., Srinivasan, S., Sivakumar, A. (2004) A comprehensive econometric microsimulator for daily activity-travel patterns. *Transportation Research Record*, 1894, 57–66.

Bifulco, G., Cartenì, A., et al. (2010) An activity-based approach for complex travel behaviour modelling. *European Transport Research Review*, 2, 209–221.

Bowman, J. L. (2009) Population synthesizers. *Traffic Engineering and Control*, 49(9), 342.

Bowman, J.L., Ben-Akiva, M.E. (2001) Activity-based disaggregate travel demand model system with activity schedules. *Transport Research Part A*, 35(1), 1–28.

Bowman, J.L., Bradley, M.A. (2008) Activity-based model: approaches used to achieve integration among trips and tours thoughtout the day, Paper presented at the European Transport Conference, Leeuwenhorst, The Netherlands, October 6-8.

Bowman, J.L., Bradley, M.A., Gibb, J. (2006) The Sacramento activity-based travel demand model: estimation and validation results, Paper presented at the European Transport Conference, Strasbourg, France, September 18-20.

Bradley, M., Bowman, J.L., Griesenbeck, B. (2010) SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution. *Journal of Choice Modeling*, 3 (1), 5-31.

Bricka, S., Bhat, C.R. (2006) A Comparative analysis of GPS-based and travel survey-based data. *Transportation Research Record*, 1972, 9–20.

Cascetta, E., Russo, F. (1997) Calibrating aggregate travel demand models with traffic counts: Estimators and statistical performance. *Transportation*, 24(3), 271–293.

Census (1992) Census of population and housing, 1990; summary tape file 3 on CD-ROM *Technical Documentation/prepared by the Bureau of the Census*. The Bureau, Washington.

Cools, M., Moons, E., Wets, G. (2010) Calibrating activity-based models with external origin–destination information: overview of possibilities. *Transportation Research Record*, 2175, 98–110.

Deming, W.E., Stephan, F.F. (1940) On the least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11(4), 427–444.

Guo, J.Y., Bhat, C.R. (2007) Population synthesis for microsimulating travel behavior. *Transportation Research Record*, 2014(12), 92–101.

Ireland, C.T. and Kullback, S. (1968) Contingency tables with given marginals. *Biometrika*, 55(1), 179–188.

MATSim-T (2010) Multi Agent Transportation Simulation Toolkit, http://www.Matsim.org. Accessed on 26/01/2013.

Mosteller, F. (1968) Association and estimation in contingency tables. *Journal of the American Statistical Association*, 63, 1–28.

Siripirote, T., Sumalee, A., Lam, W.H.K., Shao, H. (2012) Estimation of activity-based model parameters from travel diary survey: A case study of major city in Thailand. *Proceedings of the 17th International Conference of Hong Kong Society for Transportation Studies*, 427–434.

Siripirote, T., Sumalee, A., Watling, D.P., Shao, H. (2013) Updating of travel behavior model parameters and estimation of vehicle trip chain based on plate scanning. Paper accepted for publication in *Journal of Intelligent Transportation Systems*.

Stephan, F. (1942) An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13, 166–178.

Swiss federal statistical office (2000) Public use samples (PUS): Excerpts for general use from the Swiss federal population censuses 1970-2000, http://www.portal-stat.admin.ch/ pus/ files/index_e.html.

Thai national statistical office (2010), Census of population and housing, 2010, http:// www.nso.go.th. Accessed on 26/01/2013.

UrbanSim (2012) *Open Platform for Urban Simulation*, http://www.urbansim.org. Accessed on 26/01/2013.

Vovsha, P., Bradley, M., Bowman, J.L. (2004) Activity-based travel forecasting models in the United States: Progress since 1995 and Prospects for the Future, Paper presented at the EIRASS Conference on Progress in Activity-Based Analysis, Maastricht, The Netherlands, May 28-31.

Ye, X., Konduri, K., Pendyala, R.M., Sana, B., Waddell, P. (2009) A methodology to match distributions of both household and person attributes in the generation of synthetic populations, Paper presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C., January 11-15.

Yen, J.Y. (1971) Finding the K shortest loopless paths in a network. *Management Science*, 17, 712–716.