Estimation of Link Choice Probabilities Using Monte Carlo Simulation and Maximum Likelihood Estimation Method

Mundher Ali Seger[1,2]*, Lajos Kisgyörgy[1]

[1] Department of Highway and Railway Engineering, Faculty of Civil Engineering, Budapest University of Technology and Economics, 1111 Budapest, Műegyetem rkp. 3. Hungary
[2] Civil Engineering Department, University of Technology, Sana'a Street, Baghdad, Iraq
* Corresponding author, e-mail: mundher.seger@gmail.com

Abstract

Studying the uncertainty of traffic flow takes significant importance for the transport planners because of the variation and fluctuation of temporal traffic flow on all links of the transport network. Uncertainty analysis of traffic flow requires identifying and characterizing two sets of parameters. The first set is the link choice set, which involves the Origin-Destination pairs using this link. The second set is the link choice probabilities set, which includes proportions of the travel demand for the Origin-Destination pairs in the link choice set. For this study, we developed a new methodology based on Monte Carlo simulation for link choice set and link choice probabilities in the context of route choice modeling. This methodology consists of two algorithms: In the first algorithm, we used the sensitivity analysis technique the variance-based method to identify the set of Origin-Destination pairs in each link. In the second algorithm, we used a Gaussian process based on the Maximum Likelihood framework to estimate the link choice probabilities. Furthermore, we applied the proposed methodology in a case study over multiple scenarios representing different traffic flow conditions. The results of this case study show high precision results with low errors' variances.

The key contributions of this paper: First, the link choice set can be detected by using sensitivity analysis. Second, the link choice probabilities can be determined by solving an optimization problem in the Maximum likelihood framework. Finally, the prediction errors' parameters of traffic assignment model can be modeled as a Gaussian process.

Keywords

traffic assignment, choice sets, link choice, parameter estimation, sensitivity analysis

1 Introduction

The fundamental premise in traffic assignment is choosing a route for each OD pair. Numerous factors influence a route choice; these include journey time, trip distance, monetary cost, congestion, type of roadway characteristics and perceptions of the drivers' themselves, etc. Containment of all these factors and assumption that all drivers have full knowledge about the transport system is questionable. Accordingly, the finding of route choice set for each traffic assignment model is a difficult task and not practical. Therefore approximations are inevitable in estimating the route choice set [1]. That approximation in the calculation of the route choice set caused uncertainty in the traffic assignment model on each link of the transport network especially in links with higher traffic density [2, 3].

Given the stochastic nature of the traffic flow on transport networks, where one of its causes due to route choice behavior by the drivers. Therefore, it is essential to estimate the choice set in the traffic assignment model to forecast driver's behavior under hypothetical scenarios, to predict future traffic conditions on transport networks and to understand the uncertainty in traffic assignment model. The importance of knowing the route choice set is increasingly recognized as a solution to uncertainty problems in traffic assignment model [4, 5].

Route choice sets are usually designated using choice set generation algorithms (e.g. shortest paths or labeling), which compute a set of routes based on properties of the transport network. These algorithms can carry two types of errors in the choice set generation. First, the false negative errors, which occur when the algorithm is not able to recreate the chosen alternatives. The created alternatives might not match the preferences and behavior of the drivers, and

as a result, the chosen route is not regenerate. The impact of this error decreases when the ability of choice set generation algorithm to capture the drivers' preferences and behavior increases. Second, false positive errors happen when a choice set generation algorithm also creates routes that are unconsidered by the divers, resulting in a too wide choice set. In result, using choice set generation algorithms potentially comes with several imperfections [6].

Accordingly, the evaluation of the significance and reality of generated route choice sets is difficult in practice because the actual choice sets are unknown to the transport modelers because it includes unlimited possible routes for each OD pair on the transport network. Several researchers, such as Ramming [7], Hoogendoorn-Lanser [8], Bekhor et al. [9], Hoogendoorn-Lanser and Nes [10], Bovy and Fiorenzo-Catalano [11] and Prato and Bekhor [12], have stated different methods to measure of quality and efficiency of the generated route choice sets. The empirical analysis shows that no choice set generation algorithm can fully reproduce observed paths.

Route choice sets are generally used to study for three application purposes: (i) Analysis of travel behavior, where the planner is interested to know of travel alternatives, characteristics, number, variety, composition, etc. (ii) Estimating parameters of choice models (i.e., estimating parameters of utility functions of route choice models at individual level). (iii) Predicting of choice probabilities for alternatives routes (i.e., determining route flow and link flow for travel demand analysis using route choice models) [11, 13].

This study focuses on the analysis of link flow and its uncertainty characteristics. We innovated a new methodology depending on MC simulation processes to estimate the route choice probabilities for all links on the transport network, and using the methodology's results to find the error's parameters of traffic flow in the traffic assignment model. The fundamental difference between this method and the rest of other methods; that all route choice models don not give adequate information for each link independently, contrary to this method which provides data about what are OD pairs used each link individually and their proportions. In this context, the new methodology based on both sensitivity analysis GA technique to find the generation set of OD pairs and maximum likelihood estimation MLE framework to estimate rote choice probabilities of these OD pairs in each link in the transport network.

SA is widely applied in scientific research where models are developed from engineering sciences, applied sciences to financial applications and risk analysis [14, 15].

The primary purpose of SA is to identify the contribution of the variability in inputs to the variability in the output, in purpose to obtain the most significant input parameters. Furthermore, the SA can identify variables where the reduction in uncertainty will have the maximum contribution to increase model accuracy [16]. SA is a robust way to quantitatively analysis the pattern of outputs behavior regarding disturbances or irregularity of model inputs and parameters [17]. Based on the SA results, the influential and the non-influential OD pairs are identified for each link on the network.

Parameter estimation plays a principal role in complex mathematical models, as it helps us to determine optimal parameters to create a reliable model [18]. Barnard and Bayes [19] described that for most mathematical models that have a large number of parameters, and the observed data was erroneous. Accordingly, parameter estimation can be performed in MLE framework, where the state estimate is the parameter which maximizes the likelihood function. The MLE framework gives estimates of the unknown quantities that maximize the probability of obtaining the observed set of data. MLE used in various scientific fields; such as communication systems, psychometrics, pharmacology, genetics, astrophysics, etc. [20]. This work uses MLE to estimate choice probabilities sets for links by using a linear Gaussian model.

This paper has three key contributions: First, we show that the link choice can be detected by using sensitivity analysis SA. Second, we show that the link choice probabilities can be determined by solving an optimization problem in the MLE framework. Finally, we show how the prediction errors' parameters of traffic assignment model can be modelled as a Gaussian process (GP).

The remainder of this paper is organized as follows: In Section 2, materials background about route choice modelling, and the literature of related topic used in this study which includes Sensitivity Analysis (SA), Maximum Likelihood Estimation (MLE) technique, and Linear Gaussian model technique. In Section 3, the proposed methodology is introduced. Finally, in Section 4, we present the results of the application of the proposed methodology on the case study in Ajka, Hungary.

2 Materials background
2.1 Route choice modelling
Choice set generation consists of finding all feasible routes that a traveller might consider for travelling from his origin to his destination. Choice set generation for route choice

modelling is known as a problematic problem compared to other choice modelling problems such as mode choice or destination choice [21].

It is well known that in a route choice context, finding route choice is complex for various reasons and involves several steps before the actual model estimation. The route choice model depends, on the one hand, on the attributes of the available routes, such as the type of road, speed limit, travel time, number of traffic lights etc. On the other hand, the characteristics and preferences of the divers also influence the choice. Some of the drivers like high speeds on freeways while others prefer small beautiful roads, some avoid left turns or traffic lights and so on. Several aspects of the route choice problem make it particularly complicated [22, 23].

Over the years various approaches have been suggested for modelling route choice behavior, this approaches can be subdivided into two categories deterministic and stochastic; Fig. 1 illustrates the situation [22].

In this chart, (U) represents the universal set that includes all possible routes for an OD pair. In the deterministic approach, (M) represents the master set of known routes generated by the researcher using deterministic or stochastic methods in order to approximate a driver's preferences and awareness set. And then, if the choice set generation method is deterministic, the probability of a driver choosing route (i) simply equals P($i|C$) where (C) is the individual final viable choice set, where ($C \subseteq M$). On the other hand, if the choice set generation method is

a probabilistic approach where all non-empty subsets (G) are considered, (G) is all the non-empty subsets of the master set (M). The probability of the probabilistic approach would be as follow (Eq. (1)):

$$P(i) = \sum_{C \in G} P(i|C)P(C).$$ (1)

Where $P(i)$ is the probability of a driver choosing route (i) from master set M, $P(i|C)$ is the probability of a driver is choosing route (i) from a given choice set (C), $P(C)$ probability of a driver's choice set being (C).

As noted above, the main purpose of route choice models is predicting probabilities of alternatives routes/links. However, generation of route choice sets suffers from several problems; First, it is unclear which criteria should be adopted to determine the appropriate alternatives, like the choice set, should have sufficient variety and include alternatives that are logically relevant for the study. Second, the route choice models usually consider trips between origin zones to destination zones, while estimation usually concentrates on individual trips. As a result, choice sets utilized for prediction should be much more significant to account for the differences in spatial and individual characteristics [6, 7, 13].

2.2 Sensitivity Analysis (SA)
SA permits the identification of the parameter or set of parameters that have the most significant influence on the model output. It gives useful insight into which of input variables contribute most to the variability of the model output [24, 25]. SA has been used in various scientific applications can be categorized as; understanding the input-output relationship, identifying the significant and influential parameters that rise model outputs and magnitudes, determining the uncertainty in setting parameters of the models structural that contribute to the overall variability in the model output, leading future experimental designs, etc. [15, 24, 26]. SA is classified into two types local and global methods. The first method is local sensitivity analysis (LSA) concentrates on the sensitivity of the output concerning small perturbations around the input values. In LSA, the values of the input parameters of interest are perturbed, while the other parameters are fixed on their nominal values. The second method is global sensitivity analysis (GSA); this method takes one or a blend of input parameters and studies the model output over the entire spaces of the input parameters [26].
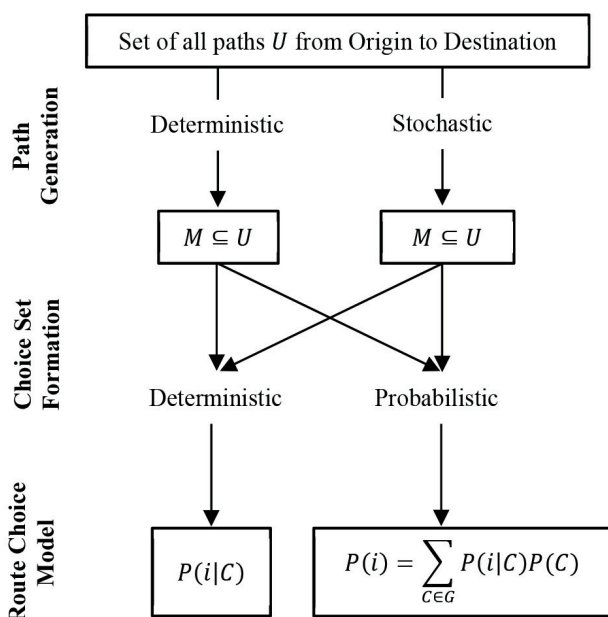

Fig. 1 Choice set generation [22]

GSA has been widely used in scientific researches and engineering design to get more information about complex model behaviour [27]. One of the most widely GSA methods used is variance-based sensitivity analysis, also known as analysis of variance (Sobol's method).

The Sobol method refers to ways of quantifying the contribution of each input parameters to the total variance of the output [28–30].

2.3 Maximum Likelihood Estimation (MLE)
MLE is one of a widely common technique used in statistical, engineering, economics, and many different applications to estimate the parameters of a model [31]. MLE can be implemented for most mathematical models regardless of the number of parameters in these models. The MLE framework gives values to the parameters that maximize the likelihood function [20].

Generally, the MLE method estimates a set of parameters that give as the simultaneous solution to the set of equations arising when the derivatives of the log likelihood concerning each of the parameters are set at zero. After that, the MLE technique involves obtaining all second derivatives of the likelihood for each parameter by employing any practical method of solution [20, 31, 32].

2.4 Gaussian Processes (GP)
The Gaussian processes GP have been commonly used in statistics and machine-learning studies for modelling stochastic processes in regression and classification [33]. The advantage of using the GP originates principally from two fundamental properties. First, a GP is wholly determined by its mean and covariance functions. This property promotes model fitting. Second, solving the prediction problem is almost straight- forward. The best predictor of a GP is a linear regression function of the observed values and, in many cases, these functions can be computed by applying recursive formulas [34, 35].

GP provides a technique for modelling probability distributions over functions based on the framework of Bayesian regression [33, 36]. Formally, GP generates data located throughout some domain such that any finite subset of the range follows a Normal multivariate distribution.

3 Methodology
In the literature of traffic assignment, all route choice models are looking for answers for the following two questions. The first question is what the possible routes for each OD pair are? And the second is what the probabilities

of these routes are? But in our new methodology, we start from a further two different questions. The first is what the OD pairs that pass by a link are? And the second is what the probabilities of these OD pairs are? The peculiarity of this methodology it is starting from the link(s) to find the other variables (i.e. OD pairs with their probabilities). Therefore, this methodology enables us to discover the characteristics of traffic flow variation of links in the case of fluctuation in travel demand.

The following methodology consists of four stages: Firstly, collecting data about travel demand of the study area. Secondly, the MC simulation process. Thirdly, sensitivity analysis to identify OD set in every link on the transport network. Finally, parameter estimation to find the probabilities of the OD set.

The relationships connecting these stages of the methodology are presented in Fig. 2. Also, all of the mathematical and logical computations of this methodology are illustrated in Appendix A.

3.1 Data
The required data for this stage consists of three components: the first is setting the transport network (TN) and traffic analysis zones (TAZ) in VISUM software. The second is defining the OD matrix. The last is specifying the sample size for the simulation process.

The notations used in this paper:
P: the total number of OD pairs,
p: an OD pair, $OD_{matrix} = \{OD_{p=1}, OD_{p=2},\ldots, OD_{p=P},\}$
L: the total number of links,
l: a link,
q_l: the traffic flow in a link (l),
Q: the traffic flow attributes, $Q = \{q_{l=1}, q_{l=2},\ldots, q_{l=L},\}$,
$\mathcal{F}_l$: the set of OD pairs for a link (l),
N: sample size.

3.2 MC simulation
In this paper, Quasi-Monte Carlo (QMC) simulation method has been used to generate data for the Sensitivity analysis stage and parameter estimation stage. QMC simulation usually used for problems that required to give a systematic sampling to get deterministic inputs instead of random ones. Hence, the results of the simulation process will be more stable and accurate [37].

In this paper, the Sobol sequence has been used to produce random data for the QMC simulation. This method generates a sequence of points in the unit hypercube $[0,1)^P$ where P is the dimension of the problem i.e., the number of
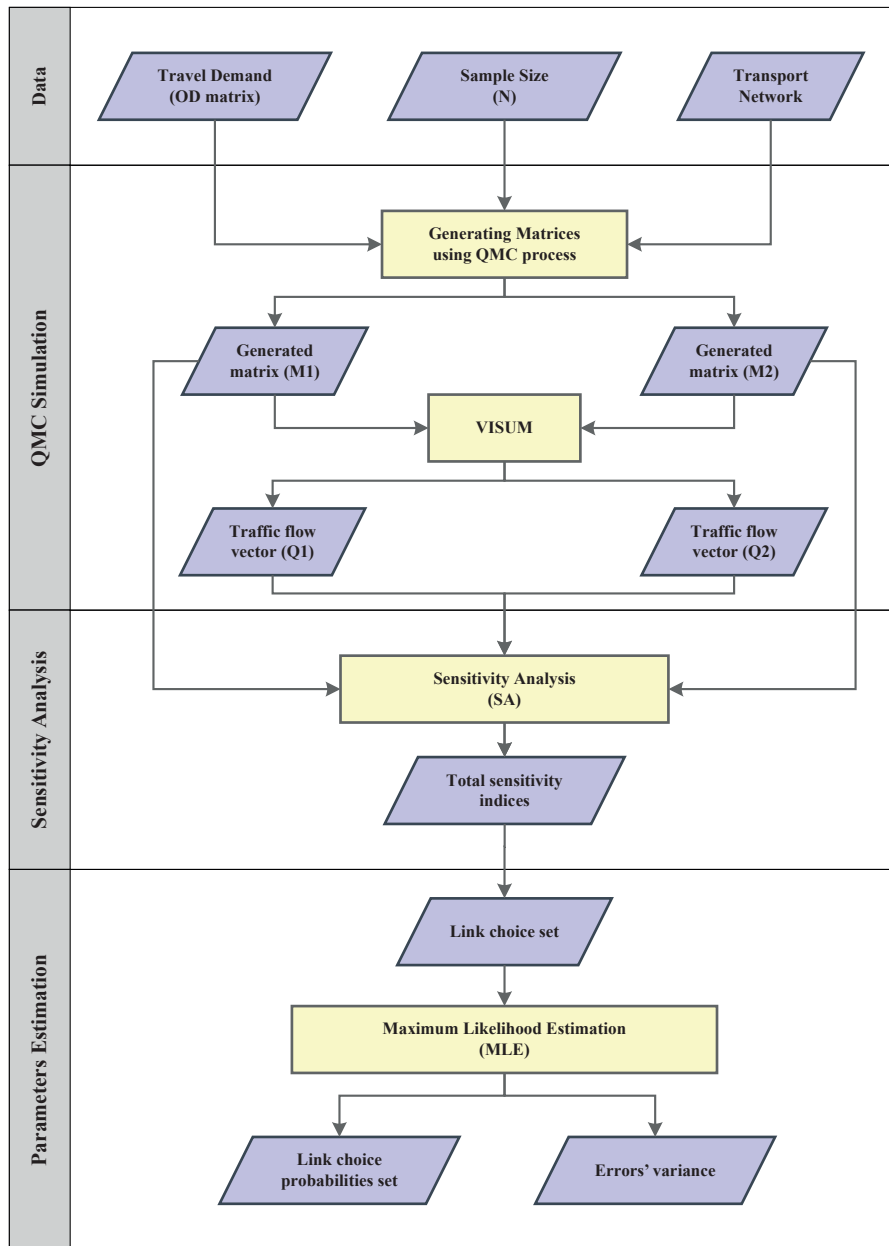
Fig. 2 Flowchart illustrates the methodology of estimation the link choice probabilities

OD pairs. In other words, each element inside the sequence is a P-dimensional vector whose components are fractions between 0 and 1.

For this methodology, a MATLAB code has been written and linked with the Component Object Model (COM) of VISUM software. Hence, by applying this code in this stage, two matrices have been generated (M_1) and (M_2) by employing the Sobol sequence, each one has a dimension (N, P). Then, the generated matrices have been executed in VISUM to produce two sets of outputs vectors: (Q_1) and (Q_2) corresponding to matrices (M_1) and (M_2), those vectors represent traffic flow in links.

3.3 Sensitivity indices (SI)

In general, SA is used to determine how "sensitive" a model is to changes in the value of the inputs corresponding to the changes in the value of the outputs. SA enable us to realize and understand the effect of the inputs on the behavior of the model outputs [26]. SA usually classified into two categories: Local Sensitivity Analysis (LSA) and Global Sensitivity Analysis (GSA). The LSA assesses changes in the model outputs regarding variations in a single input value only. While, in a GSA, all inputs values are changed simultaneously over the whole each individual input space, which allows assessing the relative

contributions of each input variable plus the interactions between input variables to the whole output variance of the model [15, 26, 38]. For this study we used GSA the method of variance-based sensitivity analysis, also known as (Sobol method), this method is based on the functional model decomposition, this method was introduced by the mathematician Sobol [28].

In this stage, Sobol method has been applied to identify the set of OD pairs that creating traffic flow in each link of the transport network. And to determine the proportion of influence for each OD pair on traffic flow variance. In general, each link has its own set of OD pairs affecting it. Moreover, the number of OD pairs in each set is different from a link to other, according to the location of the link in the transport network, the internal links have more OD pairs rather than the external links.

The Sobol method produces two types of indices: the first-order sensitivity index S_p and total sensitivity index ST_p. The first order index S_p is a measure for the variance contribution of the individual input to the total model variance. And, the total sensitivity index ST_p is a result of the main effect each individual input variable and all its interactions with the other model inputs variables.

Saltelli et al. [39] estimate the first order sensitivity index S_p and total effect index ST_p by QMC estimators.

$$S_p = \frac{\left[\left(\frac{1}{N}\right)\sum_{k=1}^{N}\left(q_k^{M_1}\right)^2 - f_0^2\right] - \left[\left(\frac{1}{2N}\right)\sum_{k=1}^{N}\left(q_k^{M_2} - q_k^{M_3}\right)^2\right]}{\left(\frac{1}{N}\right)\sum_{k=1}^{N}\left(q_k^{M_1}\right)^2 - f_0^2} \quad (2)$$

$$ST_p = \frac{\left(\frac{1}{2N}\right)\sum_{k=1}^{N}\left(q_k^{M_1} - q_k^{M_3}\right)^2}{\left(\frac{1}{N}\right)\sum_{k=1}^{N}\left(q_k^{M_1}\right)^2 - f_0^2} \quad (3)$$

Where $q_k^{M_1}$, $q_k^{M_2}$ and $q_k^{M_3}$ are model output in any iteration k corresponding to matrices M_1, M_2 and M_3 respectively, the M_3 is also a matrix formed from all the elements of the matrix M_2, except the p column (i.e. OD_p), where elements are taking from the matrix M_1.

f_0^2 is a constant calculated as follow (Eq. (4)):

$$f_0^2 = \left(\frac{1}{N}\sum_{k=1}^{N}q_k^{M_1}\right)^2. \quad (4)$$

Both of the first order sensitivity index S_p and total effect index ST_p are matrices with dimension (L, P). Generally, it possible to choose any one of these two

matrices to generate the link choice set $\mathcal{F}_l$. Consequently, the link choice set $\mathcal{F}_l$ consists of the OD pairs which have sensitivity index located within [0, 1) while the OD pairs that have zero sensitivity index will be out of the set.

3.4 Parameter estimation
In the last stage of this methodology, the route choice probabilities of each set $\mathcal{F}_l$ will be estimated using Gaussian Process GP based on the Maximum Likelihood ML framework.

The traffic flow in any link l caused by travel demand movement of a set of OD pairs and the probabilities set of these OD pairs that used this link l. Consider the Bayesian regression model the standard linear regression model with Gaussian noise; the traffic flow q_l for a link l:

$$q_l = \mathcal{F}_l\,\beta_l + \varepsilon_l. \quad (5)$$

Where: q_l is the traffic flow in the link l, $\mathcal{F}_l$ is the link choice set of OD pairs for the link l, β_l is the link choice probabilities set for the link l, ε_l is an error follows the normal distribution with a zero mean value and variance, σ^2_l hence $\varepsilon_l \sim N(0, \sigma^2_l)$. Thus, the normal density function for the ε_l:

$$\left(\varepsilon_l | o, \sigma^2_l\right) = \frac{1}{\sqrt{2\pi\sigma^2_l}} exp\left(-\frac{1}{2\sigma^2}\left(q_l - \mathcal{F}_l\beta_l\right)^2\right). \quad (6)$$

The likelihood function is a joint density of $\prod_{k=1}^{N}f\left(\varepsilon_l^k\right)$. Therefore, the likelihood function:

$$L\left(\beta_l, \sigma^2_l | q_l, \mathcal{F}_l\right) = \prod_{k=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2_l}} exp\left(-\frac{1}{2\sigma^2_l}\left(q_l^k - \mathcal{F}_l^k\beta_l\right)^2\right). \quad (7)$$

In MLE framework, we obtain the estimates of β_l and σ^2_l by maximizing the likelihood function.

First the logarithm of this function:

$$\ln L\left(\beta_l, \sigma^2_l\right) = -\frac{N}{2}\ln(2\pi) - N\ln\left(\sigma^2_l\right) \\ -\frac{1}{2\sigma^2_l}\left(q_l - \mathcal{F}_l\beta_l\right)^T\left(q_l - \mathcal{F}_l\beta_l\right). \quad (8)$$

Second derivatives the Eq. (8) with respect to β_l and σ^2_l parameters $\frac{\partial L}{\beta_l} = 0$ and $\frac{\partial L}{\partial \sigma^2_l} = 0$

The estimated probabilities set $\widehat{\beta_l}$:

$$\widehat{\beta_l} = \left(\mathcal{F}_l^T\mathcal{F}_l\right)^{-1}\mathcal{F}_l^T q_l. \quad (9)$$

The estimated variance $\widehat{\sigma^2_l}$:

$$\widehat{\sigma^2_l} = \frac{1}{N}\left(q_l - \mathcal{F}_l\widehat{\beta_l}\right)^T\left(q_l - \mathcal{F}_l\widehat{\beta_l}\right). \quad (10)$$

The predicted traffic flow $\widehat{q_l}$.

$$\widehat{q_l} = \mathcal{F}_l \widehat{\beta_l} . \tag{11}$$

The estimated error $\widehat{\varepsilon_l}$:

$$\widehat{\varepsilon_l} = q_l - \widehat{q_l} . \tag{12}$$

3.5 Synthetic example

To full understanding of the present methodology, a synthetic example for a small network consists of 4 zones and 22 links (see Fig. 3). The links' length of is proportional to the measurement in the figure and the movement is allowed in all directions at all nodes, and all the links are two-lane two-way. According to the methodology, estimating link choice probabilities for link #15 required the following stages.

- Firstly, for the data stage, a definition is required to OD matrix (see Table 1), sample size (assume $N = 1000$) and the transport network definition in VISUM software.
- Secondly, for the simulation stage, we applied the MATLAB code to generate input matrices sets and executing these matrices sets via VISUM software to produce traffic flow attributes sets.
- Thirdly, in the sensitivity analysis SA stage, the total sensitivity indices ST for OD pairs on link #15 has been presented in Table 2. The total sensitivity indices ST_p show the influence ratio of the OD pairs on traffic flow variance in link #15. Then, the link choice set $\{\mathcal{F}\}_{15}$ has been established by ditching from the OD pairs that have zero sensitivity indices (see Table 3). $\{\mathcal{F}\}_{15} = \{OD_{1-3}, OD_{2-3}, OD_{2-4}, OD_{3-1}, OD_{4-2},\}$
- Lastly, the link choice estimation stage, the link choice probabilities set $\{\beta\}_{15}$ has been estimated by applying Eq. (9), the result of this stage is presented in Table 3.
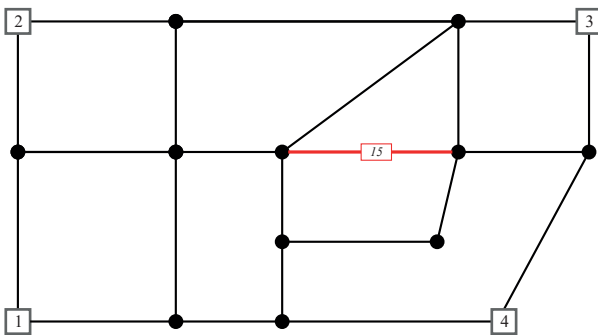

Fig. 3 Example network

Table 1 OD matrix (vph)

	D_1	D_2	D_3	D_4
O_1	-	250	400	250
O_2	200	-	250	200
O_3	150	200	-	150
O_4	300	250	200	-

Table 2 Total sensitivity indices of OD pairs for link # 15

	D_1	D_2	D_3	D_4
O_1	-	0.00	0.60	0.00
O_2	0.00	-	0.07	0.18
O_3	0.10	0.00	-	0.00
O_4	0.00	0.05	0.00	-

Table 3 The link choice set and link choice probabilities for link # 15

Link choice set $\{\mathcal{F}\}_{15}$	OD_{1-3}	OD_{2-3}	OD_{2-4}	OD_{3-1}	OD_{4-2}
Traffic demand (vph)	100	250	200	150	150
Link choice probabilities set $\{\beta\}_{15}$	0.38	0.04	0.20	0.13	0.03
Traffic flow per each OD pair (vph)	38	10	40	19.5	4.5
Predicted traffic flow $\hat{q}_{15}$ (vph)			112		

In this synthetic example, the traffic flow in link #15 produced by proportions $\{\beta\}_{15}$ of travel demand of a set of OD pairs $\{\mathcal{F}\}_{15}$. Practically, from the results in (Table 3), the estimated traffic flow $\hat{q}_{15}$ can be calculated by applying Eq. (11). Furthermore, if we have real traffic flow data, we can find the errors' parameters and validate the results by applying Eq. (12).

4 Case study

The new methodology has been implemented in a small city Ajka located in Hungary. Fig. 4 shows the transport network NT and traffic analysis zones TAZ of Ajka. The study area Ajka consists of 25 zones (22 internal zones and three external zones), 26 nodes and 28 two-directions links (i.e. 56 links).

In this paper, we suffice to show the analysis results for one link only; we select link #02 (l_{02}). This link summarizes the traffic flow pattern of other links in this study area. Moreover, we tested ten scenarios by changing the standard-deviation σ parameter of the MC simulation process to evaluate the effect of variation on the link choice sets $\{\mathcal{F}\}_{02}$ and link choice probabilities set $\{\beta\}_{02}$ for the selected link.

In Table 4 we can notice that values of the link choice probabilities of several OD pairs were decreased after increasing standard-deviation $\{\beta\}_{02}$ values, this means that

Table 4 Link choice set and link choice probabilities set of the link for various value of

link choice set $\{\mathcal{F}\}_{02}$	Link choice probabilities set $\{\beta\}_{02}$									
	$\sigma = 0.05$	$\sigma = 0.10$	$\sigma = 0.15$	$\sigma = 0.20$	$\sigma = 0.25$	$\sigma = 0.30$	$\sigma = 0.35$	$\sigma = 0.40$	$\sigma = 0.45$	$\sigma = 0.50$
OD_{04-01}	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.993	0.985
OD_{05-01}	1.000	1.000	1.000	1.000	1.000	0.993	0.976	0.988	0.959	1.000
OD_{06-01}	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.981
OD_{07-01}	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	1.000
OD_{11-01}	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.997	0.992	0.978
OD_{12-01}	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000
OD_{16-01}	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.996	1.000	1.000
OD_{04-02}	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000	1.000	1.000
OD_{05-02}	1.000	1.000	1.000	1.000	1.000	0.976	1.000	1.000	1.000	1.000
OD_{06-02}	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000
OD_{07-02}	1.000	1.000	1.000	1.000	1.000	1.000	0.978	1.000	1.000	0.964
OD_{11-02}	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.995	1.000	1.000
OD_{12-02}	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000
OD_{16-02}	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000
OD_{17-02}	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
OD_{04-03}	1.000	1.000	1.000	1.000	1.000	1.000	0.995	1.000	0.981	0.997
OD_{05-03}	1.000	1.000	1.000	1.000	1.000	0.996	0.988	1.000	1.000	1.000
OD_{11-03}	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.994	0.989	0.985
OD_{12-03}	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	1.000	0.987
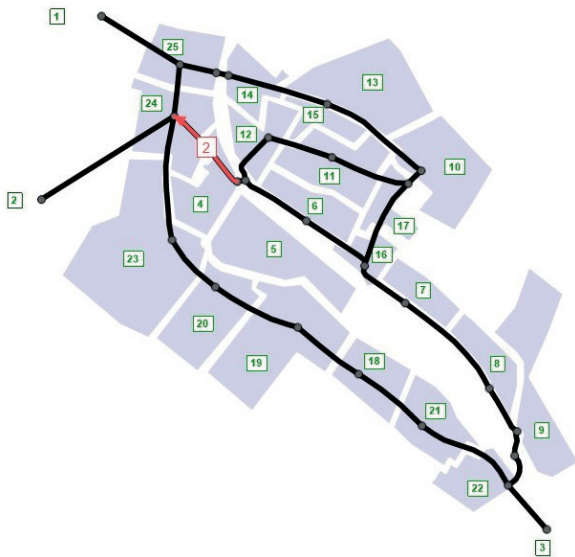
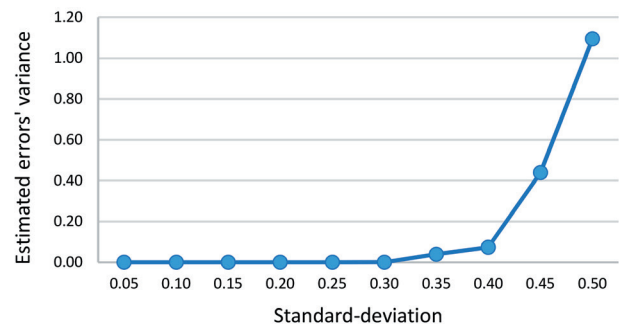**Fig. 4** Study area (Ajka, Hungary), showing transport network NT and traffic analysis zones TAZ

Fig. 5 Errors' variance in link # 02 for a sequential grow in standard-deviation

that the errors' variances $\widehat{\sigma^2}_{02}$ are near zero for σ equals from 0.05 to 0.30, after that there are increasing of errors' variances $\widehat{\sigma^2}_{02}$ to reach 1.094 at σ equal 0.50. In practical, all links were suffering from bias in errors' variances but in different values ranged between 0.5 to 12. Also, Fig. 6 shows the errors' distribution in link # 02 for a scenario of standard-deviation ($\sigma = 0.2$), we can see the estimated errors' $\widehat{\varepsilon_{02}}$ follow the Gaussian distribution.

5 Conclusions
This paper presents a new methodology for route choice modelling by adopting the links as a focal and initiate point for the purpose of finding a choice set and estimate choice probabilities. The reason for using links instead of
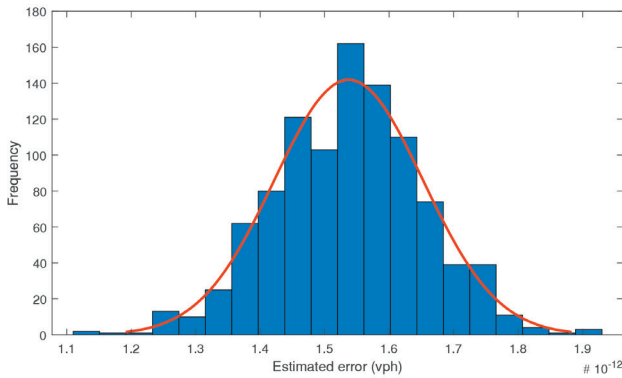
a part of the travel demand of these OD pairs moved from the link l_{02} to other links. In general, for this case study, the links choice sets $\{\mathcal{F}\}_{l=1.56}$ have contained ranged between 1 OD pair as a minimum to 34 OD pairs as a maximum.

Fig. 5 shows the estimated errors' variance $\widehat{\sigma^2}_{02}$ for various value of standard deviation σ, in this figure, we see

Fig. 6 Errors' distribution in link # 02 for a scenario $\sigma = 0.2$

routes is to enable us to discover the uncertainty about the traffic flow for each link on the transport network by comparing the predicted traffic flow obtained from this methodology and the observed traffic flow.

This paper has three key contributions:

- Sensitivity analysis was used to identify the link choice set.
- Maximum likelihood ML approach was applied to determine the link choice probabilities.
- The Gaussian process GP was applied to model the prediction errors of the methodology.

Our key findings provide a practical solution for traffic flow uncertainty analysis. With the suggested methodology the distribution of the traffic flow for each link can be known which enable us to handle the results of the traffic assignment model for different analysis scenarios like congestion, fuel, travel time. Our methodology also provides a tool for estimating the error of the modelling results, which is very useful for the decision makers.

We have proved the applicability of the suggested method using a synthetic test case and a real case. The synthetic case study shows that our methodology approaches the real values very precisely with low error variances. And the case study demonstrates the practical applicability of the proposed methodology.

Moreover, future research will be using this methodology of link choice to construct the route choice set and comparing the results with the different choice set generation methodologies.

Nomenclature

The notation used in this paper is listed below:

Acronyms

COM	Component object model
GP	Gaussian process
GSA	Global sensitivity analysis
LSA	Local sensitivity analysis
MC	Monte Carlo
MLE	Maximum likelihood estimation
OD	Origin destination
QMC	Quasi Monte Carlo
SA	Sensitivity analysis
SI	Sensitivity indices

Constants and indices

k	Iteration sample
f_0^2	Constant
l	Link
L	Number of links
N	Sample size
OD_{matrix}	Origin destination matrix
p	OD pair
P	Total number of OD pairs
S_p	First order sensitivity index
ST_p	Total sensitivity index

Variables and sets

i	Choosing route
C	Choice set
$P(i)$	Probability of a driver choosing route
$P(i\|C)$	Probability of a driver is choosing route (i) from a given choice set (C)
$P(C)$	Probability of a driver's choice set being (C)
U	Universal set that includes all possible routes for an OD pair
M	Master set of known routes
G	All the non-empty subsets of the master set (M).
ε_l	Error in (q_l)
$\hat{\varepsilon}_l$	Estimated error in (q_l)
σ_l^2	Errors' variance
$\widehat{\sigma_l^2}$	Estimated errors' variance
q_l	Traffic flow in a link (l)
$\hat{q}_l$	Estimated traffic flow in a link (l)
$\mathcal{F}_l$	Link choice set for the link (l)
β_l	Link choice probabilities set for the link (l)
$\hat{\beta}_l$	Estimated link choice probabilities set for the link (l)
M_1, M_2, M_3	Input matrices generated using QMC and OD_{matrix}
Q_1, Q_2, Q_3	Traffic flow attributes corresponding to input matrices M_1, M_2, M_3

References

[1] de Dios Ortúzar, J., Willumsen, L. G. "Modelling Transport", 4th ed., John Wiley & Sons, Chichester, UK, 2011.
https://doi.org/10.1002/9781119993308

[2] Avineri, E., Prashker, J. N. "Violations of Expected Utility Theory in Route-Choice Stated Preferences: Certainty Effect and Inflation of Small Probabilities", Transportation Research Record: Journal of the Transportation Research Board, 1894(1), pp. 222–229, 2004.
https://doi.org/10.3141/1894-23

[3] Prato, C. G. "Route choice modeling: Past, present and future research directions", Journal of Choice Modelling, 2(1), pp. 65–100, 2009.
https://doi.org/10.1016/S1755-5345(13)70005-8

[4] Sikka, N. "Understanding travelers' route choice behavior under uncertainty", PhD Thesis, University of Iowa, 2012.
https://doi.org/10.17077/etd.49ibdit5

[5] Telgen, M. G. "Realistic route choice modeling", Master's Thesis, University of Twente, 2010. [online] Available at: https://essay.utwente.nl/59718/1/MA_thesis_M_Telgen.pdf [Accessed: 10 May 2019]

[6] Ton, D., Duives, D., Cats, O., Hoogendoorn, S. "Evaluating a data-driven approach for choice set identification using GPS bicycle route choice data from Amsterdam", Travel Behaviour and Society, 13, pp. 105–117, 2018.
https://doi.org/10.1016/j.tbs.2018.07.001

[7] Ramming, M. S. "Network Knowledge and Route Choice", PhD Thesis, Massachusetts Institute of Technology, 2002. [online] Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.457.5449&rep=rep1&type=pdf [Accessed: 10 May 2019]

[8] Lanser, S. "Modelling Travel Behaviour in Multi-model Networks", PhD Thesis, Delft University of Technology, 2005. [online] Available at: http://resolver.tudelft.nl/uuid:3e013e1d-5bd6-40d6-b1d5-87fc53f3b3d9 [Accessed: 10 May 2019]

[9] Bekhor, S., Ben-Akiva, M. E., Ramming, M. S. "Evaluation of choice set generation algorithms for route choice models", Annals of Operations Research, 144(1), pp. 235–247, 2006.
https://doi.org/10.1007/s10479-006-0009-8

[10] Hoogendoorn-Lanser, S., van Nes, R. "On the Use of Choice Sets for Estimation and Prediction in Route Choice", Transport & Planning, Delft University of Technology, Amsterdam, Netherlands, 2006. [pdf] Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.564.659&rep=rep1&type=pdf [Accessed: 10 May 2019]

[11] Bovy, P. H. L., Fiorenzo-Catalano, S. "Stochastic route choice set generation: Behavioral and probabilistic foundations", Transportmetrica, 3(3), pp. 173–189, 2007.
https://doi.org/10.1080/18128600708685672

[12] Prato, C. G., Bekhor, S. "Applying Branch-and-Bound Technique to Route Choice Set Generation", Transportation Research Record: Journal of the Transportation Research Board, 1985(1), pp. 19–28, 2007.
https://doi.org/10.3141/1985-03

[13] Van Nes, R., Hoogendoorn-Lanser, S., Koppelman, F. S. "Using choice sets for estimation and prediction in route choice", Transportmetrica, 4(2), pp. 83–96, 2008.
https://doi.org/10.1080/18128600808685686

[14] Pannell, D. J. "Sensitivity analysis of normative economic models: theoretical framework and practical strategies", Agricultural Economics, 16(2), pp. 139–152, 1997.
https://doi.org/10.1016/S0169-5150(96)01217-0

[15] Borgonovo, E., Plischke, E. "Sensitivity analysis: A review of recent advances", European Journal of Operational Research, 248(3), pp. 869–887, 2016.
https://doi.org/10.1016/j.ejor.2015.06.032

[16] Cannavó, F. "Sensitivity analysis for volcanic source modeling quality assessment and model selection", Computers & Geosciences, 44, pp. 52–59, 2012.
https://doi.org/10.1016/j.cageo.2012.03.008

[17] de Jong, G., Daly, A., Pieters, M., Miller, S., Plasmeijer, R., Hofman, F. "Uncertainty in traffic forecasts: literature review and new results for The Netherlands", Transportation, 34(4), 375–395, 2007.
https://doi.org/10.1007/s11116-006-9110-8

[18] Trucano, T. G., Swiler, L. P., Igusa, T., Oberkampf, W. L., Pilch, M. "Calibration, validation, and sensitivity analysis: What's what", Reliability Engineering & System Safety, 91(10–11), pp. 1331–1357, 2006.
https://doi.org/10.1016/j.ress.2005.11.031

[19] Barnard, G. A. "Studies in the history of probability and statistics: ix. Thomas bayes's essay towards solving a problem in the doctrine of chances: Reproduced with the permission of the Council of the Royal Society from The Philosophical Transactions (1763), 53, 370-418", Biometrika, 45(3–4), pp. 293–295, 1958.
https://doi.org/10.1093/biomet/45.3-4.293

[20] Myung, I. J. "Tutorial on maximum likelihood estimation", Journal of Mathematical Psychology, 47(1), pp. 90–100, 2003.
https://doi.org/10.1016/S0022-2496(02)00028-7

[21] Bovy, P. H. L., Stern, E. "Route Choice: Wayfinding in Transport Networks", Springer, Dordrecht, Netherlands, 1990.
https://doi.org/10.1007/978-94-009-0633-4

[22] Frejinger, E., Bierlaire, M., Ben-Akiva, M. "Sampling of alternatives for route choice modeling", Transportation Research Part B: Methodological, 43(10), pp. 984–994. 2009.
https://doi.org/10.1016/j.trb.2009.03.001

[23] Fiorenzo-Calatano M. S. "Choice set generation in multi-modal transportation networks", PhD Thesis, Delft University of Technology, 2007. [online] Available at: http://resolver.tudelft.nl/uuid:ef3b9c22-b979-4f46-9b02-110c82d67535 [Accessed: 10 May 2019]

[24] Iooss, B., Lemaître, P. "A Review on Global Sensitivity Analysis Methods", In: Dellino, G., Meloni, C. (eds.) Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications, Springer, Boston, MA, USA, 2015, pp. 101–122.
https://doi.org/10.1007/978-1-4899-7547-8_5

[25] Sobol, I. M., Kucherenko, S. S. "Global sensitivity indices for nonlinear mathematical models. Review", Wilmott, 2005(1), pp. 56–61, 2005.
https://doi.org/10.1002/wilm.42820050114

[26] Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S. "Global Sensitivity Analysis. The Primer", 1st ed., John Wiley & Sons, Chichester, UK, 2007.
https://doi.org/10.1002/9780470725184

[27] Chen, W., Jin, R., Sudjianto, A. "Analytical Variance-Based Global Sensitivity Analysis in Simulation-Based Design Under Uncertainty", Journal of Mechanical Design, 127(5), pp. 875–886, 2005.
https://doi.org/10.1115/1.1904642

[28] Sobol, I. M. "Sensitivity analysis for nonlinear mathematical models", Mathematical Modelling and Computational Experiments, 1(4), pp. 407–414, 1993. [online] Available at: https://pdfs.semanticscholar.org/d339/b9cc42d6a7286d96814e6713fd13cdde87e7.pdf [Accessed: 10 May 2019]

[29] Bratley, P., Fox, B. L. "Algorithm 659: Implementing Sobol's quasirandom sequence generator", ACM Transactions on Mathematical Software, 14(1), pp. 88–100, 1988.
https://doi.org/10.1145/42288.214372

[30] Homma, T., Saltelli, A. "Importance measures in global sensitivity analysis of nonlinear models", Reliability Engineering & System Safety, 52(1), pp. 1–17, 1996.
https://doi.org/10.1016/0951-8320(96)00002-6

[31] Mai, A. T., Bastin, F., Toulouse, M. "On Optimization Algorithms for Maximum Likelihood Estimation", Interuniversity Research Center on Enterprise Networks, Logistics and Transportation, Quebec, Canada, CIRRELT-2014-64, 2014. [pdf] Available at: https://www.cirrelt.ca/DocumentsTravail/CIRRELT-2014-64.pdf [Accessed: 10 May 2019]

[32] Jain, V., Doshi, P., Banerjee, B. "Model-Free IRL using Maximum Likelihood Estimation", THINC Lab, Department of Computer Science University, University of Georgia, Athens, GA, USA, 2019. [pdf] Available at: http://thinc.cs.uga.edu/data/jdbAAAI19.pdf [Accessed: 10 May 2019]

[33] Seeger, M. "Gaussian Processes for Machine Learning", International Journal of Neural Systems, 14(02), pp. 69–106, 2004.
https://doi.org/10.1142/S0129065704001899

[34] Andrianakis, Y., Challenor, P. G. "Parameter estimation and prediction using Gaussian Processes", University of Southampton, Southampton, UK, MUCM Technical report 09/05, 2009. [pdf] Available at: http://www.mucm.ac.uk/Pages/Downloads/Technical%20Reports/09-05%20YA%203.2.3%20Parameter%20estimation%20and%20prediction%20using%20Gaussian.pdf [Accessed: 10 May 2019]

[35] Snelson, E. L. "Flexible and efficient Gaussian process models for machine learning", PhD Thesis, University of Cambridge, 2007. [online] Available at: http://www.gatsby.ucl.ac.uk/~snelson/thesis.pdf [Accessed: 10 May 2019]

[36] Do, C. B., Lee, H. "Gaussian processes", Stanford University, Stanford, CA, USA, pp. 1–14, 2007. [pdf] Available at: http://cs229.stanford.edu/section/cs229-gaussian_processes.pdf [Accessed: 05 December 2018]

[37] Jank, W. "Quasi-Monte Carlo sampling to improve the eeciency of Monte Carlo EM", Computational Statistics & Data Analysis, 48(4), pp. 685–701, 2005.
https://doi.org/10.1016/j.csda.2004.03.019

[38] Caswell H. "Introduction: Sensitivity Analysis – What and Why?", In: Sensitivity Analysis: Matrix Methods in Demography and Ecology, Springer, Cham, Switzerland, 2019, pp. 3–12.
https://doi.org/10.1007/978-3-030-10534-1_1

[39] Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S. "Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index", Computer Physics Communications, 181(2), pp. 259–270, 2010.
https://doi.org/10.1016/j.cpc.2009.09.018

Appendix (A)

Algorithm of parameter estimation of link choice probabilities.

Data stage:

$Z_{i=1,n}; \forall Z, i \in \mathbb{N}$ ▷ Zone definition.

$O_{i=1,n}; O \subset Z$ ▷ Origin definition.

$D_{j=1,n}; D \subset Z$ ▷ Destination definition.

$OD_{(i,j)}; \forall OD_{(i,j)} \in \mathbb{R}$ ▷ Matrix of Origin Destination (OD).

$OD_{(i,j)}^{obs.}; \forall OD_{(i,j)}^{obs.} \in \mathbb{R}$ ▷ Observed values of OD matrix.

$f_{TO \to D}$ ▷ Traffic Assignment function.

$L;$ ▷ Number of links.

$P = Z^2; P \in \mathbb{N}$ ▷ Number of OD pairs.

$N; N \in \mathbb{N}$ ▷ Number of QMC samples.

MC Simulation stage:

01: $\left\{ \varphi_k^{M_1} \right\}^{N \times P} \leftarrow QMC[\text{Sobol sequence}]$ ▷ Generating samples for the 1st of input matrices $\{M_1\}$.

02: $\left\{ \varphi_k^{M_2} \right\}^{N \times P} \leftarrow QMC[\text{Sobol sequence}]$ ▷ Generating samples for the 2nd of input matrices $\{M_2\}$.

03: **For** $k = 1$ to N **do**

04:     **For** $i = 1$ to n **do**

05: **For** $j = 1$ to n **do**

06:    $\left\{OD_{(i,j)}^k\right\}^{M_1} \leftarrow OD_{(i,j)}^{obs.} * \left\{\varphi_k^{M_1}\right\}$ $\triangleright$ Get the 1st set of input OD matrices $\{M_1\}$.

07:    $\left\{OD_{(i,j)}^k\right\}^{M_2} \leftarrow OD_{(i,j)}^{obs.} * \left\{\varphi_k^{M_2}\right\}$ $\triangleright$ Get the 2nd set of input OD matrix $\{M_2\}$.

08: **end for**

09: **end for**

10: **end for**

11: **For** $k = 1$ to N **do** $\triangleright$ Run VISUM.

12: $\left\{q_{l=1,L}^k\right\}^{M_1} \leftarrow f_{TO \to D}\left\{OD_{(i,j)}^k\right\}^{M_1}; \forall q_l^k \in \mathbb{R}$ $\triangleright$ Get the 1st set of traffic flow attributes $\{Q_1\}$.

13: $\left\{q_{l=1,L}^k\right\}^{M_2} \leftarrow f_{TO \to D}\left\{OD_{(i,j)}^k\right\}^{M_2}; \forall q_l^k \in \mathbb{R}$ $\triangleright$ Get the 2nd set of traffic flow attributes $\{Q_2\}$.

14: **end for**

Sensitivity indices stage:

15: **For** $l = 1$ to L **do** $\triangleright$ The first loop for links.

16: **For** $p = 1$ to P **do** $\triangleright$ The second loop for OD pairs.

17: **For** $k = 1$ to N **do** $\triangleright$ The third loop for samples.

18:   $\left\{OD_{(p=1:P,k)}^k\right\}^{V_1} \leftarrow reshape\left\{OD_{(i=1:n,j=1:n)}^k\right\}^{M_1}$ $\triangleright$ Reshape matrix (M_1) to vector (V_1).

19:   $\left\{OD_{(p=1:P)}^k\right\}^{V_2} \leftarrow reshape\left\{OD_{(i=1:n,j=1:n)}^k\right\}^{M_2}$ $\triangleright$ Reshape matrix (M_2) to vector (V_2).

20:   $\left\{OD_{(p=1:P)}^k\right\}^{V_3} \leftarrow \left\{OD_{(p=1:P)}^k\right\}^{V_2}$ $\triangleright$ Generate vector (V_3).

21:   $\left\{OD_{(p)}^k\right\}^{V_3} \leftarrow \left\{OD_{(p)}^k\right\}^{V_1}$ $\triangleright$ Replace the (p) column of (V_3) by the same column from (V_1).

22:   $\left\{OD_{(i=1:n,j=1:n)}^k\right\}^{M_3} \leftarrow reshape\left\{OD_{(p=1:P)}^k\right\}^{V_3}$ $\triangleright$ Reshape vector (V_3) to matrix (M_3).

23:   $\left\{q_{l=1,L}^k\right\}^{M_3} \leftarrow f_{TO \to D}\left\{OD_{(i,j)}^k\right\}^{M_3}; \forall q_l^k \in \mathbb{R}$ $\triangleright$ Run VISUM to get the 3rd set of traffic flow attributes (Q_3).

24: **end for**

25: $f_0^2 \leftarrow \left(\dfrac{1}{N}\sum_{k=1}^{N}\left\{q_l^k\right\}^{M_1}\right)^2$ $\triangleright$ Constant.

26: $S_{(l,p)} \leftarrow \dfrac{\left[\left(\dfrac{1}{N}\right)\sum_{k=1}^{N}\left(\left\{q_l^k\right\}^{M_1} - f_0^2\right)^2\right] - \left[\left(\dfrac{1}{2N}\right)\sum_{k=1}^{N}\left(\left\{q_l^k\right\}^{M_2} - \left\{q_l^k\right\}^{M_3}\right)^2\right]}{\left(\dfrac{1}{N}\right)\sum_{k=1}^{N}\left(\left\{q_l^k\right\}^{M_1}\right)^2 - f_0^2}$ $\triangleright$ First order sensitivity index.

27: $ST_{(l,p)} \leftarrow \dfrac{\left(\dfrac{1}{2N}\right)\sum_{k=1}^{N}\left(\left\{q_l^k\right\}^{M_1} - \left\{q_l^k\right\}^{M_3}\right)^2}{\left(\dfrac{1}{N}\right)\sum_{k=1}^{N}\left(\left\{q_l^k\right\}^{M_1}\right)^2 - f_0^2}$ $\triangleright$ Total sensitivity index.

28: **end for**

29: **end for**

Parameters Estimation stage:

30: **For** $l = 1$ to L **do** $\triangleright$ The first loop for links.

31: **For** $p = 1$ to P **do** $\triangleright$ The second loop for OD pairs.

32:      If $ST_{(l,p)} \neq 0$

33:        **For** $k = 1$ to N **do**   ▷ The third loop for samples.

34:        $\left\{\mathcal{F}_{(p,k)}\right\}^l = \left\{OD_{(p,k)}\right\}^{V_1}$ ▷ Link Choice set.

35:        **end for**

36:    **end for**

37: $\left\{\hat{\beta}_{(p)}\right\}^l = \left(\left(\left\{\mathcal{F}_{(p,k=1,N)}\right\}^l\right)^T \left\{\mathcal{F}_{(p,k=1,N)}\right\}^l\right)^{-1} \left(\left\{\mathcal{F}_{(p,k=1,N)}\right\}^l\right)^T \left\{q_{(l,k=1,N)}\right\}^{M_1}$ ▷ Link choice probabilities set.

38: $\widehat{\sigma^2}_l = \frac{1}{N}\left(\left\{q_{(l,k=1,N)}\right\}^{M_1} - \left\{\mathcal{F}_{(p,k=1,N)}\right\}^l \left\{\hat{\beta}_{(p)}\right\}^l\right)^T \left(\left\{q_{(l,k=1,N)}\right\}^{M_1} - \left\{\mathcal{F}_{(p,k=1,N)}\right\}^l \left\{\hat{\beta}_{(p)}\right\}^l\right)$ ▷ Errors' variance.

39: **end for**

40: **End**