Iranian Journal of Numerical Analysis and Optimization
Vol. 10, No. 2, (2020), pp 105–130
DOI:10.22067/ijnao.v10i2.79882

**Research Article**

# Maximum probability O-D matrix estimation in large-sized networks

M. Abareshi*

**Abstract**

We propose a maximum probability model to estimate the origin-destination trip matrix in the networks, where the observed traffic counts of links and the target origin-destination trip demands are independent discrete random variables with known probabilities. The problem is formulated by using the least squares approach in which the objective is to maximize the probability that the sum of squared errors between the estimated values and the observed (target) ones does not exceed a pre-specified threshold. An enumeration solution approach is proposed to solve the problem in small-sized networks, while a normal approximation based on the central limit theorem is applied in large-sized networks to transform the problem into a deterministic nonlinear fractional model. Some numerical examples are provided to illustrate the efficiency of the proposed method.

**AMS(2010):** Primary 90B06, Secondary 90B15, 90C06, 90C32.

**Keywords:** Transportation; Origin-destination trip matrix; Least squares approach; Probabilistic traffic counts; Fractional programming.

## 1 Introduction

Origin-destination $(O-D)$ trip matrix estimation problem is one of the most important issues in transportation networks, in which the number of trips between each $O-D$ pair is evaluated and inserted in a matrix called $O-D$ trip matrix. The traditional methods used for investigating the $O-D$ trip matrix, including such direct measurement methods as roadside interviews and flagging techniques, are so costly and mistakable. Conducting these methods to update the $O-D$ matrix information is not easily possible due

*Corresponding author
Received 3 April 2019; revised 19 February 2020; accepted 14 July 2020

Maryam Abareshi
Department of Applied Mathematics, Faculty of Mathematics and Computer Sciences, Hakim Sabzevari University, Sabzevar, Iran. e-mail: abareshi66@gmail.com

to limitations of budgeting and implementation problems. In recent years, using such available information as the observed traffic counts of links or a target $O-D$ matrix, obtained by a sample survey or an outdated model, the problem is formulated as a mathematical model that estimates the $O-D$ matrix in a cheaper and faster way.

Determining the $O-D$ matrix using the traffic counts of links can be interpreted as the inverse of the traffic assignment ($TA$) problem. In the $TA$ problem, the demands of $O-D$ pairs are given previously and assigned to the paths connecting them so that the total users' cost is minimized. As a result, the flows on paths and the traffic counts of links would be determined. In the $O-D$ matrix estimation problem, the purpose is to estimate the $O-D$ demands by using the observed traffic counts of links. The solution methods for this purpose are divided into two basic classes including (see [5]):

1. The proportional assignment models, in which the congestion effects are not considered and the only reasons to pick and choose a route are the traveler and route characteristics.

2. The equilibrium assignment models, in which the traffic congestion has considerable influences on choosing routes. In this case, the Wardrop's first principle will satisfy that the traffic network is in equilibrium when no traveler can achieve a lower travel cost by switching to another route; see [34].

Considering either the proportional or equilibrium assignment, there are four main mathematical models to estimate the $O-D$ matrix:

1. Entropy maximization model [33, 36, 18], in which by introducing a logarithm-based function, the most unbiased $O-D$ trip matrix consistent with the existing evidence is estimated.

2. Least squares approach [9, 11] with the purpose of minimizingthe sum of squared deviations between the estimated (true) variables and the observed (target) ones.

3. Maximum likelihood approach [29], which maximizes the likelihood of observing the target $O-D$ matrix and the observed link traffic counts conditional on the estimated (true) $O-D$ matrix.

4. Bayesian inference approach [22], in which Bayes' theorem is applied to combine information obtained from both the target $O-D$ matrix and the observed traffic counts of links to estimate the true $O-D$ matrix.

Abareshi, Zaferanieh, and Keramati [3] studied a maximum entropy ($ME$) path flow estimator for disaggregated flows between $O-D$ pairs with a pre-specified level for each disaggregation. Sun et al. [32] applied the $ME$ approach on a subnetwork $O-D$ matrix estimation model by using the total traffic generations (attractions) along with some fixed $O-D$ demands of

the subnetwork as the constraints. Applying a convex combination method, the resulted nonlinear problem was converted to the classical linear transportation problem where a tabular method was implemented to solve it. Xie, Kockelman, and Waller [37] proposed an elastic $O-D$ flow table estimation problem for subnetwork analysis. They proposed a combined maximum entropy-least squares estimator, by which the $O-D$ flows were distributed over the subnetwork in terms of the maximum entropy principle, while demand function parameters were estimated so that the least sum of squared errors was achieved.

Sun et al. [31] introduced two bi-level models to reconstruct $O-D$ demands under congested network using both the observed link and route travel times. Their proposed models aimed to minimize the distances between the observed and estimated traffic information ($O-D$, link and route travel times) in the upper level, and optimize the stochastic user equilibrium ($SUE$) model in the lower level, in which no driver can unilaterally change routes to improve his/her *perceived*, rather than actual, travel times.

In many real applications, the precise values of the observable data as link traffic counts or target $O-D$ trips in a network might be unavailable. For example, the observed number of travelers moving between cities may differ in various situations due to the weather conditions or coming the rush hours; consequently, some measurement errors may occur with known or unknown probability distributions. Therefore, it is desirable to treat the observed information as random or time-dependent variables with certain or uncertain parameters; see [13, 26]. In such cases, the use of statistical modeling to consider explicitly the presence of measurement and sampling errors in the observed information is developed.

Hoang, Vu, and Lo [16] formulated the information-based stochastic user equilibrium dynamic traffic assignment problem in a network, where the real-time information accounted for the (stochastic) uncertainty in demand and network capacity. Jones et al. [17] developed new methods for network assessment and control. Applying partial sensor and survey data while imposing equilibrium conditions during the data collection phase, they considered explicit account of demand variability and uncertainty.

Ma and Qian [20], using day-to-day traffic data collected over many years, estimated the mean and variance/covariance matrix of the $O-D$ demands. They [21] also presented a data-driven framework that estimates the day-to-day dynamic $O-D$ demands using high-granular traffic counts and speed data collected over many years. Pitomberia-Neto, Loureiro, and Carvalho [27] estimated the $O-D$ flows using link traffic volumes over a sequence of days and applied a dynamic hierarchical Bayesian model for estimating the day-to-day $O-D$ demands. Ching, Scholtes, and Zhang [13] assumed that the demands between nodes (zones) over a fixed period of time are formulated as independent random variables with unknown means. They considered both Poisson and normal distributions as the density functions

for the random travel demands and applied some numerical algorithms to estimate the optimal solution.

Shao et al. [28] proposed a bi-level optimization problem using a weighted least squares model in the upper level and a $TA$ model in the lower level to estimate the peak hour $O-D$ traffic demand variations from day-to-day hourly traffic counts throughout the whole year. Abareshi, Zaferanieh, and Safi [4] introduced a Markov chain $O-D$ matrix estimation problem in which the average time between two incoming streams to or outgoing streams from nodes in consecutive time periods was considered as a Markov chain. Besides, a normal distribution with pre-determined parameters in each period was used for traffic counts of links. They proposed a bi-level model where in the upper level, the network flow pattern with the maximum probability in the $n$th period was to be estimated, while a traffic assignment problem considering the equilibrium conditions was solved in the lower level.

Random variables in location problems have also received lots of attention, recently. Abareshi and Zaferanieh [1] considered prior probabilities to serve the demands of nodes by different facilities and introduced a bi-level model by applying the minimum information ($MI$) approach to determine the most probable allocation solution in the customer's point of view. Berman and Wang [7] studied locating $p$ medians to serve clients with discrete probabilistic demand weights, where the purpose was maximizing the probability that the total weighted distance does not exceed a given threshold value. To overcome the difficulty of evaluating the objective function, they [7] suggested to use a normal approximation of the problem based on the central limit theorem ($CLT$) when the size of the network is large enough. Berman and Wang [6] also considered four probabilistic network location problems with independent discrete demand weights and proposed efficient algorithms to solve the problem. Abareshi and Zaferanieh [2] studied the 1-median problem in the case, where the demand weights of nodes and the travel times of links were both discrete random variables.

In this paper, the observed traffic counts of links and the entries of the target $O-D$ matrix are considered to be independent random variables, where unlike [13], the corresponding probabilities are known. In this case, the squared errors between the estimated and the observed values are also independent random variables. Therefore, instead of minimizing the sum of squared errors, it is suggested to maximize the probability that this amount is less than or equal to a pre-threshold; see [6, 7]. Indeed, a predetermined bound is given for the estimated sum of squared errors. The purpose is to estimate the traffic flows on paths in an equilibrium approach to maximize the probability that the sum of squared errors between the estimated and observed (target) values does not exceed the given upper bound.

The problem is first examined for small-sized networks and a mixed-integer quadratically constrained ($MIQC$) model is proposed, which could be solved by using semidefinite programming ($SDP$) to reformulate the model as a quadratic convex problem; see [15]. We also propose an enumeration

method to solve the problem and compare the results with the ones obtained by $MIQC$ methods. Then, the $CLT$ is applied to the large-sized networks whereby the $MIQC$ model is transformed into a nonlinear fractional programming ($NLFP$) problem. The resulted problem would be solved by such appropriate methods as parametric algorithm or linearization approaches; see [30]. The main contributions of the paper are summarized below:

- A maximum probability approach is proposed to estimate the $O-D$ trip matrix in a probabilistic network in which the observed link traffic counts and the target $O-D$ matrix are discrete random variables with known probabilities
- The problem is formulated as an $MIQC$ model being solved via both $MIQC$ methods and an enumeration approach in small-sized networks, while a normal approximation based on the $CLT$ is applied to large-sized networks to transform the problem into a nonlinear fractional programming model.
- Three different cases are investigated by which the problem in large-sized networks is solved via the parametric algorithm, generating a sequence which superlinearly converges to the solution.
- Providing some numerical examples in both small and large-sized networks, the validity of the model as well as the solution approach is verified.

The rest of the paper is organized as follows: Problem formulation and the proposed $MIQC$ model as well as the enumeration approach are given in Section 2. In Section 3, the $CLT$ is applied to find the solution in large-sized networks, which results in an $NLFP$ problem. In Section 4, the solution approach for the resulted fractional model is stated. In Section 5, some numerical examples are provided to examine the added value of the proposed methods. Summary and conclusions are given in the last section.

## 2 Problem formulation

Let $G=(N,E)$ be a network with the node set $N$ and the link set $E$. The observed traffic counts of links as well as the entries of the target $O-D$ matrix are assumed to be independent random variables with known discrete probability distributions. Let $State=\{1,2,\ldots,H\}$ be the set of all realizations of the network. The vectors of the link traffic counts and target $O-D$ matrix in state $h\in State$ are, respectively, denoted by $\overline{V}^h$ and $\overline{X}^h$. Therefore, $\bar{v}_e^h$ and $\bar{x}_{rs}^h$ denote the observed traffic counts of link $e\in E$ and the target number of trips between the $O-D$ pair $(r,s)\in N\times N$ in the state $h$.

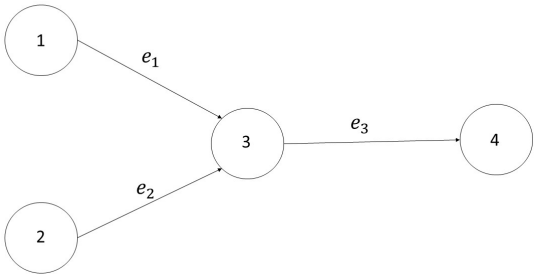



Figure 1: Small network

Note that if there is neither origin nor destination in the network, the traffic counts of links cannot be treated independently. For example, in Figure 1, the flow of link $e_3$ is equal to the sum of the traffic counts of links $e_1$ and $e_2$, that is, $v_{e_1}+v_{e_2} = v_{e_3}$. Then counting traffic on the link $e_3$ is redundant, while two counts $v_{e_1}$ and $v_{e_2}$ are independent. Moreover, if node 3 is considered as an origin or a destination, then flows in the network are independent values. Therefore, the expression $v_{e_1} + v_{e_2} = v_{e_3}$ is not satisfied; see [35].

Ching, Scholtes, and Zhang [13] considered the traffic volumes between different zones ($O - D$ pairs) over a fixed period of time as independent random variables with unknown means. In this paper, due to the assumption of existence of origin and destination nodes, both the link traffic counts and the target $O-D$ matrix components are assumed to be independent random variables. A threshold $D$ is given and the purpose is to estimate the $O - D$ matrix that maximizes the probability that the sum of squared errors between the estimated and the observed information is less than or equal to $D$.

The mathematical formulation of the maximum probability least squares ($MPLS$) problem to estimate the flow pattern in a probabilistic network is proposed as follows:

$$\max_f Z(f) = P_r\Big(\sum_{e\in E}(v_e - \bar{v}_e)^2 + \sum_{rs}(x_{rs} - \bar{x}_{rs})^2 \le D\Big), \tag{1}$$

$$s.t. \quad \sum_{rs}\sum_{k\in K_{rs}} f_{rs}^k \delta_{e,k} = v_e \qquad \text{for all } e \in E, \tag{2}$$

$$\sum_{k\in K_{rs}} f_{rs}^k \delta_{k,rs} = x_{rs} \qquad \text{for all } (r,s) \in N \times N, \tag{3}$$

$$f_{rs}^k \ge 0 \qquad \text{for all } k \in K_{rs}, (r,s), \tag{4}$$

where $\bar{v}_e$ and $\bar{x}_{rs}$ are, respectively, the probabilistic observed traffic count of link $e$ and target demand between $O - D$ pair $(r, s)$. Also, the decision

variables $v_e$ and $x_{rs}$ represent the corresponding estimated values. The set $K_{rs}$ includes the $K$-shortest paths connecting $O-D$ pair $(r,s)$, which is determined by Eppstein's $K$-shortest path ranking algorithm; see [14]. The variable $f_{rs}^k$ denotes the traffic flow on the path $k$ connecting $O-D$ pair $(r,s)$. The notation $\delta_{e,k}$ is the arc-path incident indicator, that is, $\delta_{e,k}=1$ if the link $e$ is a part of path $k$ and zero, otherwise. Similarly, $\delta_{k,rs}=1$ if the path $k$ is used to connect the $O-D$ pair $(r,s)$ and zero otherwise. Constraint (2) states that the estimated traffic count of link $e$ equals the sum of flows of all paths passing through this link. Besides, by constraint (3), the travel demand between each $O-D$ pair $(r,s)$ equals the sum of flows on all paths connecting them.

One of the fundamental methods to estimate the solution in probabilistic cases, is using the expected value model ($E$-model) that converts the probabilistic model into a deterministic problem by applying the expected values of the objective function or constraints; see [12]. Indeed in real applications, applying this approach may be useless and result in ineffective solutions; see the following example.

**Small Example.** Consider the network with two links, given in Figure 2. There are two paths including links $e_1$ and $e_2$ connecting $O-D$ pair $(1,2)$. Two realizations for the network with probabilities 0.7 and 0.3 are specified. The target $O-D$ matrix and the observed link traffic counts in the first and second realizations are assumed to be $\bar{x}_{1,2}^1=6.5,\ \bar{v}_{e_1}^1=3,\ \bar{v}_{e_2}^1=2$ and $\bar{x}_{1,2}^2=7.5,\ \bar{v}_{e_1}^2=4,\ ,\bar{v}_{e_2}^2=5$, respectively.

To estimate the optimal flow pattern with the least squared errors, either each of the distinct realizations or the expected values of the observed variables, instead of their random amounts, may be considered. The optimal flow pattern applying the least squares approach in the first sate is $v_{e_1}^*=3.5,\ v_{e_2}^*=2.5$ with the total sum of squared errors 0.75 in the first state and 8.75 in the second one. Also the optimal solution with the least squared errors in the second state is $v_{e_1}^{**}=3.5,\ v_{e_2}^{**}=4.5$ with the total sum of squared errors 6.75 in the first state and 0.75 in the second one.

For the first state optimal solution, $v_{e_1}^*=3.5,\ v_{e_2}^*=2.5$, the probability of the total sum of the squared errors being less than 1 is 0.7, while this value for the second state optimal solution, $v_{e_1}^{**}=3.5,\ v_{e_2}^{**}=4.5$, is 0.3. If the expected values of $\bar{x}_{1,2},\ \bar{v}_{e_1}$, and $\bar{v}_{e_2}$ are used, then the problem can be written as follows:

$$\min_f Z_{exp}(f)=\sum_e (v_e-\mu_e)^2+\sum_{rs}(x_{rs}-\mu_{rs})^2, \tag{5}$$

subject to constraints (2), (3), and (4). The notations $\mu_e$ and $\mu_{rs}$ are used for the expected values of random variables $\bar{v}_e$ and $\bar{x}_{rs}$, respectively. Let the

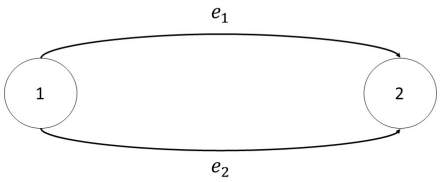



Figure 2: Network with two links

optimal flow vector of problem (5) be denoted by $f^*_{exp}$. Then, the total sum of squared errors in the first state is 1.47 and in the second one is 4.67. Hence the total sum of squared errors is larger than 1 in both cases.

If the purpose is to estimate a solution with the least sum of squared errors in both states simultaneously, then $f^*_{exp}$ is preferred to the solutions of the first and second states. But, if there is a pre-determined upper bound $D = 1$ for the total sum of squared errors, then the solution $v^*_{e_1} = 3.5,\ v^*_{e_2} = 2.5$ satisfies this condition with probability 0.7 and obviously is preferred to the other ones.

Let $Z_h(f) = \sum_e (v_e - \bar{v}^h_e)^2 + \sum_{rs}(x_{rs} - \bar{x}^h_{rs})^2$, for each state $h \in State$, where $v_e$ and $x_{rs}$ are defined by constraints (2) and (3), respectively. We define the characterization function $Y_h(f)$ as follows:

$$Y_h(f) = \begin{cases} 1 & Z_h(f) \le D, \\ 0 & otherwise. \end{cases}$$

Therefore problem (1) is rewritten as follows:

$$\max_f Z(f) = \sum_h Y_h(f) P^h_r, \tag{6}$$

where $P^h_r$ is the probability of state $h$. Let the optimal path flow vector for the following least squares problem, in state $h \in State$, be represented by $f^*_h$:

$$\min_f Z_h(f) = \sum_e (v_e - \bar{v}^h_e)^2 + \sum_{rs}(x_{rs} - \bar{x}^h_{rs})^2,$$

subject to constraints (2), (3), and (4). The flow vector $f^*_h$ can be estimated by using such solution methods for the least squares $O - D$ matrix estimation problem as Lagrangian dual approach or active-set constraints; see [10, 11, 24].

**Definition 1.** The set of indices $h \in State$ with $Z_h(f^*_h) \le D$ is defined as the candidate set and denoted by $C_{set} = \{h \in State;\ Z_h(f^*_h) \le D\}$.

If $h \notin C_{set}$, then $Y_h(f_h^*) = 0$, and consequently $Y_h(f) = 0$ for all vectors $f$. The definition of the characterization function $Y_h(f)$ for each vector $f$ results in the following constraint using the binary variable $y_h$, where $y_h$ is used instead of $Y_h(f)$ to simplify the model:

$$\sum_{e\in E}(v_e - \bar{v}_e^h)^2 + \sum_{rs}(x_{rs} - \bar{x}_{rs}^h)^2 - D \leq M(1 - y_h), \tag{7}$$

where $M$ is chosen sufficiently large. Inequality (7) together with the objective function $\max_f Z(f) = \sum_h y_h P_r^h$ is equivalent to the definition of $Y_h(f)$. Therefore, the following $MIQC$ problem is resulted:

$$\begin{aligned}
MIQC: \quad & \max_f Z(f) = \sum_h y_h P_r^h, \qquad (8)\\
s.t. \quad & \sum_{e\in E}(v_e - \bar{v}_e^h)^2 + \sum_{rs}(x_{rs} - \bar{x}_{rs}^h)^2 - D \leq M(1 - y_h),\\
& \sum_{rs}\sum_{k\in K_{rs}} f_{rs}^k \delta_{e,k} = v_e \qquad \text{for all } e \in E,\\
& \sum_{k\in K_{rs}} f_{rs}^k \delta_{k,rs} = x_{rs} \qquad \text{for all } (r,s) \in N \times N,\\
& f_{rs}^k \geq 0 \qquad \text{for all } k,(r,s), \qquad y_h \in \{0,1\} \quad \text{for all } h \in State.
\end{aligned}$$

There are some solution approaches to solve problem (8). Galli and Letchford [15] studied the possibility of extending the reformulation approach, using the $SDP$, proposed in [8] for equality-constrained $0-1$ quadratic programs, to the more general case of mixed-integer quadratically constrained quadratic ($MIQCQ$) problems. The proposed reformulation strengthened the continuous relaxation of the problem, while the optimal solution remained unchanged. Misener and Floudas [23] studied the global optimization of $MIQCQ$ problems by using a branch and bound algorithm applying a relaxation approach. They integrated the edge-concave relaxation with the piecewise-linear programming to tightly underestimate the $MIQCQ$ problems.

We propose an enumeration solution approach to solve problem (1) and compare the results with the ones obtained by solving $MIQC$ problem (8). First, it is required to find the vectors $f$ that minimize the maximum of the functions $Z_h(f)$, where indices $h$ belong to the subsets of $C_{set}$. Let $S_i \subseteq C_{set}$ be a subset with cardinality $i$, the optimal solution of the following problem should be determined:

$$\min_f \max_{h\in S_i} Z_h(f) = \sum_{e\in E}(v_e - \bar{v}_e^h)^2 + \sum_{rs}(x_{rs} - \bar{x}_{rs}^h)^2, \tag{9}$$

subject to constraints (2), (3), and (4). Problem (9) could be reformulated as a quadratically constrained ($QC$) problem:

$$
\begin{aligned}
QC: \quad & \min_{f} \quad \theta, && (10)\\
s.t. \quad & \sum_{rs} \sum_{k \in K_{rs}} f_{rs}^{k} \delta_{e,k} = v_e \qquad \text{for all } e \in E,\\
& \sum_{k \in K_{rs}} f_{rs}^{k} \delta_{k,rs} = x_{rs} \qquad \text{for all } (r,s) \in N \times N,\\
& \sum_{e \in E} (v_e - \bar{v}_e^h)^2 + \sum_{rs} (x_{rs} - \bar{x}_{rs}^h)^2 \leq \theta \qquad \text{for all } h \in S_i,\\
& f_{rs}^{k} \geq 0 \qquad \text{for all } k, (r,s).
\end{aligned}
$$

Problem (10) would be solved by using such nonlinear programming methods as Lagrangian dual approach or other quadratically constrained programming techniques; see [19, 25]. Let the optimal solution of problem (10) be $f_{S_i}^*$. If $Y_h(f_{S_i}^*) = 1$ for all $h \in S_i$, then $Z(f_{S_i}^*) = \sum_{h \in S_i} P_r^h$. By comparing the values of $Z(f_{S_i}^*)$ for all subsets $S_i \subseteq C_{set}$, the optimal solution of problem (1) is determined. Note that if a solution $f^*$ is obtained, then any subset $S_i$ that the sum of its elements' probabilities is less than $Z(f^*)$ can be ignored.

In Section 5, the solution of problem (1) is obtained for Yang's network by both solving the resulted $MIQC$ problem (8) and applying the proposed enumeration method using the $QC$ problem (10). In Section 3, a normal approximation is applied to investigate the problem in large-sized networks, which results in an $NLFP$ problem equivalent to problem (1).

## 3 Normal approximation

In this section, the $MPLS$ problem in large-sized networks, having an enormous number of links and $O-D$ pairs, is studied. Let the expected mean and variance of the probabilistic observed traffic count $\bar{v}_e$ be, respectively, denoted by $\mu_e$ and $\sigma_e^2$, and the expected mean and variance of $\bar{x}_{rs}$ be represented by $\mu_{rs}$ and $\sigma_{rs}^2$. Using Lemmas 1 and 2, first the mean and variance of term $\sum_{e \in E} (v_e - \bar{v}_e)^2 + \sum_{rs} (x_{rs} - \bar{x}_{rs})^2$ are determined. Then applying the $CLT$, problem (1) is reduced to an $NLFP$ problem.

**Lemma 1.** The mean of term $(v_e - \bar{v}_e)^2$ is $\hat{\mu}_e = (v_e - \mu_e)^2 + \sigma_e^2$, and its variance is $\hat{\sigma}_e^2 = 4v_e^2\sigma_e^2 + K_e$, where $K_e = var(\bar{v}_e^2)$.

*Proof.* Using the expansion of term $(v_e - \bar{v}_e)^2$ together with equality $\sigma_e^2 = E(\bar{v}_e^2) - \mu_e^2$ yields:

$$E(v_e - \bar{v}_e)^2 = E(v_e^2 - 2v_e\bar{v}_e + \bar{v}_e^2) = v_e^2 - 2v_e\mu_e + \sigma_e^2 + \mu_e^2 = \hat{\mu}_e$$

and

$$var(v_e - \bar{v}_e)^2 = 0 + 4v_e^2\sigma_e^2 + var(\bar{v}_e^2).$$

□

**Lemma 2.** The mean and variance of term $(x_{rs} - \bar{x}_{rs})^2$ are $\hat{\mu}_{rs} = (x_{rs} - \mu_{rs})^2 + \sigma_{rs}^2$ and $\hat{\sigma}_{rs}^2 = 4x_{rs}^2\sigma_{rs}^2 + K_{rs}$, respectively, where $K_{rs} = var(\bar{x}_{rs}^2)$.

The proof of Lemma 2 is similar to Lemma 1 and has been omitted. Applying the $CLT$ and noting the independence of the probabilistic variables $\bar{v}_e$s and $\bar{x}_{rs}$s, the term $W = \sum_{e\in E}(v_e - \bar{v}_e)^2 + \sum_{rs}(x_{rs} - \bar{x}_{rs})^2$ has a normal distribution with the expected mean $\mu_W = \sum_e \hat{\mu}_e + \sum_{rs}\hat{\mu}_{rs}$ and variance $\sigma_W^2 = \sum_e \hat{\sigma}_e^2 + \sum_{rs}\hat{\sigma}_{rs}^2$. Hence,

$$\frac{\sum_{e\in E}(v_e - \bar{v}_e)^2 + \sum_{rs}(x_{rs} - \bar{x}_{rs})^2 - \mu_W}{\sigma_W} \sim \text{normal}(0, 1).$$

Therefore,

$$\begin{aligned} &P_r\Big(\sum_{e\in E}(v_e - \bar{v}_e)^2 + \sum_{rs}(x_{rs} - \bar{x}_{rs})^2 \le D\Big) \\ &= P_r\Big(\frac{\sum_{e\in E}(v_e - \bar{v}_e)^2 + \sum_{rs}(x_{rs} - \bar{x}_{rs})^2 - \mu_W}{\sigma_W} \le \frac{D - \mu_W}{\sigma_W}\Big) \\ &= \phi\Big(\frac{D - \mu_W}{\sigma_W}\Big), \end{aligned}$$

where $\phi$ is the cumulative distribution function of the standard normal distribution. Since $\phi$ is an increasing function, problem (1) would be rewritten as the following $NLFP$ problem:

$$\max_f Z_{frac}(f) = \frac{D - \sum_e \hat{\mu}_e - \sum_{rs}\hat{\mu}_{rs}}{\sqrt{\sum_e \hat{\sigma}_e^2 + \sum_{rs}\hat{\sigma}_{rs}^2}}, \tag{11}$$

subject to constraints (2), (3), and (4), where variables $\hat{\mu}_e$, $\hat{\sigma}_e^2$, $\hat{\mu}_{rs}$ and $\hat{\sigma}_{rs}^2$ are the same as given in Lemmas 1 and 2. Problem (11) could be solved by using such nonlinear fractional programming methods as parametric algorithm or linearization approaches; see [30]. In the next section, a parametric method is employed to solve problem (11) in three cases.

## 4 Solution approach

Consider the following fractional programming problem:

**Algorithm 1** Parametric method

1: Take $\lambda = \lambda_1$ such that $F(\lambda_1) \geq 0$. (There is such a $\lambda_1$, since $F(0) \geq 0$ considering the assumption that $f(x) \geq 0$ for at least one $x \in S$.)
2: Solve problem (13). If $|F(\lambda)| \leq \epsilon$, stop. Otherwise, go to Step (3).
3: Set $\lambda = \dfrac{f(x^*)}{g(x^*)}$, where $x^*$ is the optimal solution of problem (13) obtained in Step (2). Go to Step (2).

$$P : \max_x \{q(x) = \frac{f(x)}{g(x)} | x \in S\}, \tag{12}$$

where $f$ and $g$ are real-valued differentiable functions on $S \subseteq R^n$. The parametric method is used to solve problem (12) in the case that $S \subseteq R^n$ is a compact and nonempty set and functions $f, g : S \to R$ are continuous with $g(x) > 0$ for all $x \in S$ and $f(x) \geq 0$ for at least one $x \in S$; see [30]. The equivalent parametric problem is introduced as follows:

$$Q(\lambda) : \max_x \{f(x) - \lambda g(x) | x \in S\}, \tag{13}$$

where $\lambda$ is a real-valued parameter. Let $F(\lambda)$ be the optimal value of the objective function in problem $Q(\lambda)$. The reader is referenced to [30] to see the proofs of the following lemmas and more details.

**Lemma 3.** Let $\bar{x}$ be the optimal solution to problem (12) and $\bar{\lambda} = \dfrac{f(\bar{x})}{g(\bar{x})}$. Then

1. $F(\lambda) < 0$ if and only if $\lambda > \bar{\lambda}$,
2. $F(\lambda) = 0$ if and only if $\lambda = \bar{\lambda}$,
3. $F(\lambda) > 0$ if and only if $\lambda < \bar{\lambda}$.

**Corollary 1.** If $F(\lambda) = 0$, then the optimal solution to problem (13) is also optimal to problem (12).

**Lemma 4.** The function $F(\lambda)$ is continuous, convex, and strictly decreasing on $R$.

Therefore, problem (12) is solved by determining the root of $F(\lambda) = 0$. It can be shown that the root is unique and nonnegative; see [30]. The methods used for solving the equation $F(\lambda) = 0$ will generate the solution algorithms for the $NLFP$ problem (12). The solution steps are outlined in Algorithm 1. The sequence generated by Algorithm 1 superlinearly converges to the root of $F(\lambda) = 0$; see [30]. Next, we give the solution method for problem (11) by introducing three subproblems under specified conditions.

1: **The First case:** $D - \sum_e \hat{\mu}_e - \sum_{rs} \hat{\mu}_{rs} > 0$**.** In this case, the numerator and denominator of the objective function in problem (11) are both continuous, differentiable, and positive functions. We use the parametric method described in Algorithm 1 to solve problem (11). In Step (2) of Algorithm 1, the following problem should be solved:

$$\max_f D' - \sum_e (v_e - \mu_e)^2 - \sum_{rs} (x_{rs} - \mu_{rs})^2 - \lambda \sqrt{\sum_e 4v_e^2\sigma_e^2 + \sum_{rs} 4x_{rs}^2\sigma_{rs}^2 + k''}, \tag{14}$$

subject to constraints (2), (3), and (4). In this problem, $D' = D - \sum_e \sigma_e^2 - \sum_{rs} \sigma_{rs}^2$ and $k'' = \sum_{rs} K_{rs} + \sum_e K_e$. Using such nonlinear constrained programming methods as penalty methods or augmented Lagrangian approaches, problem (14) could be solved; see [25]. However, if the first case is infeasible, the second one should be investigated.

2: **The Second case:** $D - \sum_e \hat{\mu}_e - \sum_{rs} \hat{\mu}_{rs} = 0$**.** In this case, the following least squares programming problem is considered:

$$\min_f \sum_e (v_e - \mu_e)^2 + \sum_{rs} (x_{rs} - \mu_{rs})^2, \tag{15}$$

subject to constraints (2), (3), and (4). Problem (15) could be solved by using such methods for least squares $O-D$ matrix estimation problems as Lagrangian dual approach or active-set constraints; see [10, 11, 24]. If the optimal value of problem (15) equals $D'$, then the obtained solution would be also optimal for problem (11) with the objective value equal to zero, else the third case should be considered.

3: **The third case:** $D - \sum_e \hat{\mu}_e - \sum_{rs} \hat{\mu}_{rs} < 0$**.** This case would be considered if neither the first nor the second case is feasible. Here, the parametric method is used with small modifications, to solve problem (12) in which $f(x)$ is negative while $g(x)$ is positive. Consider the problem:

$$Q'(\lambda) : \max_x \{f(x) + \lambda g(x) | x \in S\}. \tag{16}$$

Let $F'(\lambda)$ be the optimal value of problem (16). Then, Lemma 5, Corollary 2, and Lemma 6 are applied to find the optimal solution. For more detail see [30].

**Lemma 5.** Let $\bar{x}$ be an optimal solution to problem (12) in the case that the numerator is negative and $\bar{\lambda} = -\dfrac{f(\bar{x})}{g(\bar{x})}$. Then

(a) $F'(\lambda) < 0$ if and only if $\lambda < \bar{\lambda}$,

(b) $F'(\lambda) = 0$ if and only if $\lambda = \bar{\lambda}$,

(c) $F'(\lambda) > 0$ if and only if $\lambda > \bar{\lambda}$.

*Proof.* We only examine case (a), the other cases are similarly proved. Let $F'(\lambda) < 0$; then $f(x') + \lambda g(x') < 0$ for all $x' \in S$. Since $g(\cdot)$ is a positive function, then $-\dfrac{f(x')}{g(x')} > \lambda$ for all $x' \in S$. Therefore, $-\dfrac{f(\bar{x})}{g(\bar{x})} > \lambda$ and consequently $\bar{\lambda} > \lambda$.

Conversely, if $\bar{\lambda} > \lambda$, then $-\dfrac{f(\bar{x})}{g(\bar{x})} > \lambda$, and consequently $\dfrac{f(\bar{x})}{g(\bar{x})} < -\lambda$. Since $\bar{x}$ is the optimal solution to problem (12), then $\dfrac{f(x)}{g(x)} < -\lambda$ for all $x \in S$. Hence, $f(x) + \lambda g(x) < 0$ for all $x \in S$ and consequently $F'(\lambda) < 0$. □

**Corollary 2.** Let $F'(\lambda) = 0$; then the optimal solution to problem (16) is also optimal to problem (12).

**Lemma 6.** The function $F'(\lambda)$ is continuous, convex, and strictly increasing on $R$.

*Proof.* The continuity and convexity are proved similar to Lemma 4; see [30]. Let $\lambda_1 < \lambda_2$ and $x_1$ and $x_2$ be the optimal solutions to problems $Q'(\lambda_1)$ and $Q'(\lambda_2)$, respectively. Then $g(x) > 0$ yields

$$F'(\lambda_1) = f(x_1)+\lambda_1 g(x_1) < f(x_1)+\lambda_2 g(x_1) \leq f(x_2)+\lambda_2 g(x_2) = F'(\lambda_2).$$

□

In this case, Algorithm 1 is modified in Step (1) by choosing $\lambda$ such that $F'(\lambda) \leq 0$. In Step (2), the problem $Q'(\lambda)$ should be solved and in Step (3) the parameter $\lambda$ is updated by equality $\lambda = -\dfrac{f(x^*)}{g(x^*)}$.

Using Algorithm 1 with the mentioned modifications, problem (11) could be solved in the third case, where in the second step of Algorithm 1, the following problem should be solved:

$$\max_f D' - \sum_e (v_e - \mu_e)^2 - \sum_{rs} (x_{rs} - \mu_{rs})^2 + \lambda \sqrt{\sum_e 4v_e^2\sigma_e^2 + \sum_{rs} 4x_{rs}^2\sigma_{rs}^2 + k''},$$

subject to constraints (2), (3), and (4), by using nonlinear constrained programming methods such as penalty methods or augmented Lagrangian approaches; see [25].

**Efficiency of the method**

Applying the proposed methods in Sections 2 and 3 for small and large-sized networks, respectively, the nonlinear problem (1) with probabilistic variables in the objective function would be reduced to deterministic models, problems (10) and (11).

It should be noted that by applying the enumeration method proposed in Section 2, $2^{|C_{set}|}$ nonlinear constrained problems should be solved where $C_{set}$ is the candidate set introduced in Section 2. Therefore, in the worst case that $C_{set} = State$, $2^H$ $NP$-hard problems are to be solved, which cause the running time of the solution method to increase dramatically specially for large networks. Whereas, using the $CLT$ approach for medium and large-scale networks, it is required to solve just one deterministic fractional model, problem (11), instead of several nonlinear models, while simultaneously the number of constraints is also decreased.

In order to obtain the solution of problem (11), three different cases are considered in a certain order, where in the case of feasibility of each one, the remaining cases are neglected. The involved problems in cases 1 and 3 are readily solved by applying the parametric method, generating a sequence that superlinearly converges to the solution. The least squares problem (15) in the second case could be also solved by active-set constraints or Lagrangian approaches; see [24]. It is worth mentioning that, applying the $CLT$, an approximation of the optimal value of original problem (1) is provided, where the reliability of the solution would be investigated via the numerical examples in the next section.

## 5 Numerical examples

In this section, some numerical examples are provided to illustrate the efficiency of the proposed approaches discussed in the previous sections.

**Eexample 2.** In the first example, a small-sized network is studied, using the method stated in Section 2. Consider Yang's network given in Figure 3. It is assumed, there are five realizations for the observed information in the network with probability vector $Probability = \{0.2, 0.1, 0.2, 0.2, 0.3\}$. The necessary data, including the free flow travel times of links and the observed (target) flows in different states, are summarized in Tables 1 and 2. Ten equilibrium paths are picked up by Eppstein's $K$-shortest paths ranking algorithm [14], corresponding to four $O - D$ pairs $(1,3), (1,4), (2,3), (2,4)$ shown in Table 2.

The optimal path flow vectors $f^*$ obtained by solving $MIQC$ problem (8) and the proposed enumeration ($ENM$) method using problem (10), are given in Table 3, under column headings $MIQC$ and $ENM$, respectively. Since

$MATLAB$ solvers could not provide a feasible solution for $MIQC$ problem (8), the corresponding model was implemented in $CPLEX$ 12.6, while the $QC$ problem (10) was solved by $MATLAB$ $R2014a$. The optimal solutions with their corresponding objective values $Z(f^*)$ as well as the states that $Y_h(f^*) = 1$ for different values of $D$ are inserted in Table 3 (Note that the optimal solution is not unique).

As it is seen in Table 3, both methods $MIQC$ and $ENM$ reach the same values of the objective function for different amounts of threshold $D$. In addition, increasing the value of $D$ results in increasing the $MIQC$ and $ENM$ objective functions. In other words, when the threshold for the sum of squared errors rises, the total probability of states $h$ with $Y_h(f^*) = 1$ grows.

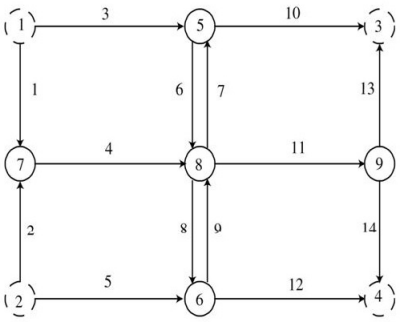



Figure 3: Yang's network

Next, to examine the stability of the resulted solutions in the case of occurring some errors in the observed information, we have made some modifications in the values of the observed link traffic counts and target $O-D$ demands. To this purpose, first, the values of $\bar{v}_e^h$ and $\bar{x}_{rs}^h$ for all links $e$ and $O-D$ pairs $(r, s)$ are multiplied by $\alpha = 1.5, 0.5$ in states $h = 1, 4$. The solutions for two values $D = 50000$ and $D = 95000$ are obtained by the $ENM$ algorithm and represented in Table 4. Then to investigate the effect of probabilities, the probability vector is changed to $Probability = [0.05\ 0.8\ 0.05\ 0.05\ 0.05]$ and the solutions are again obtained for $\alpha = 1.5, 0.5$.

As it is seen in Table 4, multiplying the values of $\bar{v}_e^2$ and $\bar{x}_{rs}^2$ by $\alpha = 1.5$ does not change the solution in both cases $D = 50000$ and 95000 compared with that of Table 3. While, when the values of $\bar{v}_e^4$ and $\bar{x}_{rs}^4$ are multiplied by 1.5, the solution changes for both $D = 50000$ and $D = 95000$. This is because, as it is perceived by Table 3, the value of the function $Z_h(f^*)$ corresponding to the optimal solution $f^*$ was less than $D = 50000$ for state $h = 4$; therefore, any change in the observed information in state $h = 4$ would affect the objective function. In other words, since in case $D = 50000$, the candidate set $C_{set}$ consists of just $h = 4$, multiplying the values $\bar{v}_e^4$ and

Table 1: The information of links in Yang's network

| Index of link | From | To | Free flow travel time | The observed traffic counts in states 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 5 | 13.18 | 130 | 100 | 120 | 150 | 100 |
| 2 | 1 | 7 | 4.29 | 100 | 210 | 200 | 180 | 180 |
| 3 | 2 | 6 | 11.49 | 180 | 80 | 80 | 150 | 130 |
| 4 | 2 | 7 | 4.46 | 130 | 50 | 80 | 130 | 100 |
| 5 | 5 | 3 | 13.24 | 200 | 80 | 120 | 100 | 90 |
| 6 | 5 | 8 | 3.00 | 30 | 90 | 90 | 90 | 100 |
| 7 | 6 | 4 | 12.16 | 200 | 150 | 100 | 110 | 130 |
| 8 | 6 | 8 | 5.00 | 10 | 180 | 180 | 150 | 100 |
| 9 | 7 | 8 | 11.89 | 230 | 60 | 60 | 100 | 120 |
| 10 | 8 | 5 | 4.00 | 30 | 330 | 330 | 130 | 200 |
| 11 | 8 | 6 | 4.00 | 50 | 85 | 120 | 140 | 150 |
| 12 | 8 | 9 | 13.05 | 210 | 150 | 150 | 130 | 130 |
| 13 | 9 | 3 | 4.19 | 115 | 140 | 140 | 100 | 100 |
| 14 | 9 | 4 | 3.11 | 90 | 200 | 200 | 100 | 140 |

$\bar{x}_{rs}^4$ by 1.5 for all links $e$ and $O-D$ pairs $(r,s)$, cause the candidate set to be empty and consequently, the optimal objective value would be equal to zero. While, the state $h=2$ did not even belong to the candidate set $C_{set}$ in Table 3. By multiplying its observed values by 1.5, the value of $Z_2(f_2^*)$ is still greater than $D$ which means that $2 \notin C_{set}$ for both $D=50000$ and 95000.

However, the case is different for $\alpha=0.5$. Multiplying $\alpha=0.5$ in $\bar{v}_e^2$ and $\bar{x}_{rs}^2$ brings the state $h=2$ to $C_{set}$ for both $D=50000, 95000$. Although, since the probability of state $h=2$ is less than the others, if it is necessary, the model might neglect it to find a solution with greater probability. When $D$ increases to 95000, the model finds a solution with $Z_h(f^*) \leq D$ for $h=2,3,4,5$.

Next, we have changed the used probability vector to $Probability = [0.05\ 0.8\ 0.05\ 0.05\ 0.05]$ to see the influence of probabilities on the optimal solution. Even, using the new vector of probabilities, the solutions corresponding to $\alpha=1.5$ are the same as previous, since the candidate set has not changed. But, for $\alpha=0.5$, where the candidate set includes the state $h=2$, the model attempts to find a solution with $Z_2(f^*) \leq D$. Note in the case that $\bar{v}_e^4$ and $\bar{x}_{rs}^4$ are multiplied by 0.5, the candidate set dose not include the second state for $D=50000$ and 95000.

Generally, changing the observed information of some states, might or might not result in changing the solution. Indeed, it depends on the state itself, whether it is in the candidate set or not, as well as its observed (target) values and probability.

**Example 3.** In this example, a large-sized network is investigated by using

Table 2: Target $O-D$ demands with the equilibrium shortest paths in Yang's network

| Index | Target demands in states 1 | 2 | 3 | 4 | 5 | From | To | Path | Path track |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 150 | 150 | 100 | 200 | 120 | 1 | 3 | 1 | $1 \to 5 \to 3$ |
| 2 | 120 | 200 | 120 | 150 | 130 | 1 | 4 | 2 | $1 \to 5 \to 8 \to 6 \to 4$ |
| | | | | | | | | 3 | $1 \to 5 \to 8 \to 9 \to 4$ |
| | | | | | | | | 4 | $1 \to 7 \to 8 \to 6 \to 4$ |
| | | | | | | | | 5 | $1 \to 7 \to 8 \to 9 \to 4$ |
| 3 | 820 | 140 | 180 | 90 | 90 | 2 | 3 | 6 | $2 \to 7 \to 8 \to 5 \to 3$ |
| | | | | | | | | 7 | $2 \to 7 \to 8 \to 9 \to 3$ |
| | | | | | | | | 8 | $2 \to 6 \to 8 \to 5 \to 3$ |
| | | | | | | | | 9 | $2 \to 6 \to 8 \to 9 \to 3$ |
| 4 | 300 | 80 | 80 | 100 | 170 | 2 | 4 | 10 | $2 \to 6 \to 4$ |

the normal approximation stated in Sections 3 and 4. We have considered *Anahiem* network shown in Figure 4 with 419 nodes and 914 links. Figure 4 and such necessary information as the free flow travel times and capacities of links are available in address $http://www.bgu.ac.il/\ bargera/tntp/$. Using Eppstein's $K$-shortest path ranking algorithm [14], the equilibrium cyclic free paths for 80, 90, 100, and 110 selected $O-D$ pairs are determined. We have considered different samples of sizes 200, 300, 400, and 500 for network realizations, taking random values with a discrete uniform distribution in the interval $[0, 8500]$, for the observed link traffic counts and target $O-D$ demands.

The proposed approximation method based on the $CLT$ is written in $MATLAB\ R2014a$ and applied to find the optimal solutions for different values of $D$ and different numbers of $O-D$ pairs, applying taken random samples. To investigate the acceptability of the obtained solutions, the estimated optimal solution vector $f^*$ obtained by problem (11) is given to problem (6) and $Z(f^*)$ is calculated. Then the values of $\phi(\frac{D-\mu_W}{\sigma_W})$ for $f^*$ and $Z(f^*)$ are compared. The absolute errors between $Z(f^*)$ and $\phi(\frac{D-\mu_W}{\sigma_W})$ for different sample sizes ($SS$) and different values of $D$ are shown in Table 5. The small absolute errors indicate the validity of the estimated solutions by the $CLT$ approach.

The optimal values of parameter $\lambda$ in Algorithm 1 obtained in one of the three cases discussed in Section 4, for 80 $O-D$ pairs, are inserted in Table 6. As the table indicates, the optimal value of $\lambda$ mostly belongs to the interval $[0, 2]$. In other words, the numerator in problem (11) is at most two times of the denominator. However, this may change for different problems.

Table 3: The estimated path flows with their objective functions values for different amounts of $D$

| | Optimal flows | | | | | |
|---|---|---|---|---|---|---|
| Paths | $D = 50000$ | | $D = 65000$ | | $D = 80000$ | |
| | $ENM$ | $MIQC$ | $ENM$ | $MIQC$ | $ENM$ | $MIQC$ |
| 1 | 111.81 | 95.16 | 67.55 | 77.74 | 67.55 | 79.89 |
| 2 | 60.24 | 51.94 | 50.87 | 49.75 | 50.87 | 50.94 |
| 3 | 35.46 | 35.93 | 32.25 | 44.50 | 32.25 | 49.44 |
| 4 | 53.88 | 42.78 | 61.62 | 47.1 | 61.62 | 47.93 |
| 5 | 20.52 | 31.51 | 31.18 | 42.38 | 31.18 | 46.67 |
| 6 | 11.95 | 26.74 | 34.43 | 41.22 | 34.43 | 46.23 |
| 7 | 23.59 | 29.30 | 17.89 | 38.60 | 17.89 | 44.49 |
| 8 | 46.60 | 40.54 | 49.35 | 45.67 | 49.35 | 46.77 |
| 9 | 53.66 | 46.83 | 31.80 | 41.75 | 31.80 | 44.38 |
| 10 | 54.67 | 61.11 | 81.92 | 49.42 | 81.92 | 85.49 |
| States | $\{4\}$ | $\{4\}$ | $\{4,5\}$ | $\{4,5\}$ | $\{4,5\}$ | $\{4,5\}$ |
| $Z(f^*)$ | 0.2000 | 0.2000 | 0.5000 | 0.5000 | 0.5000 | 0.5000 |
| Paths | $D = 95000$ | | $D = 11000$ | | $D = 125000$ | |
| | $ENM$ | $MIQC$ | $ENM$ | $MIQC$ | $ENM$ | $MIQC$ |
| 1 | 28.30 | 32.37 | 33.59 | 42.53 | 33.59 | 64.83 |
| 2 | 36.13 | 36.51 | 53.37 | 44.52 | 53.37 | 45.89 |
| 3 | 34.41 | 32.82 | 29.95 | 35.03 | 29.95 | 42.52 |
| 4 | 66.80 | 64.44 | 81.51 | 77.77 | 81.51 | 56.01 |
| 5 | 54.42 | 54.69 | 58.09 | 50.32 | 58.09 | 50.38 |
| 6 | 38.55 | 36.15 | 10.15 | 23.73 | 10.15 | 41.40 |
| 7 | 11.53 | 17.73 | 9.73 | 16.52 | 9.73 | 32.42 |
| 8 | 127.70 | 125.08 | 131.02 | 107.1 | 131.02 | 67.56 |
| 9 | 28.26 | 25.73 | 18.44 | 27.88 | 18.44 | 41.73 |
| 10 | 11.15 | 18.61 | 25.73 | 38.12 | 25.73 | 61.162 |
| States | $\{3,4,5\}$ | $\{3,4,5\}$ | $\{2,3,4,5\}$ | $\{2,3,4,5\}$ | $\{2,3,4,5\}$ | $\{2,3,4,5\}$ |
| $Z(f^*)$ | 0.7000 | 0.7000 | 0.8000 | 0.8000 | 0.8000 | 0.8000 |

Table 4: The changes in the solution under some modifications in the observed information

| | $\bar{v}_e^2 = 1.5\bar{v}_e^2$, | $\bar{x}_{rs}^2 = 1.5\bar{x}_{rs}^2$ | $\bar{v}_e^4 = 1.5\bar{v}_e^4$, | $\bar{x}_{rs}^4 = 1.5\bar{x}_{rs}^4$ |
|---|---|---|---|---|
| Paths | $D = 50000$ | $D = 95000$ | $D = 50000$ | $D = 95000$ |
| 1 | 111.81 | 28.30 | – | 28.30 |
| 2 | 60.24 | 36.14 | – | 35.50 |
| 3 | 35.46 | 36.41 | – | 35.05 |
| 4 | 53.88 | 66.81 | – | 67.45 |
| 5 | 20.53 | 54.42 | – | 53.78 |
| 6 | 11.95 | 38.55 | – | 37.00 |
| 7 | 23.60 | 11.53 | – | 13.08 |
| 8 | 46.61 | 127.71 | – | 129.26 |
| 9 | 53.67 | 28.27 | – | 26.72 |
| 10 | 54.67 | 11.16 | – | 11.16 |
| $Z(f^*)$ | 0.2 | 0.7 | 0 | 0.5 |
| States | $\{4\}$ | $\{3,4,5\}$ | $\{\}$ | $\{3,5\}$ |
| | $\bar{v}_e^2 = 0.5\bar{v}_e^2$, | $\bar{x}_{rs}^2 = 0.5\bar{x}_{rs}^2$ | $\bar{v}_e^4 = 0.5\bar{v}_e^4$, | $\bar{x}_{rs}^4 = 0.5\bar{x}_{rs}^4$ |
| Paths | $D = 50000$ | $D = 95000$ | $D = 50000$ | $D = 95000$ |
| 1 | 111.81 | 27.45 | 55.91 | 51.06 |
| 2 | 60.24 | 37.62 | 30.12 | 45.91 |
| 3 | 35.46 | 30.79 | 17.74 | 26.45 |
| 4 | 53.88 | 62.70 | 26.95 | 55.53 |
| 5 | 20.53 | 54.72 | 10.26 | 32.64 |
| 6 | 11.95 | 31.54 | 5.97 | 32.03 |
| 7 | 23.60 | 15.57 | 11.80 | 19.13 |
| 8 | 46.61 | 127.64 | 23.31 | 49.52 |
| 9 | 53.67 | 22.32 | 26.83 | 19.57 |
| 10 | 54.67 | 11.28 | 27.33 | 80.33 |
| $Z(f^*)$ | 0.2 | 0.8 | 0.2 | 0.5 |
| States | $\{4\}$ | $\{2,3,4,5\}$ | $\{4\}$ | $\{4,5\}$ |
| $Probability = [0.05\ 0.8\ 0.05\ 0.05\ 0.05]$ | | | | |
| | $\bar{v}_e^2 = 1.5\bar{v}_e^2$, | $\bar{x}_{rs}^2 = 1.5\bar{x}_{rs}^2$ | $\bar{v}_e^4 = 1.5\bar{v}_e^4$, | $\bar{x}_{rs}^4 = 1.5\bar{x}_{rs}^4$ |
| Paths | $D = 50000$ | $D = 95000$ | $D = 50000$ | $D = 95000$ |
| 1 | 111.81 | 28.30 | – | 28.30 |
| 2 | 60.24 | 36.14 | – | 35.50 |
| 3 | 35.46 | 36.41 | – | 35.05 |
| 4 | 53.88 | 66.81 | – | 67.45 |
| 5 | 20.53 | 54.42 | – | 53.78 |
| 6 | 11.95 | 38.55 | – | 37.00 |
| 7 | 23.60 | 11.53 | – | 13.08 |
| 8 | 46.61 | 127.71 | – | 129.26 |
| 9 | 53.67 | 28.27 | – | 26.72 |
| 10 | 54.67 | 11.16 | – | 11.16 |
| $Z(f^*)$ | 0.05 | 0.15 | 0 | 0.10 |
| States | $\{4\}$ | $\{3,4,5\}$ | $\{\}$ | $\{3,5\}$ |
| | $\bar{v}_e^2 = 0.5\bar{v}_e^2$, | $\bar{x}_{rs}^2 = 0.5\bar{x}_{rs}^2$ | $\bar{v}_e^4 = 0.5\bar{v}_e^4$, | $\bar{x}_{rs}^4 = 0.5\bar{x}_{rs}^4$ |
| Paths | $D = 50000$ | $D = 95000$ | $D = 50000$ | $D = 95000$ |
| 1 | 16.80 | 27.45 | 55.91 | 51.06 |
| 2 | 26.69 | 37.62 | 30.12 | 45.91 |
| 3 | 14.98 | 30.79 | 17.74 | 26.45 |
| 4 | 40.76 | 62.70 | 26.95 | 55.53 |
| 5 | 29.05 | 54.72 | 10.26 | 32.64 |
| 6 | 6.41 | 31.54 | 5.97 | 32.03 |
| 7 | 3.53 | 15.57 | 11.80 | 19.13 |
| 8 | 64.18 | 127.64 | 23.31 | 49.52 |
| 9 | 10.55 | 22.32 | 26.83 | 19.57 |
| 10 | 12.87 | 11.28 | 27.33 | 80.33 |
| $Z(f^*)$ | 0.8 | 0.95 | 0.05 | 0.10 |
| States | $\{2\}$ | $\{2,3,4,5\}$ | $\{4\}$ | $\{4,5\}$ |

Table 5: The absolute errors between $Z(f^*)$ and $\phi(\frac{D-\mu_W}{\sigma_W})$

| $O-D$ | 80 | | | | | 90 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SS\ $D$ ($10^{10}$) | 1.7 | 1.75 | 1.8 | 1.85 | 1.9 | 1.7 | 1.75 | 1.8 | 1.85 | 1.9 |
| 200 | 0.08 | 0.04 | 0.04 | 0.04 | 0.03 | 0.08 | 0.07 | 0.02 | 0.09 | 0.05 |
| 300 | 0.05 | 0.03 | 0.06 | 0.07 | 0.02 | 0.06 | 0.07 | 0.01 | 0.05 | 0.06 |
| 400 | 0.08 | 0.05 | 0.03 | 0.07 | 0.04 | 0.07 | 0.07 | 0.03 | 0.05 | 0.04 |
| 500 | 0.08 | 0.06 | 0.07 | 0.05 | 0.04 | 0.05 | 0.10 | 0.03 | 0.06 | 0.05 |
| $O-D$ | 100 | | | | | 110 | | | | |
| SS\ $D$ ($10^{10}$) | 1.7 | 1.75 | 1.8 | 1.85 | 1.9 | 1.7 | 1.75 | 1.8 | 1.85 | 1.9 |
| 200 | 0.06 | 0.09 | 0.05 | 0.06 | 0.08 | 0.03 | 0.08 | 0.07 | 0.06 | 0.08 |
| 300 | 0.07 | 0.06 | 0.00 | 0.10 | 0.07 | 0.04 | 0.08 | 0.06 | 0.04 | 0.08 |
| 400 | 0.03 | 0.09 | 0.03 | 0.06 | 0.07 | 0.03 | 0.07 | 0.04 | 0.03 | 0.06 |
| 500 | 0.04 | 0.08 | 0.03 | 0.06 | 0.08 | 0.03 | 0.07 | 0.05 | 0.01 | 0.08 |

The comparison between the average value of $Z(f^*)$ and $\phi(\frac{D-\mu_W}{\sigma_W})$, for different values of $D$ and different numbers of $O-D$ pairs, are given by Figures 5a–5d. As it is seen, the value of $\phi(\frac{D-\mu_W}{\sigma_W})$ is a reasonable approximation for $Z(f^*)$, the objective function of problem (6).

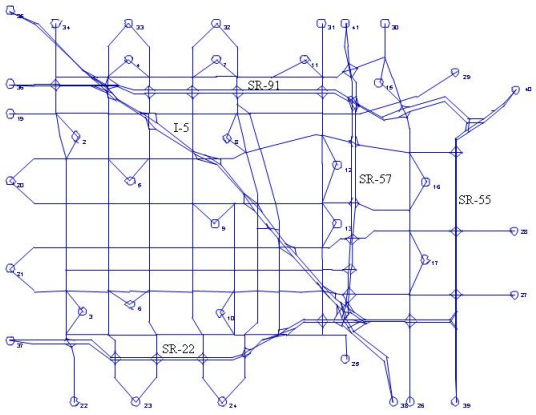



Figure 4: Anahiem network

**Example 4.** In the last example, in order to examine the efficiency of the $CLT$ method in comparison to other existing nonlinear solvers, we consider a 50-node network with 998 links and 56 $O-D$ pairs, where the network topology and necessary data are constructed randomly. Indeed, the existence

Table 6: The optimal values of parameter $\bar{\lambda}$ for 80 $O-D$ pairs

| $SS \setminus D$ | 1.7 ($10^{10}$) | 1.75 ($10^{10}$) | 1.8 ($10^{10}$) | 1.85 ($10^{10}$) | 1.9 ($10^{10}$) |
|---|---|---|---|---|---|
| 200 | 0.96 | 0.41 | 0.32 | 1.02 | 1.58 |
| 300 | 1.02 | 0.39 | 0.22 | 0.96 | 1.55 |
| 400 | 1.00 | 0.39 | 0.22 | 0.88 | 1.54 |
| 500 | 1.02 | 0.35 | 0.28 | 0.91 | 1.55 |

Table 7: The comparison between the $CLT$ method and $MIQC$ solvers

| $D$ | 1.9 ($10^8$) | | 1.8 ($10^8$) | | 1.7 ($10^8$) | |
|---|---|---|---|---|---|---|
| Solver | $RT$ $(sec)$ | $OBJF$ | $RT$ $(sec)$ | $OBJF$ | $RT$ $(sec)$ | $OBJF$ |
| $CLT$ | 54.13 | 1.00 | 46.71 | 0.97 | 43.47 | 0.50 |
| *BONMIN* | 11.75 | 0.98 | – | *Infeasible* | – | *Infeasible* |
| $BARON$ | – | *Infeasible* | – | *Infeasible* | – | *Infeasible* |
| $OQNLP$ | – | *Infeasible* | – | *Infeasible* | – | *Infeasible* |

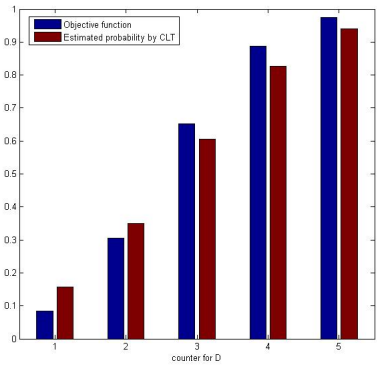
or nonexistence of links between nodes is determined by binary random variables, while the free flow travel times of links are randomly selected from the interval $[10, 30]$. Using Eppstein $K$-shortest paths ranking algorithm [14], 65 paths are picked up. In addition, 200 realizations are considered for the network where the values of the observed link traffic counts and target $O-D$ demands are taken from the interval $[0, 750]$ by the discrete uniform distribution.

Problem (1) is solved via both applying the proposed normal approximation based on the $CLT$ and employing $MIQC$ solvers including $BONMIN$, $BARON$, $OQNLP$ to problem (8). The corresponding running time ($RT$) and the objective function value ($OBJF$) for different values of $D$ are inserted in Table 7. The sign $-$ indicates that the solver could not find any feasible solution.

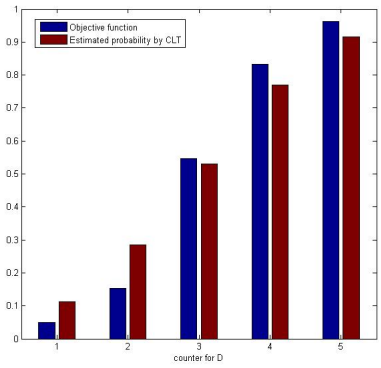
As it is seen by Table 7, in most cases, the $MIQC$ solvers did not find any feasible solution, while the $CLT$-based method was always able to find the optimal solution in an acceptable time. We have also examined such other existing solvers as $LINDOGLOBAL$, $COUNNE$, and $DICOPT$ on smaller and larger networks, where almost in all cases, the solvers did not provide any solution. Since the results were similar to Table 7, they have not been reported again.

## 6 Summary and conclusions

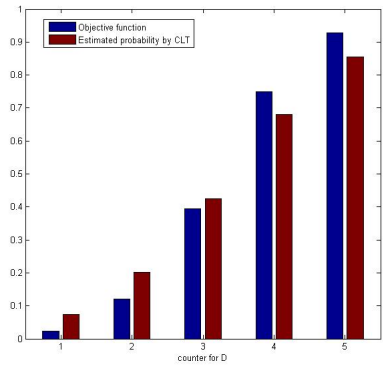
The available data in such transportation problems as the $O-D$ matrix estimation may be treated as nondeterministic information due to some external conditions. In this paper, we have considered different states for the



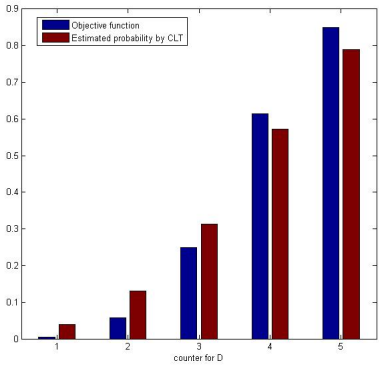

(a) Comparison for 80 $O-D$ pairs




(b) Comparison for 90 $O-D$ pairs




(c) Comparison for 100 $O-D$ pairs




(d) Comparison for 110 $O-D$ pairs

Figure 5: Comparison between $\phi(\frac{D-\mu_W}{\sigma_W})$ and $Z(f^*)$

observed link traffic counts and target $O-D$ matrix. The purpose was to estimate the solution that maximizes the probability of the sum of the squared errors being less than or equal to a pre-selected threshold. We investigated the problem in small and large-sized networks. An enumeration-based solution algorithm of the exponential time complexity was presented for small networks, where the validity of the solutions was verified via comparing with $MIQC$ solutions. As it was shown through examples, both the $MIQC$ solvers and enumeration algorithm provided equal objective function values in the optimality. Due to the large running time of the enumeration method, a normal approximation was proposed for large-sized networks. Using the $CLT$, the probabilistic problem in large-sized networks was transformed into one deterministic nonlinear fractional programming model. To reduce the complexity of the resulted model, three cases were considered to be solved in a

certain order where in the case of feasibility of each one, the next problems would be neglected. A parametric algorithm was implemented to solve the problems in the first and third cases, where the obtained sequence would superlinearly converge to the optimal solution. Providing some numerical examples, the approximated objective function by the $CLT$ approach was compared with its true value where considering the small differences between them, the efficiency of the $CLT$ approach was verified.

## References

1. Abareshi, M. and Zaferanieh, M. *A bi-level capacitated P-median facility location problem with the most likely allocation solution*, Transport. Res. B-Meth. 123 (2019), 1–20.

2. Abareshi, M. and Zaferanieh, M. *The network 1-median with discrete demand weights and travel times*, Iranian journal of numerical analysis and optimization, 9(1) (2019), 69–92.

3. Abareshi, M., Zaferanieh, M. and Keramati, B. *Path flow estimator in an entropy model using a nonlinear L-shaped algorithm*, Netw. Spat. Econ. 17 (2017), 293–315.

4. Abareshi, M., Zaferanieh, M. and Safi M. R. *Origin-destination matrix estimation problem in a Markov chain approach*, Netw. Spat. Econ. 19 (2019), 1069–1096.

5. Abrahamsson, T. *Estimation of origin-destination matrices using traffic counts - A literature survey*, IIASA Interim Report. IIASA, Laxenburg, Austria: IR-98-021 (1998).

6. Berman, O. and Wang, J. *Probabilistic location problems with discrete demand weights*, Networks, 44 (2004), 47–57.

7. Berman, O. and Wang, J. *The network p-median problem with discrete probabilistic demand weights*, Comput. Oper. Res. 37 (2010), 1455–1463.

8. Billionnet, A., Elloumi, S. and Plateau, M. C. *Improving the performance of standard solvers for quadratic 0-1 programs by a tight convex reformulation: the QCR method*, Discrete. Appl. Math. 157 (2009), 1185–1197.

9. Carey, M., Hendrickson, C. and Siddharthan, K. *A method for direct estimation of origin/destination trip matrices*, Transport. Sci. 15 (1981), 32–49.

10. Carey M. and Revelli, R. *Constrained estimation of direct demand functions and trip matrices*, Transport. Sci. 20 (1986), 143–152.

11. Cascetta, E. *Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator*, Transport. Res. B-Meth. 18 (1984), 289–299.

12. Charnes, A. and Cooper, W.W. *Deterministic equivalents for optimizing and satisficing under chance constraints*, Oper. Res. 11 (1963), 18–39.

13. Ching, W., Scholtes, S. and Zhang S. *Numerical algorithms for dynamic traffic demand estimation between zones in a network*, Eng. Optimiz. 36 (2004), 379–400.

14. Eppstein D. *Finding the K-shortest paths*, SIAM J. Comput. 28 (1998), 652–673.

15. Galli, L. and Letchford, A. N. *Reformulating mixed-integer quadratically constrained quadratic programs*, Working Paper. The Department of Management Science, Lancaster University, 2011.

16. Hoang, N.H., Vu, H.L. and Lo, H.K. *An informed user equilibrium dynamic traffic assignment problem in a multiple origin-destination stochastic network*, Transport. Res. B-Meth. 115 (2018), 207–230.

17. Jones, L.K., Gartner, N.H., Shubov, M., Stamatiadis, C. and Einstein, D. *Modeling origin-destination uncertainty using network sensor and survey data and new approaches to robust control*, Transport. Res. C-Emer. 94 (2018), 121–132.

18. Jornsten, K. and Nguyen, S. *On the estimation of a trip matrix from network data*, Technical report. LiTH-MAT-R-79-36, Department of Mathematics, University of Linkoping, Sweden, 1979.

19. Lu, C., Fang, S.C., Jin, Q., Wang, Z. and Xing, W. *KKT solution and conic relaxation for solving quadratically constrained quadratic programming problems*, SIAM J. Optim. 21 (2011), 1475–1490.

20. Ma, W. and Qian, Z. *Statistical inference of probabilistic origin-destination demand using day-to-day traffic data*, Transport. Res. C-Emer. 88 (2018), 227–256.

21. Ma, W. and Qian, Z. *Estimating multi-year 24/7 origin-destination demand using high-granular multi-source traffic data*, Transport. Res. C-Emer. 96 (2018), 96–121.

22. Maher, M.J. *Inferences on trip matrices from observations on link volumes: A Bayesian statistical approach*, Transport. Res. B-Meth. 17 (1983), 435–447.

23. Misener, R. and Floudas, C.A. *Global optimization of mixed-integer quadratically-constrained quadratic programs (MIQCQP) through piecewise-linear and edge-concave relaxations*, Math. Program. 136 (2012), 155–182.

24. Nie, Y., Zhang, H.M. and Recker, W.W. *Inferring origin-destination trip matrices with a decoupled GLS path flow estimator*, Transport. Res. B-Meth. 39 (2005), 497–518.

25. Nocedal, J. and Wright, S. *Numerical Optimization*, 2nd ed, Springer-Verlag New York, 2006.

26. Parry, K. and Hazelton, M.L. *Bayesian inference for day-to-day dynamic traffic models*, Transport. Res. B-Meth. 50 (2013), 104–115.

27. Pitombeira-Neto, A.R., Loureiro, C.F.G. and Carvalho, L.E. *A dynamic hierarchical Bayesian model for the estimation of day-to-day origin-destination flows in transportation networks*, Netw. Spat. Econ. 20 (2020), 499–527.

28. Shao, H., Lam, W.H.K., Sumalee, A., Chen, A. and Hazelton, M. L. *Estimation of mean and covariance of peak hour origin-destination demands from day-to-day traffic counts*, Transport. Res. B-Meth. 68 (2014), 52–75

29. Spiess, H. *A maximum likelihood model for estimating origin-destination matrices*, Transport. Res. B-Meth. 21 (1987), 395–412.

30. Stancu-Minasian, I.M. *Fractional programming, theory, methods and applications*, 1st ed, Springer Netherlands, 1997.

31. Sun, C., Chang, T., Luan, X., Tu, Q. and Tang, W. *Origin-destination demand reconstruction using observed travel time under congested network*, Netw. Spat. Econ. https://doi.org/10.1007/s11067-020-09496-4 (2020).

32. Sun, C., Chang, Y., Shi, Y., Cheng, L. and Ma, J. *Subnetwork origin-destination matrix estimation under travel demand constraints*, Netw. Spat. Econ. 19 (2019), 1123–1142.

33. Van Zuylen, H. and Willumsen, L.G. *The most likely trip matrix estimated from traffic counts*, Transport. Res. B-Meth. 14 (1980), 281–293.

34. Wardrop, J. *Some theoretical aspects of road traffic research*, Proceedings of the institution of civil engineers, (1952) 325–378.

35. Willumsen, L.G. *Estimation of an $O-D$ matrix from traffic counts- A review*, Working paper, Institute of Transport Studies, University of Leeds, Leeds, UK 1978.

36. Willumsen, L.G. *Simplified transport models based on traffic counts*, Transportation, 10 (1981), 257–278.

37. Xie, C., Kockelman, K.M. and Waller, S.T. *A maximum entropy-least squares estimator for elastic origin-destination trip matrix estimation*, Transport. Res. B-Meth. 45 (2011), 1465–1482.