Sampling approach on spatial variation for travel demand forecasting

Riki Kawase, Junji Urata, Takamasa Iryo
Department of Civil Engineering, Kobe University, Japan

1 Introduction

Estimating an origin-destination (OD) travel demand pattern from a limited dataset is a traditional transportation problem. It is still vital these days even though a variety of observation measures that can directly measure each traveler's origin and destination using GPS or other techniques are available. It is mostly requested when we need to forecast future demand patterns. The traditional four-way step method is a typical example, where an OD demand pattern is estimated from its peripheral distribution (i.e. the number of generation and attraction trips). Even in the modern context, forecasting process of an origin-destination travel pattern should rely on how the numbers of attraction and generation trips evolves in the future if their changes are caused by changes of land use patterns. We especially consider to predict travel demand patterns a few days or weeks after a major disaster, in which damaged areas generate travel demand for recovery and transportation of relief goods. In such cases, we cannot rely on an OD demand pattern in a normal situation, which can be observed by GPS or other techniques. Instead, we need to estimate an amount of travel demand in each damaged area, and then distribute it to an OD table assuming a certain travel behavior model.

Estimating an OD demand pattern from a limited dataset surely incurs a variety of errors, and hence attempting a point estimation (i.e. estimating a single number for each cell of an OD demand matrix) is not appropriate. There are a number of possible error sources such as intra-personal heterogeneity, spatial variation, day-to-day variability, observation errors of transportation planners, and limitations in forecasting future demands. Zhao and Kockelman (2002)[1)] focused on uncertainty propagation in a contemporary transport demand model using point estimation. They showed that the variability in model inputs is highly correlated across outputs and the mis-predictions of travel demand in the early stages will surely amplify the mis-predictions of traffic flows.

This study presents a novel approach to obtain an interval estimation of an OD demand pattern instead of a point estimation. For point estimation, we should be able to use the generalized least squares estimator (GLS) approach using observed data (e.g. Zuylen and Willumsen (1980)[2)] and Cascetta (1984)[3)]) when we can use a variety of datasets e.g. so-called big-data. However, the GLS approach is hard to apply for future prediction with stochastic variations. Our approach estimates an OD demand pattern set, which includes a lot of likely OD demand patterns using data of trip generations and attractions. The set can include both the short-term and long-term future stochastic variability (Figure 1). The proposed destination choice model based on a discrete choice model evaluates spatial variations as random factors. Our algorithm randomly samples OD demand patterns using the Monte Carlo method. The computational cost to generate enough range of sampling is high; therefore, we implement a parallel computing algorithm to reduce the computation time.

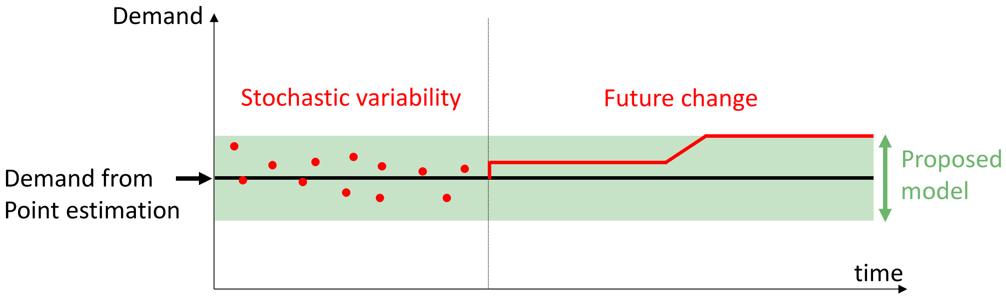

Figure 1 Stochastic variability and future changes in the proposed model

2 Proposed Approach

2.1 Formulation of Destination Choice Model

This subsection describes the formulation of the destination choice model with spatial variation. We assume that the spatial variation of the travel cost $c_{i\to j}$ is superior to other factors in the destination choice. This spatial variation is represented by normal distribution. Thus, the formulation that follows is similar to a mixed logit model. Standardly, a variation of a mixed logit model is derived from individual n, but the variation of our approach is derived from zones $i(=1,2,...,M)$ and $j(=$

$1, 2, ..., M$).

An observed utility from zone i to j is defined as:

$$V_{i \to j} = G_j - \log\left(c_{i \to j}\right), \tag{1}$$

where G_j is a measure of attraction for the destination j. A utility of the destination choice is defined as follows to introduce a spatial variation:

$$U^n_{i \to j} = V_{i \to j} + \eta_{i \to j} + \varepsilon^n_{i \to j}, \tag{2}$$

where $\varepsilon^n_{i \to j}$ is an IID Gunbel distribution, and $\eta_{i \to j}$ indicates a spatial variation defined as:

$$\eta_{i \to j} = -\nu_{i \to j} \log\left(c_{i \to j}\right). \tag{3}$$

$\nu_{i \to j}$ is normally distributed: $\nu_{i \to j} \sim N\left(0, \phi\right)$. We can obtain a destination choice probability $p_{i \to j}$ using a sampled $\eta_{i \to j}$. Although the spatial variation is normally distributed here, this condition can be changed.

This model assumes that variation of the destination choice mainly depends on travel cost. We hypothesize that a planner cannot know the randomness from each destination j, in addition to a random utility of each individual n. Moreover, this randomness is proportional to the travel cost because this variation will be correlated with zonal recognition, distance of OD pairs and future mode.

We can apply the network Generalised Extreme Value (GEV) model (Daly and Bierlaire (2006)[4)]), to capture the correlation between destinations. The Network GEV model allows complex correlation structures of destination choice situation.

2.2 Sampling Algorithm

Figure 2 shows a flowchart of our proposed sampling algorithm. First, the value of the travel cost $c_{i \to j}$, expected number of trips generated $E\left[O_i\right]$, and attracted $E\left[D_j\right]$ are given as exogenous parameters. We assume that planners can determine these parameters and the variances of O_i and D_j in future or unobserved situations. In addition, the number of exogenous parameters is considerably less than that in the GLS approach using a lot of the observed data. Using a distribution of O_i and D_j can also weaken the influence of these parameters of expected number on the output.

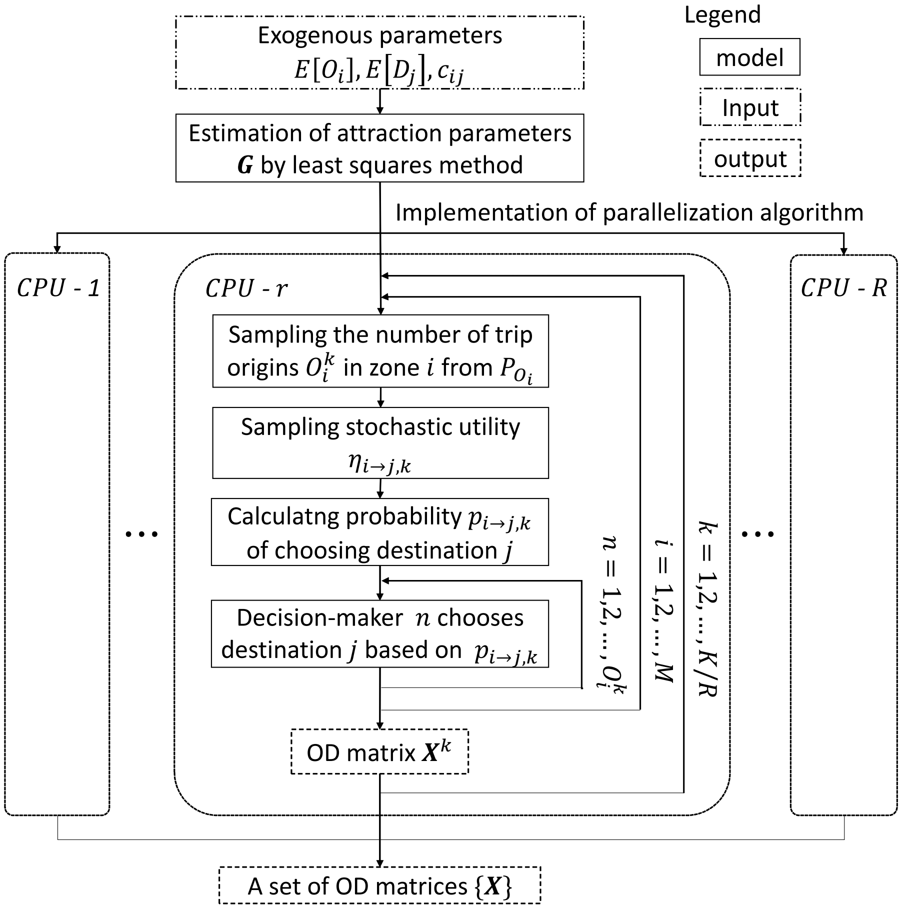

Figure 2 Sampling algorithm for a set of OD flow patterns

The measures of attraction $\boldsymbol{G}$ are calculated as follows using the exogenous parameters:

$$\min_{\boldsymbol{G}} RSS(\boldsymbol{G}) = \sum_j \left[E[D_j] - \sum_i \left[E[O_i] \cdot p_{i\to j} \right] \right]^2 . \tag{4}$$

We let $\phi = 0$, in Eq. (3), obtain $\boldsymbol{G}$ to calculate $p_{i\to j}$.

O_i^k, which is the k-th number of trips generated from zone i, is then sampled from a Poisson distribution P_{O_i}. Meanwhile, the destination choice probability $p_{i\to j,k}$ is calculated using a sampled $\eta_{i\to j,k}$. The destinations of O_i^k individuals are determined by $p_{i\to j,k}$. After the calculation for each origin i, we can then obtain an OD flow pattern $\boldsymbol{X}^k$. Finally, we can obtain a set of OD flow patterns $\{\boldsymbol{X}\}$ by iterating these sampling K times. This sampling can be parallelized because the generation of $\boldsymbol{X}^k$ is independent from $\boldsymbol{X}^h$.

3 Numerical Example

3.1 Settings

In this section, the Philadelphia Network[5)6)] is used to verify the reproducibility of the proposed model. The network is comprised of two states, the Pennsylvania and the New Jersey. The network consists of 13,389 nodes and 40,003 links. We divide the Philadelphia area into square grids to generate 588 zones in accordance with the traffic assignment problem (for convenience). The travel cost $c_{i \to j}$ is calculated by the user equilibrium assignment model (UE) using the Frank-Wolfe algorithm. The number of total trips is 13,338,626. This numerical example discards the intra-zonal trip because calculating the intra-zonal travel cost by UE is not possible.

Figure 3 shows the tree structure of the destination choice model. In this case, the upper nest regarding the travel cost and the lower nest regarding the state correspond the Generalized Nested Logit (GNL) model and Nested Logit (NL) model, respectively. The GEV generator function $G_i\left(Y_{i \to j}\right)$ can be written as follows:

$$G_i\left(Y_{i \to j}\right) = \sum_q \left(\sum_l \left(\alpha_i^{lq} \sum_{j \in J_l} \left(Y_{i \to j}\right)^{1/\mu_i^l} \right)^{\mu_i^l / \mu_i^q} \right)^{\mu_i^q} \tag{5}$$

$$Y_{i \to j} = \exp\left(V_{i \to j} + \eta_{i \to j}\right) \quad , \tag{6}$$

where J_l is a set of the destination j which is included in the NL nest l. The parameters μ_i^q and μ_i^l are respectively the nest-specific coefficients of the GNL and NL $\left(0 \leq \mu_i^q \leq \mu_i^l \leq 1\right)$. The parameter α_i^{lq} $\left(0 \leq \alpha_i^{lq} \leq 1\right)$ which allocates a proportion of

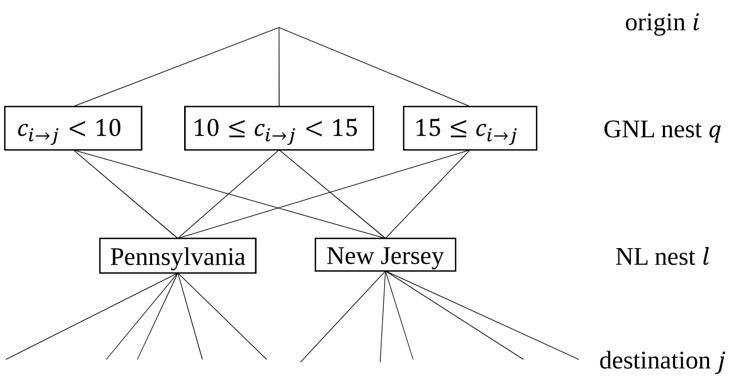

Figure 3 Tree structure for the destination choice

the NL nest l to the GNL nest q, is defined as:

$$\alpha_i^{lq} = \frac{|J_q \cap J_l|}{|J_l|}, \tag{7}$$

where J_q is a set of the destination j which is included in the GNL nest q. The parameters are set as $\mu_i^q = 0.2$, $\mu_i^l = 0.6$ and $\phi = 0.1$ in our calculation.

3.2 Results

We obtain 10,000 OD flow patterns using the proposed algorithm. Table 1 presents comparison between our sampling pattern and the real OD flow pattern. Table 1 presents that 93.5% of OD pairs is in the 95% confidence interval. A root mean square error (RMSE) shows that the volumes of the OD pairs with a low travel cost are extremely different from the volume of the real OD pairs.

Table 1 Comparison of the real OD flow pattern and the sampled OD flow patterns

	Percentage of the OD pairs in the 95% confidence interval of the smapled OD volume (%)	RMSE between the real OD volume and the median value of the sampled OD volume (vehicles)	the number of OD pairs (pairs)
All OD pairs	93.5	110.3	345,156
OD pairs with travel cost of 10 minutes or less	67.3	816.9	3,439

Figures 4 and 5 show 95% confidence intervals of the volume of each OD pair in our sampled set. The horizontal axis shows the IDs of the OD pairs arranged from the left in ascending order of the real OD volume. Figure 4 shows that the sampled OD volumes increase as the real OD volume increases. Figure 5, which is an enlarged view of Figure 4, depicts the same tendency. However, the range of OD pairs with a very large traffic volume is slightly smaller than that of the real OD volume. Most of its travel costs are less than 10 minutes, and the problem persists when reproducing the real OD flow pattern.

Finally, we show the calculation efficiency of our parallelized sampling algorithm. Figure 6 illustrates the speed-up ratios, which are based on the computation time of 125 CPUs. This speed-up ratio is very close to the ideal one, indicating that our algorithm is highly parallelized.

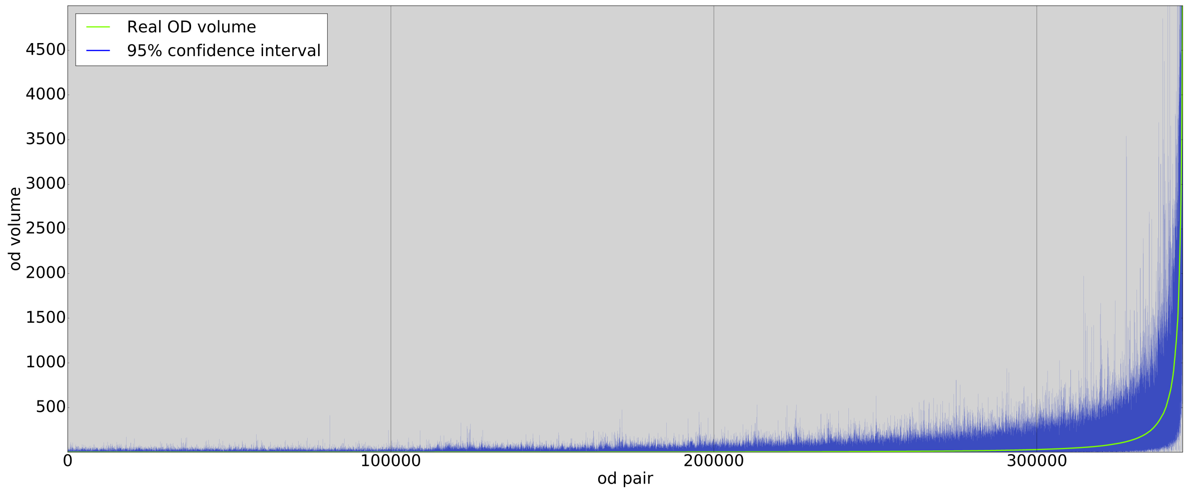

Figure 4   Confidence interval of OD volume (up to 5000 vehicles)

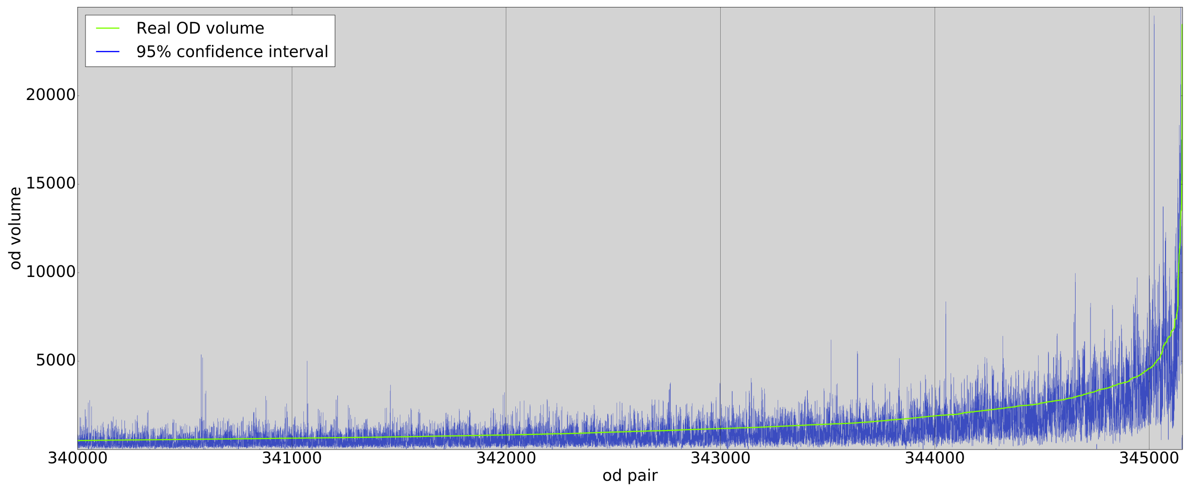

Figure 5   Confidence interval of the OD volume (enlarged view after the $340,000^{th}$ OD volume)

4 Conclusions

This proposed approach can present a set of OD flow patterns, including an uncertainty of demand forecasting. The limited dataset and stochastic variability are usually problems for travel demand forecasting. Our approach can obtain an interval estimation of an OD flow pattern instead of a point estimation. In consideration of uncertainty of observations, the interval estimation is superior to the point estimation. Our parallelized algorithm could decrease the large computational cost caused

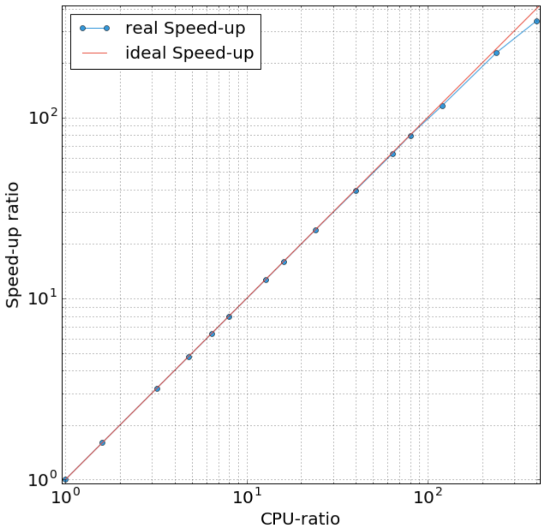

Figure 6 Speed-up ratio by the sampling algorithm (number of sample: 10,000,000)

by the large number of samples.

References

[1] Zhao, Y., Kockelman, K. M.: The propagation of uncertainty through travel demand models: An exploratory analysis, The Annals of Regional Science, Vol. 36, pp. 145-163, 2002.

[2] Zuylen, H.J.V., Willumsen, L.G.: The most likely trip matrix estimated from traffic counts, Transportation Research Part B, Vol.14 (3), pp.281-293, 1980.

[3] Cascetta, E.: Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator, Transportation Research Part B, Vol. 18 (4), pp.289-299, 1984.

[4] Daly, A., Bierlaire, M..: A general and operational representation of generalised extreme value models, Transportation Research Part B, Vol. 40, pp.285-305, 2006.

[5] Boyce, D., Ralevic-Dekic, B., Bar-Gera, H.: Convergence of traffic assignments: how much is enough?, Journal of Transportation Engineering, Vol. 130 (1), pp. 49-55, 2004.

[6] Transportation Network Test Problems: `http://www.bgu.ac.il/~bargera/tntp/`