Accurate OD Traffic Matrix Estimation Based on Resampling of Observed Flow Data

doi: info:doi/10.23919/APSIPA.2018.8659531

Accurate OD Traffic Matrix Estimation Based on Resampling of Observed Flow Data

Simon Kase*, Masato Tsuru† and Masato Uchida‡
* iD Corporation, Hokkaido, Japan
E-mail: s-kase@intelligent-design.co.jp
† Kyushu Institute of Technology, Fukuoka, Japan
E-mail: tsuru@cse.kyutech.ac.jp
‡ Waseda University, Tokyo, Japan
E-mail: m.uchida@waseda.jp

***Abstract*—It is important to observe the statistical characteristics of global flows, which are defined as series of packets between networks, for the management and operation of the Internet. However, because the Internet is a diverse and large-scale system organized by multiple distributed authorities, it is not practical (sometimes impossible) to directly measure the precise statistical characteristics of global flows. In this paper, we consider the problem of estimating the traffic rate of every unobservable global flow between corresponding origin-destination (OD) pair (hereafter referred to as "individual-flows") based on the measured data of aggregated traffic rates of individual flows (hereafter referred to as "aggregated-flows"), which can be easily measured at certain links (e.g., router interfaces) in a network. In order to solve the OD traffic matrix estimation problem, the prior method uses an inverse function mapping from the probability distributions of the traffic rate of aggregated-flows to those of individual-flows. However, because this inverse function method is executed recursively, the accuracy of estimation is heavily affected by the initial values of recursion and variation of the measurement data. In order to solve this issue and improve estimation accuracy, we propose a method based on a resampling of measurement data to obtain a set of solution candidates for OD traffic matrix estimation. The results of performance evaluations using a real traffic trace demonstrate that the proposed method achieves better estimation accuracy than the prior method.**

I. Introduction

Communication infrastructure has become the foundation for various socioeconomic activities and plays an important role as a lifeline for supporting the daily lives of people. Therefore, the influence of network failures on people has increased significantly. In order to avoid large-scale and long-term network failures, it is important to develop technology to accurately view the state of network utilization. However, the Internet is a diverse and large-scale system operated and managed by multiple distributed authorities, meaning it is not easy to directly measure the state of network utilization. Therefore, it is necessary to estimate the state of network utilization, which is difficult (sometimes impossible) to measure directly because of issues regarding the independence of network operation and management.

Network tomography is a technology addressing this issue [1]. In this paper, we focus on a network tomography method for estimating the unobservable traffic rate of every individual-flow between corresponding OD pairs (hereafter referred to as an OD traffic matrix or ODTM) based on a measurement of the aggregated traffic rate of individual-flows. The rate of aggregated-flow can be easily measured at some certain links (e.g., router interfaces) in a network. This is a more cost-effective method for network operation and management compared to the method of directly measuring the traffic rate of every individual-flow by investigating the source and destination IP addresses of individual packets passing through a router. Therefore, improving the accuracy of ODTM estimation is extremely important not only for network operation and management in normal situations, but also in emergency situations, such as the detection of network failures and identification of the causes of network failures. Although a hybrid approach was also proposed to utilize both direct monitoring of the traffic rates of individual-flows and indirect monitoring of those of the aggregated-flows [2], our present paper focuses on the approach that does not use the information about the traffic rates of individual-flows.

For ODTM estimation, the validity of the assumptions of the probabilistic model expressing the statistical state of individual-flow rates, which are difficult to directly measure, heavily affects estimation accuracy. For example, Cao et al. used a model assuming that each individual-flow rate follows an independent normal distribution [3]. Zhang et al. used a model that assumes a certain proportional relationship between aggregated-flow rate and individual-flow rate [4]. However, it is known that when these assumptions do not hold, estimation accuracy decreases considerably.

Tsuru et al. proposed a ODTM estimation method called the inverse function method by using a discrete probabilistic model with high degree of freedom [5]. This method calculates an inverse function mapping from the probability distributions of the rate of aggregated-flows to those of individual-flows. This method has been shown to be feasible under the assumptions that the probability distributions of the rate of individual-flows are independent of each other and the probability that an individual-flow rate becomes zero is positive. As long as these assumptions hold, the inverse function method can uniquely identify the probability distributions of the rate of individual-flows according to the measurement data of the rate of aggregated-flows. However, because the inverse function

method is executed recursively and the degree of freedom of the discrete probabilistic model used in this method is very high, the accuracy of estimation is heavily influenced by the initial values of recursion and variation of the measurement data.

In order to avoid such difficulties and improve estimation accuracy, we propose a method based on the resampling of measurement data. Specifically, we generate a large number of replicated measurement data through repeated resampling of the measurement data of aggregated-flows and generate various probabilistic models for individual-flows by applying the inverse function method using these replicated data. The generated probabilistic models provide a set of solution candidates for the ODTM estimation problem, which are expected to be distributed around the true solution. We then reconstruct the probability distribution of aggregated-flows from the estimation result and evaluate the consistency of the reconstructed aggregated-flows with the observed aggregated-flows.

The remainder of this paper is organized as follows. Section II reviews the prior inverse function method and its issues. Section III describes the proposed method and presents an evaluation using a real traffic trace. Section IV is the conclusion of this paper.

II. Inverse Function Method

A. Principle

We consider the simple network model shown in Fig. 1, where three individual-flows pass through routers 1 and 2. The individual flows from network A to C, A to B, and B to C are labeled as 0, 1, and 2, respectively. In the following, based on this simple network model, we will explain the principle of the inverse function method, which is a method for estimating the probability distributions of every individual-flow rate based on N samples measured at each router. Although this paper focuses on this simple network model, the inverse function method can be applied to more general path-topologies [5]. Here, the flow rate is defined as the number of passed packets (or bytes) within a unit measurement period. It has an integer value in the range of $\{0,\ 1,\ \ldots,\ M\}$.

Let $X_i,\ (i = 0,\ 1,\ 2)$ be the discrete random variable for the rate of individual-flow i within a unit measurement period and $Y_j,\ (j = 1,\ 2)$ be the discrete random variable for the rate of aggregated-flow j within a unit measurement period. Then, we have

$$Y_j = X_0 + X_j, \quad (j = 1,\ 2).$$

Additionally, we define Y_{12} as

$$Y_{12} = \max\{Y_1,\ Y_2\}.$$

The probability distributions of $X_i,\ (i = 0,\ 1,\ 2), Y_j,\ (j = 1,\ 2)$ and Y_{12} are defined as

$$\begin{aligned} P_{X_i}(m) &= \Pr\{X_i = m\}, \\ F_{X_i}(m) &= \Pr\{X_i \le m\}, \\ F_{Y_j}(m) &= \Pr\{Y_j \le m\}, \\ F_{Y_{12}}(m) &= \Pr\{Y_{12} \le m\}, \end{aligned}$$

where $m = 0,\ 1,\ \ldots,\ M$. If we assume that $X_i,\ (i = 0,\ 1,\ 2)$ are independent of each other, then we can define the relationship between the probability distributions of individual-flows (unobservable) and those of aggregated-flows (observable) as follows:

$$F_{Y_j}(m) = \sum_{k=0}^{m} P_{X_j}(m-k) F_{X_j}(k), \tag{1}$$

$$F_{Y_{12}}(m) = \sum_{k=0}^{m} P_{X_0}(m-k) F_{X_1}(k) F_{X_2}(k). \tag{2}$$

If we assume that $P_{X_i}(0) > 0,\ (i = 0,\ 1,\ 2)$ are satisfied, then $F_{Y_1}(0) > 0,\ F_{Y_2}(0) > 0,\ F_{Y_{12}}(0) > 0$ are also satisfied based on the assumption of independence among $X_i,\ (i = 0,\ 1,\ 2)$. In this case, Eqs. (1) and (2) have the inverse functions and $P_{X_0},\ P_{X_1},\ P_{X_2}$ can be calculated from $F_{Y_1},\ F_{Y_2},\ F_{Y_{12}}$ recursively as follows [5]: This is why this method is called the inverse function method.

If $m = 0$

$$\begin{aligned} P_{X_0}(0) &= \frac{F_{Y_1}(0) F_{Y_2}(0)}{F_{Y_{12}}(0)}, \\ P_{X_1}(0) &= \frac{F_{Y_{12}}(0)}{F_{Y_2}(0)}, \\ P_{X_2}(0) &= \frac{F_{Y_{12}}(0)}{F_{Y_1}(0)}. \end{aligned}$$

If $m \ge 1$

$$\begin{aligned} P_{X_0}(m) &= \frac{-b(m) - \sqrt{b(m)^2 - 4ac(m)}}{2a}, \\ a &= \frac{P_{X_1}(0) P_{X_2}(0)}{P_{X_0}(0)}, \\ b(m) &= \frac{F_{Y_{12}}(0) - P_{X_1}(0) B_2(m) - P_{X_2}(0) B_1(m)}{P_{X_0}(0)}, \\ c(m) &= \frac{B_1(m) B_2(m)}{P_{X_0}(0)} - C(m), \\ B_j(m) &= F_{Y_j}(m) - \sum_{k=1}^{m-1} P_{X_0}(m-k) F_{X_j}(k), \quad (j = 1,\ 2), \\ C(m) &= F_{Y_{12}}(m) - \sum_{k=1}^{m-1} P_{X_0}(m-k) F_{X_1}(k) F_{X_2}(k), \\ P_{X_1}(m) &= \frac{-P_{X_1}(0) P_{X_0}(m) + B_1(m)}{P_{X_0}(0)} - F_{X_1}(m-1), \\ P_{X_2}(m) &= \frac{-P_{X_2}(0) P_{X_0}(m) + B_2(m)}{P_{X_0}(0)} - F_{X_2}(m-1). \end{aligned}$$

By using the probability distributions identified by the inverse function method, we can calculate various statistics,

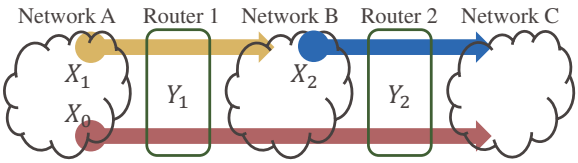

Fig. 1. Network Topology

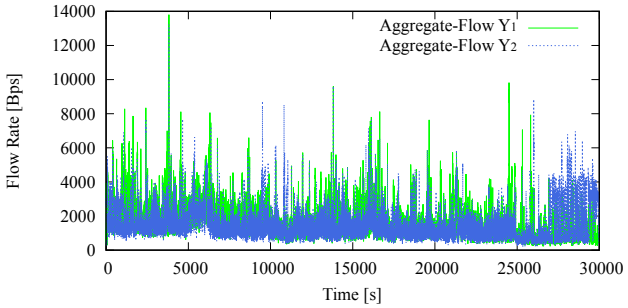

Fig. 2. Example rates of aggregated-flow Y_1 and Y_2

such as the expected value and variance of individual-flows $X_i,\ (i = 0,\ 1,\ 2)$. In this paper, we consider the problem of estimating the expected values of individual flows.

B. Minimum Effective Flow Rate

In the formulation of the inverse function method, it is assumed that $P_{X_0}(0) > 0$, $P_{X_1}(0) > 0$, and $P_{X_2}(0) > 0$ (i.e., $F_{Y_1}(0) > 0$, $F_{Y_2}(0) > 0$, and $F_{Y_{12}}(0) > 0$) hold. However, these assumptions do not hold in general. One way to avoid this problem is to introduce a minimum effective flow rate that is defined by the observed minimum flow rate and estimate the incremental quantity from the minimum effective flow rate [6]. In this paper, we consider a method based on this concept, which is described below.

Let $y_{j,n}$ be the measurement sample of aggregated-flows $Y_j,\ (j = 1,\ 2)$ at time stamp n, where $n = 1,\ 2,\ \ldots, N$ and $S_j = \{y_{j,1},\ y_{j,2},\ \ldots,\ y_{j,N}\}$ be the set of measurement samples. Then, the empirical distribution derived from S_j can be expressed as

$$\widehat{F}_{Y_j}(m) = \frac{1}{N}\sum_{n=1}^{N} I(y_{j,n} \le m), \quad (j = 1,\ 2) \tag{3}$$

$$\widehat{F}_{Y_{12}}(m) = \frac{1}{N}\sum_{n=1}^{N} I(y_{12,n} \le m), \tag{4}$$

where $y_{12,n} = \max\{y_{1,n}, y_{2,n}\}$ and $I(\cdot)$ represent the indicator function that is equal to 1 when the logical statement inside the parenthesis is true and equal to 0, otherwise. Now, if $\widehat{F}_{Y_j}(0) > 0$, $(j = 1,\ 2)$, and $\widehat{F}_{Y_{12}}(0) > 0$ hold, we can implement the inverse function method by replacing $F_{Y_j}(m)$ in Eq. (1) with $\widehat{F}_{Y_j}(m)$ and $F_{Y_{12}}(m)$ in Eq. (2) with $\widehat{F}_{Y_{12}}(m)$.

If the above assumptions do not hold, we estimate the incremental quantity from the minimum effective rate. We define the incremental quantity of the measured aggregated-flow $y'_{j,n}$ based on the minimum flow rate as

$$y'_{j,n} = y_{j,n} - y_{j,\min}, \quad (j = 1,\ 2),$$

where $y_{j,\min}$ is the minimum effective rate defined as

$$y_{j,\min} = \min_{n=1,2,\ldots,N} y_{j,n}, \quad (j = 1,\ 2).$$

Then, we define $y'_{12,n}$ as

$$y'_{12,n} = \max\{y'_{1,n},\ y'_{2,n}\}.$$

Note that the minimum value in $\{y'_{12,1}, \ldots, y'_{12,N}\}$, which is defined by

$$y'_{12,\min} = \min_{n=1,2,\ldots,N} y'_{12,n},$$

does not necessarily become 0. Therefore, we define $y''_{j,n}$ and $y''_{12,n}$ as

$$y''_{j,n} = \max\{y'_{j,n} - y'_{12,\min},\ 0\}, \quad (j = 1,\ 2)$$
$$y''_{12,n} = y'_{12,n} - y'_{12,\min}.$$

The empirical distribution derived from the adjusted data set of measurement samples, written as $S''_j = \{y''_{j,1},\ y''_{j,2},\ \ldots,\ y''_{j,N}\}$, is defined as

$$\widehat{F}_{Y''_j}(m) = \frac{1}{N}\sum_{n=1}^{N} I(y''_{j,n} \le m), \quad (j = 1,\ 2)$$

$$\widehat{F}_{Y''_{12}}(m) = \frac{1}{N}\sum_{n=1}^{N} I(y''_{12,n} \le m),$$

which satisfies $\widehat{F}_{Y_j}(0) > 0$ and $\widehat{F}_{Y_{12}}(0) > 0$. Therefore, we can implement the inverse function method by replacing $F_{Y_j}(m)$ in Eq. (1) with $\widehat{F}_{Y''_j}(m)$ and $F_{Y_{12}}(m)$ in Eq. (2) with $\widehat{F}_{Y''_{12}}(m)$. The expected value of X_i derived from the probability distribution of the individual-flow resulting from the above method is represented by $\widehat{\mu}''_i$.

In this study, considering that the above process was applied to the original measurement data of the aggregated-flow rate, we estimated the expected value of the individual flow i, $\widehat{\mu}_i$ as

$$\widehat{\mu}_i = \widehat{\mu}''_i + \alpha_i, \quad i = 0,\ 1,\ 2,$$

where

$$\alpha_0 = \frac{\alpha_1 + \alpha_2}{2}, \quad \alpha_j = \frac{y''_{j,\min} + y''_{12,\min}}{2}, \quad (j = 1,\ 2).$$

C. Quantization of Unit Quantity for Estimation

Because the degree of freedom of the discrete probabilistic model used in the inverse function method is $O(M)$, it can become very large depending on the maximum value of the flow rate. Additionally, because the inverse function method is executed recursively, its estimation accuracy is heavily affected by the variation of the measurement data of aggregated-flow rate used for estimation. Therefore, in this study, we optimized the quantization width (i.e., the degree of freedom) by considering the results presented in [6].

III. Proposed Method and Its Performance Evaluation

A. Performance Improvement by Resampling

1) Proposed Method: In this paper, we propose a method to improve the estimation accuracy of the inverse method by evaluating the impact of variation in the measurement data on the estimation results. The proposed method is based on resampling measurement data. Specifically, we generate a large number of replicated measurement data through repeated resampling on measurement data from aggregated-flows. We then generate various probabilistic models for individual-flows by applying the inverse function method based on these replicated data, where the degrees of freedom of the probabilistic models are chosen randomly. These probabilistic models provide a set of solution candidates for the ODTM estimation problem, which are expected to be distributed around the true solution. We then reconstruct the probability distributions of aggregated-flows from the estimated probability distributions of individual-flows. In addition, we evaluate the consistency of the reconstructed probability distributions of aggregated-flows with the observed (empirical) probability distributions of aggregated-flows. The main concept behind the proposed method is not to mitigate the impact of variation in the measurement data, but to avoid the impact by evaluating the expected estimation accuracy of solution candidates.

Let $\widetilde{T}$ be a multiset of N samples that are randomly selected with replacement from the set $T = \{1,\ 2,\ \ldots,\ N\}$. In other words, $\widetilde{T}$ is a bootstrap sample set randomly selected from the set $T = \{1,\ 2,\ \ldots,\ N\}$. Therefore, $\widetilde{S}_j = \{y_{j,n} \mid n \in \widetilde{T}\}$ is a multiset of N samples randomly selected with replacement from the original set of samples $S_j = \{y_{j,n} \mid n \in T\}$, which is the entire set of measurement data of aggregated-flows Y_j, $(j = 1,\ 2)$. Then, the empirical distribution derived from $\widetilde{S}_j$ can be calculated as

$$\widetilde{F}_{Y_j}(m) = \frac{1}{N}\sum_{n\in\widetilde{T}} I(y_{j,n} \le m), \quad (j = 1,\ 2) \tag{5}$$

$$\widetilde{F}_{Y_{12}}(m) = \frac{1}{N}\sum_{n\in\widetilde{T}} I(y_{12,n} \le m). \tag{6}$$

Now, we get $\widetilde{\mu}_i$, which is the expected value of the individual-flows X_i, $(i = 0,\ 1,\ 2)$ in terms of the aforementioned empirical probability distribution of individual-flows derived from the inverse function method discussed in Sec. II.

Because the true probability distributions of individual-flows X_i, $(i = 0,\ 1,\ 2)$ are unknown, it is impossible to evaluate the validity of the estimator $\widetilde{\mu}_i$ directly. In other words, although the relative error between μ_i^* (the expectation of X_i in terms of the true probability distribution of the individual-flows) and $\widetilde{\mu}_i$ can be defined as

$$e^* = \sqrt{\sum_{i=0}^{2}\left(\frac{\mu_i^* - \widetilde{\mu}_i}{\mu_i^*}\right)^2}, \tag{7}$$

it is impossible to evaluate e^* in a real-world scenario because μ_i^* are unknown. Therefore, we evaluate the validity of the estimator $\widetilde{\mu}_i$ based on relative error in terms of the aggregated-flow rate, which is defined as

$$\widehat{e} = \sqrt{\sum_{j=1}^{2}\left(\frac{\widehat{\lambda}_j - (\widetilde{\mu}_0 + \widetilde{\mu}_j)}{\widehat{\lambda}_j}\right)^2}, \tag{8}$$

where $\widehat{\lambda}_j$ is the sample mean of the measurement data $S_j = \{y_{j,n} | n \in T\}$, defined as

$$\widehat{\lambda}_j = \frac{1}{N}\sum_{n\in T} y_{j,n},$$

and $\widetilde{\mu}_0 + \widetilde{\mu}_j$ are the estimators of the rate of aggregated-flow Y_j.

In the proposed method, we generate a large number of replicated measurement data of aggregated-flows, $\widetilde{S}_j$, through repeated resampling of the measurement data of aggregated-flows, S_j, and apply the above method to each replica. We then evaluate the validity of the estimation based on the relative error defined by Eq. (8) and select the estimator $\widetilde{\mu}_i$ whose evaluation results with respect to $\widehat{e}$ are the best as the ultimate estimator.

2) Results of Evaluation: We evaluated the proposed method by using a real traffic trace captured at the campus network of the Kyushu Institute of Technology, Japan. The real traffic trace of the aggregated-flows Y_1 and Y_2 used for

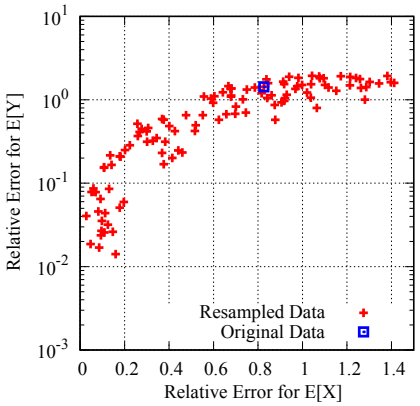

Fig. 3. Relative Error ($b = 3$)

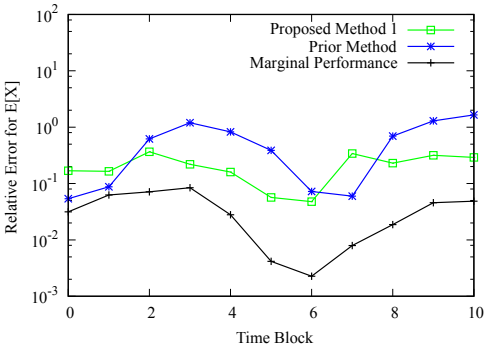

Fig. 4. Comparison of Relative Error

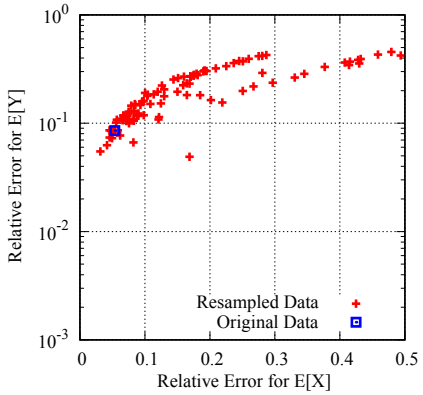

Fig. 5. Relative Error ($b = 0$)

evaluation was captured over 30,000 seconds. An example time series from the real traffic trace used in this paper is presented in Fig. 2. For evaluation, the proposed method was applied to time series data from intervals of 5,000 consecutive seconds, which constitute evaluation units (i.e., $N = 5000$). The intervals were shifted by 2,500 seconds. In other words, the real traffic trace for 5,000 seconds of aggregated-flow rate Y_j, $j = 1,\ 2$ in the time intervals $b = 0,\ 1,\ \ldots,\ 10$, denoted $S_j^{(b)} = \{y_{j,n+2500b} \mid n \in T\}$, was evaluated by the proposed method. Regarding the units of estimation for implementing the inverse function method, the quantization number (i.e., degree of freedom) q was randomly chosen between 20 and 80. Therefore, the quantization width was $\lfloor M/q \rfloor$.

Figures 3, 4, and 5 present the results of estimation obtained by the proposed method using the real traffic trace illustrated in Fig. 2.

Figure 3 presents an example of the estimation results obtained by the proposed method using the real traffic trace of aggregated-flow within a time block $S_j^{(b)}$, $(b = 3)$. The horizontal axis denotes the relative error in terms of the individual-flow rates e^* and the vertical axis denotes the relative error in terms of the aggregated-flow rate $\hat{e}$. In the proposed method, we use replicas of the measurement data of the aggregated-flow $S_j^{(b)}$ that are obtained through repeated resampling of $S_j^{(b)}$. This figure presents the results of resampling 150 times, where each $+$ mark in the figure corresponds to one replica (i.e., the number of $+$ marks is 150). From the figure, it can be seen that the estimation error varies depending on the number of replicas used in the proposed method. Additionally, we can confirm the tendency that the relative error e^* decreases as the relative error $\hat{e}$ decreases. Therefore, by using the estimation result $\tilde{\mu}_i$, $i = 0,\ 1,\ 2$ when the relative error $\hat{e}$, which can be calculated from the data of the aggregate flow rate, is minimized, it is possible to reduce the true relative error e^*, which cannot be calculated from the measurement data of the aggregate flow rate. In fact, as shown in this figure, the estimation results from the proposed method can decrease relative error compared to the estimation results obtained by the prior method, which uses the original measurement data $S_j^{(b)}$ for the implementation of the inverse function method (i.e., the method without resampling).

Figure 4 presents the time series of relative errors e^* for each time block from the proposed method (Proposed Method 1) and prior method (Prior Method). Additionally, this figure presents the marginal performance obtained when the optimal solution could be selected (Marginal Performance) as a reference. From the figure, one can see that the proposed method generally achieved higher estimation accuracy than the prior method, even in cases other than $b = 3$. However, in the cases of $b = 0, 1, 7$, the prior method achieved higher estimation accuracy than the proposed method. In order to investigate this result in greater detail, as an example, a scatter plot of the relative error e^* in terms of the individual-flow rate and the relative error $\hat{e}$ in terms of the aggregated-flow rate for the case of $b = 0$ is presented in Fig. 5. From this figure, the correlation between the relative error in terms of the individual-flow rate e^* and the relative error in terms of the aggregated-flow rate $\hat{e}$ shows a similar tendency to that in Fig. 3. However, one can see that the relative error e^* for the proposed method is larger than that for the prior method when the relative error $\hat{e}$ is minimized. This indicates that the estimation error in terms of the individual-flow rate X_0 and that in terms of the aggregated-flow rate X_j are canceled because of the relationship between the aggregated-flow rate and individual-flow rate $Y_j = X_0 + X_j$. In other words, e^* is not necessarily minimized when $\hat{e}$ is minimized. This phenomenon is caused by the fundamental indefiniteness in the ODTM estimation problem. We propose a method that can achieve high estimation accuracy in such cases in the following section.

B. Solution Filtering

1) Proposed Method: Figure 6 presents the relationship between the selected solution candidate for $b = 0$ and the true solution discussed in the previous section. The point indicated by the circle (red) is the true value and the point indicated by diamond (green) is the solution selected by the proposed method from the previous section. Figure 7 presents an enlarged view of the points surrounding the true value. The straight line shown the figure indicates a region satisfying the relationship $\mu_j^* + \mu_0^* = X_j + X_0$, $(j = 1,\ 2)$. In the proposed method from the previous section, because the point with the smallest squared distance from the straight line was selected as a solution, a point far from the true value was selected even though there are many points surrounding the true value. Therefore, we propose a method to remove such points from the set of candidate solutions.

Let the set of estimation results obtained by performing resampling l_1 times be $U_{l_1} = \{(\tilde{\mu}_0^{(j)}, \tilde{\mu}_1^{(j)}, \tilde{\mu}_2^{(j)}) | \ j = 1, \cdots, l_1\}$. Then, we calculate the relative error $\hat{e}$ for each element in U_{l_1}. We expect that there is some correlation between relative errors $\hat{e}$ and e^*. Therefore, l_2 elements are selected in ascending order with respect to the relative error $\hat{e}$ from the elements in U_{l_1}. The set of selected elements is denoted U'_{l_2} $(\subset U_{k_1})$. In this manner, the points greatly deviating from

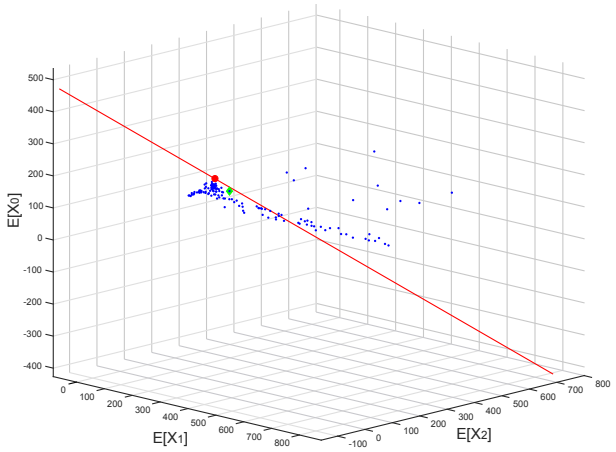

Fig. 6. Distribution of Estimation Results

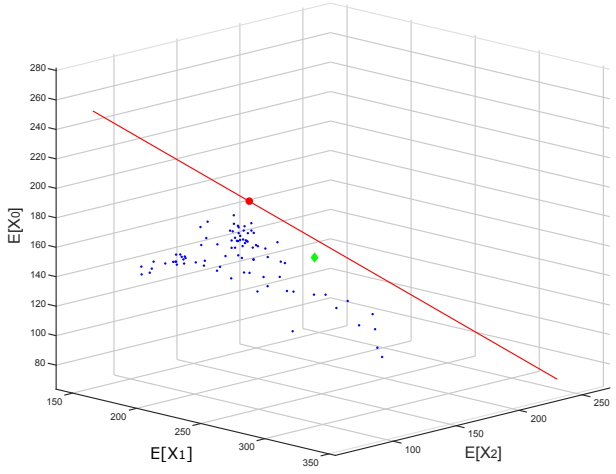

Fig. 7. Enlarged View of Fig. 6

the straight line (Fig. 6) are removed from the set of candidate solutions.

The center of the distribution of elements in the set U'_{l_2} is expected to be close to the true value, as shown in Fig. 7. Therefore, we filter the candidate solutions based on this property. Here, let $\tilde{\mu}_{i,\mathrm{med}}$ be the median of the elements in the set $U'_{i,l_2} = \{\tilde{\mu}_i | (\tilde{\mu}_0, \tilde{\mu}_1, \tilde{\mu}_2) \in U'_{l_2}\}$ and $\boldsymbol{\mu}_{\mathrm{med}} = (\tilde{\mu}_{0,\mathrm{med}}, \tilde{\mu}_{1,\mathrm{med}}, \tilde{\mu}_{2,\mathrm{med}})$ be a vector composed of these medians. Then, let U''_{l_3} $(\subset U'_{l_2})$ be the set of l_3 elements that are selected from U'_{l_2} in ascending order with respect to the distance from $\boldsymbol{\mu}_{\mathrm{med}}$. Finally, let the element in U''_{l_3} for which the relative error $\hat{e}$ is minimized be the final estimator.

2) Results of Evaluation: Figure 8 presents the time series of relative errors e^* for each time block from the above method (Proposed Method 2), where $l_1 = 150, l_2 = 50, l_3 = 40$. The parameters values were determined through a trial and error process. From this figure, one can see that the errors in the time blocks of $b = 0, 1$ are smaller than those from the prior method. Additionally, it can be seen that the error in the time block of $b = 7$ is smaller than that of Proposed Method 1,

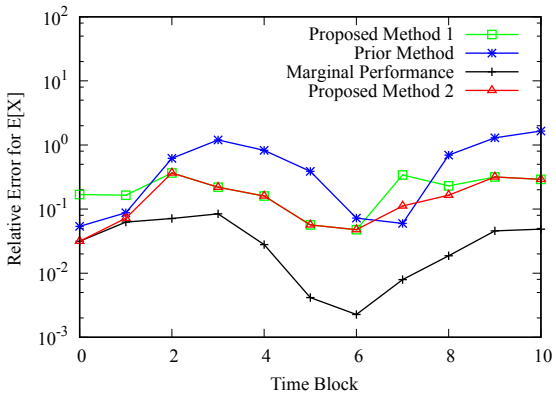

Fig. 8. Comparison of Relative Errors

although it is still inferior to the error of the prior method.

IV. Conclusions

The accuracy of ODTM estimation based on the inverse function method is heavily affected by the variation of the measurement data of the aggregated-flow rate. In this paper, we proposed a method to improve the estimation accuracy of the inverse function method by evaluating the impact of variation in measurement data on estimation results. The proposed method generates a large number of replicas through repeated resampling of the measurement data of the aggregated-flows and generates various probabilistic models for individual-flows (i.e., solution candidates) by applying the inverse function method using these replicated data. Additionally, by considering the characteristics of the solution candidates obtained through repeated resampling, the solution candidates that are considered to have low estimation accuracy are excluded from the set of candidate solutions. As a result of evaluations using a real traffic trace, we confirmed that the proposed method achieves higher estimation accuracy than the prior method in many cases.

Acknowledgment

This work was supported in part by the Japan Society for the Promotion of Science through Grants-in-Aid for Scientific Research (C) (17K00135).

References

[1] P. Tune and M. Roughan. Internet traffic matrices: A primer. In H. Haddadi and O. Bonaventure, editors, *Recent Advances in Networking*, volume 1. ACM SIGCOMM eBook, August 2013.

[2] Qi Zhao, Zihui Ge, Jia Wang, and Jun Xu. Robust traffic matrix estimation with imperfect information: Making use of multiple data sources. *SIGMETRICS Perform. Eval. Rev.*, 34(1):133–144, 2006.

[3] J. Cao, D. Davis, S. V. Wiel, and B. Yu. Time-varying network tomography: Router link data. *Journal of the American Statistical Association*, 95:1063–1075, 2000.

[4] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg. Fast accurate computation of large-scale ip traffic matrices from link loads. In *Proceedings of ACM SIGMETRICS 2003*, pages 206–217, 2003.

[5] M. Tsuru, T. Takine, and Y. Oie. Inferring traffic flow characteristics from aggregated-flow measurement. *IPSJ Journal*, 43(11):3291–3300, 2002.

[6] M. Tsuru, T. Takine, and Y. Oie. Inferring arrival rate statistics of individual flows from aggregated-flow rate measurements. In *Proceedings of IEEE SAINT 2003*, pages 257–266, 2003.