Improving Transportation Information Resilience: Error Estimation for Networked Sensor Data

June 2020

A Research Report from the National Center for Sustainable Transportation

Yueyue Fan, University of California, Davis
Han Yang, University of California, Davis
Saurabh Maheshwari, University of California, Davis
Yudi Yang, University of California, Davis

National Center for Sustainable Transportation

ITS UCDAVIS INSTITUTE OF TRANSPORTATION STUDIES

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. NCST-UCD-RR-20-18	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle Improving Transportation Information Resilience: Error Estimation for Networked Sensor Data		**5. Report Date** June 2020 **6. Performing Organization Code** N/A
7. Author(s) Yueyue Fan, PhD https://orcid.org/0000-0002-6615-2029 Han Yang Saurabh Maheshwari Yudi Yang, PhD		**8. Performing Organization Report No.** UCD-ITS-RR-20-17
9. Performing Organization Name and Address University of California, Davis Institute of Transportation Studies 1605 Tilia Street, Suite 100 Davis, CA 95616		**10. Work Unit No.** N/A **11. Contract or Grant No.** Caltrans 65A0686 Task Order 003 USDOT Grant 69A3551747114
12. Sponsoring Agency Name and Address California Department of Transportation Division of Research, Innovation and System Information, MS-83 1727 30th Street, Sacramento, CA 95816		**13. Type of Report and Period Covered** Final Report (December 2018 – November 2019) **14. Sponsoring Agency Code** USDOT OST-R

15. Supplementary Notes
DOI: https://doi.org/10.7922/G2610XK9
Dataset DOI: https://doi.org/10.25338/B8TP5Q

16. Abstract
Nowadays, the effectiveness of any smart transportation management or control strategy would heavily depend on reliable traffic data collected by sensors. Two problems regarding sensor data quality have received attention: first, the problem of identifying malfunctioning sensors; second, reconstruction of traffic flow. Most existing studies concerned about identifying completely malfunctioning sensors whose data should be discarded. This project focuses on the problem of error detection and data recovery of partially malfunctioning sensors that could provide valuable information. By integrating a sensor measurement error model and a transportation network model, the authors propose a Generalized Method of Moments (GMM) based estimation approach to determine the parameters of systematic and random errors of traffic sensors in a road network. The proposed method allows flexible data aggregation that ameliorates identification and accuracy. The estimates regarding both systematic and random errors are utilized to conduct hypothesis test on sensor health and to estimate true traffic flows with observed counts. The results of three network examples with different scales demonstrate the applicability of the proposed method in a large variety of scenarios. This research improves fundamental knowledge on transportation data analytics as well as the effective management of data and information infrastructure in transportation practice.

17. Key Words Sensor health, error estimation, information resilience, data science, traffic network	**18. Distribution Statement** No restrictions.

19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 45	**22. Price** N/A

NCST

About the National Center for Sustainable Transportation

The National Center for Sustainable Transportation is a consortium of leading universities committed to advancing an environmentally sustainable transportation system through cutting-edge research, direct policy engagement, and education of our future leaders. Consortium members include: University of California, Davis; University of California, Riverside; University of Southern California; California State University, Long Beach; Georgia Institute of Technology; and University of Vermont. More information can be found at: ncst.ucdavis.edu.

Disclaimer

Acknowledgments

This study was funded, partially or entirely, by a grant from the National Center for Sustainable Transportation (NCST), supported by the U.S. Department of Transportation (USDOT) and the California Department of Transportation (Caltrans) through the University Transportation Centers program. The authors would like to thank the NCST, the USDOT, and Caltrans for their support of university-based research in transportation, and especially for the funding provided in support of this project.

Improving Transportation Information Resilience: Error Estimation for Networked Sensor Data

A National Center for Sustainable Transportation Research Report

June 2020

Yueyue Fan, Department of Civil and Environmental Engineering, University of California, Davis

Han Yang, Department of Civil and Environmental Engineering, University of California, Davis

Saurabh Maheshwari, Department of Civil and Environmental Engineering, University of California, Davis

Yudi Yang, Department of Civil and Environmental Engineering, University of California, Davis

NCST

[page intentionally left blank]

TABLE OF CONTENTS

List of Figures

Improving Transportation Information Resilience: Error Estimation for Networked Sensor Data

EXECUTIVE SUMMARY

Reliable and resilient data serves as a critical foundation for almost all new mobility technologies in a modern transportation system, thus plays a significant role in improving the efficiency, safety, and ultimately sustainability of a transportation system. Despite of the importance of sensor data performance evaluation, research effort devoted to this subject is rather thin compared to studies that directly take sensor data as input. Most sensor health monitoring studies focused on identifying 'bad' sensors whose data should be discarded. The problem of estimating and correcting the systematic bias of sensor data has not been addressed, especially in the context of large-scale networked data pieces from heterogeneous sources.

In this project, we fill the aforementioned critical gap - by combining data science and network modeling techniques, our method provides greater modeling flexibility to incorporate spatial correlation of networked data, which leads to a better estimation quality. Main research steps include:

- **Model Creation:** We have developed a mathematical model combining moment matching and traffic network models. Using synthesized data generated for networks with different graph properties (including orders, sizes, connectivities, etc.) and different flow patterns, we test the estimation quality of the proposed method using the basic knowledge of network flow balance.
- **Model testing and validation:** Real-world traffic data and computer simulations were used for model testing and validation, including road segments and networks in urban congested areas (such as using PeMS data in Orange County and the San Francisco Bay Area). These networks can simulate urban corridors and generate large-scale and multiple types of traffic data. Note that since the true measurement errors are unknown, simple comparison between estimated and true systematic bias will not be possible. However, the estimated results may be used to provide information for inspection and maintenance activities.

This research improves fundamental knowledge on transportation data analytics as well as the effective management of data and information infrastructure in transportation practice. The open-source computer program created from this project will be shared by the California Department of Transportation to facilitate knowledge transfer from academia to practice.

1. Introduction

In modern transportation systems, reliable sensor data is heavily relied on to produce effective planning and operational strategies in coping with almost all-important issues, including road congestion, traffic safety, pollutant emission control and energy consumption. A major challenge pertaining to sensor data is to deal with corrupted data or even completely missing data that frequently and widely occurs in most established traffic monitoring systems. Many empirical case studies have evidenced that data conflicts and missing records exist in a large amount of road traffic sensors. For example, it was reported that about one third of the freeway sensors in PeMs (California Performance Measurements), a broadly referenced data system, were not working properly (Rajagopal and Varaiya, 2007). Quality control for archived data management systems (ADMS) has also been identified as a high-priority task recommended to the Federal Highway Administration (Turner, 2007).

Research efforts devoted to traffic sensor data quality in the last few decades can be divided into two categories. The first one attempts to address the issues of assessing data quality and identifying completely malfunctioning sensors. It is usually referred to as sensor health problem and predominantly treated as a pure engineering task that merely requires traffic domain expertise. The second category focuses on remedying the corrupted data in a systematic manner using all available data. The solutions to these problems are usually data oriented and statistical learning based without fully considering traffic data structure. The literature suggests both the importance of having a solid statistical basis to infer sensor quality from a network perspective as well as the necessity of assimilating useful knowledge on data rectification. On account of those matters, this article spans over the two categories via developing a statistical inference approach for data quality assessment and reconstruction based on a transportation network model.

Most existing works on sensor health problems focused on identifying completely malfunctioning sensors whose data should be directly discarded, but few paid attention to moderately malfunctioning ones whose data are significantly erroneous yet still endow useful information. The pioneer endeavors among them mainly depend on setting allowable range for observed values and checking consistency among volumes, occupancy and speeds. Over the years, studies following the same school of thoughts have evolved to include more complicated validity criteria combinations (Turochy and Smith, 2000, Hu et al., 2001, Chen et al., 2003, and Turner et al., 2004). Nowadays, they are still prevailing in practice due to its convenient implementation in a conventional database management system. The other branch of works leverages the mutual dependency of traffic data from closely located sensors and adjacent time intervals. Spatially, the correlation of traffic counts are modeled based on neighboring lane similarity (Dailey, 1993), upstream and downstream consistency (Nihan, 1997), macroscopic traffic flow conservation (Vanajakshi and Rilett,2004) and simply proximity in distance (Kwon et al., 2004). Recently, Sun et al. (2016) pointed out the limitation of earlier studies which did not fully exploit spatial correlations on the network level and proposed a new approach to identify malfunctioning sensors of all possible reasons whose data are supposed to be significantly inconsistent against data from others. This paper shares a similar network perspective in

defining spatial dependency but in a more flexible manner that requires much weaker assumption to establish. The major distinction of this work from previous studies in this area is the capability of telling the magnitude of data corruption and identifying partially malfunctioning sensors. This virtue is of evident practical value because it is beneficial for practitioners to be able to utilize information from those sensors to reconstruct traffic data.

Till date, the majority of research efforts in the area of data remediation is to handle completely missing data on the basis of uncorrupted data from other sensors. In many cases, data imputation methods based on time-series analyses and machine learning approaches are applied only after all susceptible data from malfunctioning sensors have been completely removed. Li et al. (2014)'s review pointed out that traditional prediction methods using time series model such as ARIMA to map historical and future values of traffic data failed to fully utilize the observed data succeeding to missing data occurrence. Interpolation using spatially and temporally adjacent records is prevailing in highway agencies, but forcing counts to be close to each other may underestimate the traffic variation in the corresponding dimensions. A large body of recent literature utilizes learning algorithms in searching for a pattern of traffic data, including for example, Probabilistic Principal Component Analysis in Qu et al. (2009), Fuzzy C-means Clustering in Tang et al. (2015) and Deep Learning in Duan et al. (2016). To the best of our knowledge, though being diverse in terms of employed learning models, none of those studies considers the possibility of systematic errors in the observation datasets, which could potentially mislead the learning outcomes.

According to Traffic Detector Handbook published by the US Federal Highway Administration (Klein et al., 2006), there exist different levels of sensor problems, ranging from most obvious ones such as zero call or constant call, to modest but less detectable errors, such as unbalanced sensitivity. Sensor data that are systematically deviated from the real traffic volume, out of the reasons such as counting neighboring lane traffic, missing motorcycles, more than one count for long vehicles, may still be valuable in revealing important information on the traffic flow that it is actually monitoring as well as on the other flows in the network. In order to take advantage of those sensors' data, the health monitoring task is not only to pinpoint the malfunctioning detectors, but to measure their respective levels of sensor health. It is equally important to actually carry such obtained knowledge into the steps of reconstructing traffic flow.

In this project, the health of a sensor is represented by its measurement error, which can be modeled mathematically and characterized by its inferred statistics. Measurement errors are usually divided into two components (Dunn, 1989). Systematic error is determined by the inaccuracy that is involved inherently in the observation process. It can be used to measure the level of sensor health problem and to rectify observed values. Random error is, however, natural to any type of measurement. Even with a perfectly functioning detector, the traffic counts can be ostensibly different from true values. Hence, it should not be an indicator of sensor health problem unless its scale is abnormally large, but its related knowledge is important in deriving estimator's properties and conducting statistical inference. Therefore, by integrating a sensor measurement error model and a transportation network model, we

propose a Generalized Method of Moments (GMM) based estimation approach to determine the parameters of systematic and random errors of traffic sensors in a road network. The roles and functionalities of the problem discussed in this paper are illustrated in Figure 1and highlighted in blue. Steps 1 and 2 are the detection of completely and partially malfunctioning sensors, respectively. Step 3 represents standard denoising procedure. Step 4 is to correct systematically erroneous data. Step 5 is to impute missing data.

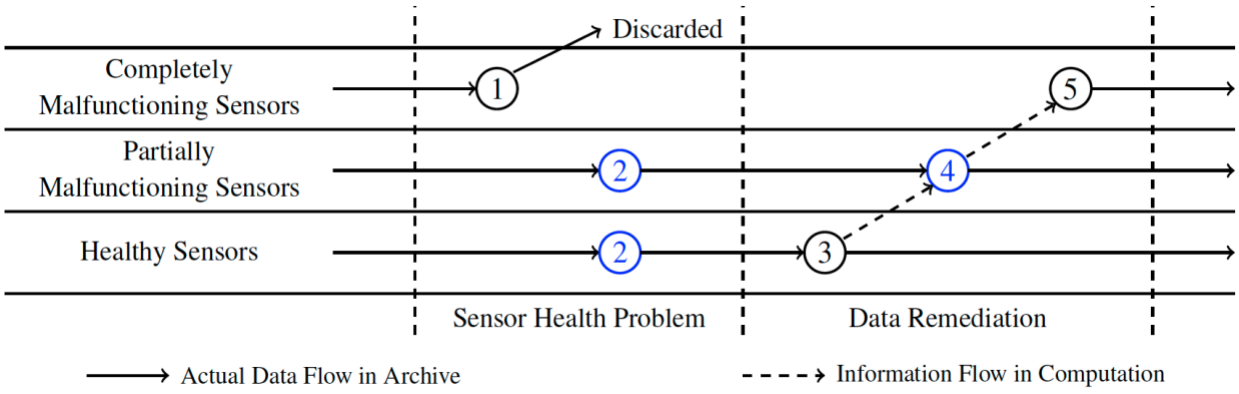


Figure 1. Sensor Data Processing Chart

The rest of the report is organized as follows. The second section provides the detail of sensor measurement error model and describes the way that flow balance law fits into structural equations which serves as a foundation to estimation. The third section introduces a main GMM approach that provides unique and statistically consistent estimates of systematic error parameters. The fourth section discusses the estimation of random error parameters and their uses on sensor health monitoring and traffic data correction. The fifth section first uses a small walk-though example, then demonstrates the numeric robustness of the method with respect to various factors, and finally employs a large-scale case study to examine the scalability. The sixth section concludes the paper with discussion and future extensions.

2. Mathematical Model

Considering the wide range of types and levels of sensor malfunction, it is difficult, if at all possible, to find a universal mathematical model to explain all possible data errors of a roadway sensor. Traffic agencies have extensively conducted univariate tests that use simple but reliable filters to identify completely broken sensors based on their individual outputs. Complementary to those existing efforts, we focus on failure types that are more subtle for two reasons. First, such failure is typically not as easily identifiable as completely broken ones thus needing more in-depth investigation. Second, data generated by those sensors may be erroneous but still can be informative if systematic error can be identified to correct these data.

In this section, we first explain basic assumptions for sensor error generation mechanism and develop mathematical models, from general to specific, for measurement errors. Then we discuss how the important network relation, i.e. flow balance law, should be utilized in the

model setting. In preparation of sensor error estimation, a system of structural equations is generated by integrating the measurement model and flow balance conditions.

2.1. Measurement Errors

Suppose road traffics are continuously monitored by sensors and the number of passing vehicles is reported based on consecutive time intervals. Let the recorded count for s-th vehicle passing the detection scope of sensor a during the measurement period in question be $1 + \in_a^s$, where the registration error $\in_a^s$ is a random variable. Note that $\in_a^s$ is by nature discrete since a passing vehicle either correctly incurs one count or mistakenly zero or more counts. Then for a measurement interval the aggregated traffic count of sensor a is expressed as,

$$V_a = Z_a + \sum_{s=1}^{Z_a} \epsilon_a^s, \tag{2.1}$$

where Z_a is the true value of traffic volume, and V_a is the measured count. Then the total error is expressed as the sum of systematic error and random error,

$$\sum_{s=1}^{z_a} \in_a^s = V_a - Z_a = \underbrace{(E[V_a|Z_a] - Z_a)}_{\text{systematic error}} + \underbrace{(V_a - E[V_a|Z_a])}_{\text{random error}}. \tag{2.2}$$

The basic assumption is that the measurement error generation mechanism is invariant among all the time intervals. It is not hard to justify by restricting estimation horizon to have a suitable time duration. Hence, the parameters that control the error generation are considered fixed over time. As a consequence, we use time independent vectors μ and σ for the parameters related to systematic error and random error, respectively. Without loss of generality, we could write the first two central moments of traffic counts conditioning on Z as deterministic functions of Z, such as,

$$E[V_a|Z_a] = f\left(Z_a; \mu_a\right) \text{ and } \text{Var}[V_a|Z_a] = \varphi(Z_a; \sigma_a^2). \tag{2.3}$$

In this general modeling framework, the exact forms of function f and φ depend on the nature of the registration error ϵ_a^s. For example, if the variance of ϵ_a^s rises when traffic volume Z_a becomes higher, then $\varphi(Z_a; \sigma_a^2)$ should increase faster than Z_a.

The main emphasis of this work is to introduce a statistical method of estimating error and reconstructing flows using networked data. Instead of delving into the detailed discussion of the choices for f and φ, we bring an additional assumption to narrow our focus to a specific model. Assume that ϵ_a^s is independent identically distributed (i.i.d.) with mean μ_a and σ_a, we obtain the statistics of the traffic counts, that is,

$$E[V_a|Z_a] = Z_a + \mu_a Z_a \text{ and } \text{Var}[V_a|Z_a] = \sigma^2 Z_a. \tag{2.4}$$

Hence, $\mu_a Z_a$ is the systematic error of the measurement. Let $U_a = V_a - E(V_a|Z_a)$ denote the random error of the measurement and $\mathrm{Var}(U_a|Z_a) = \mathrm{Var}(V_a|Z_a)$. Because now the moments are linear functions of Z, we also call μ_a the systematic error ratio and σ_a^2 the random error ratio.

A practical concern worth noting is that a road segment typically consists of multiple lanes. For sensors like video cameras or weight tubes, only one sensor unit is needed at a location to capture vehicles on all the lanes. For sensors like inductive loop detectors, multiple detectors are typically embedded in parallel across the road to capture vehicles on all the lanes. Here, we make a simplification that sensors installed on multiple lanes across a road section are considered as one single sensor integrally, referred to as a link-level sensor set. Consequently, all variables in the measurement error estimation model are link specific, indexed by a subscript a. Though link a may have sensors in completely different conditions (for example, one lane may have an overly sensitive sensor that mistakenly records vehicles passing on an adjacent lane, while another lane may have a sensor that fails to count), the model in (2.3) is able to capture the mixed effect of multiple sensors.

2.2. Network Structure

Now let us turn to an important spatial relation between measurements that should be utilized. Consider a traffic network abstracted into a directed graph $\mathcal{G} = \{\mathcal{N}, \mathcal{A}\}$. The flow balance law, i.e., the total flow entering an intermediate node i should be equal to the total flow exiting that node, can be written as,

$$\sum\nolimits_{a\in\mathcal{A}+(i)} Z_a - \sum\nolimits_{a\in\mathcal{A}-(i)} Z_a = 0, \qquad \forall i \in \mathcal{I}, \tag{2.5}$$

where $\mathcal{I} \subset \mathcal{N}$ is the sets of intermediate nodes. $\mathcal{A}^+(i)$ and $\mathcal{A}^-(i)$ are the set of entering links and exiting links, respectively. Let P be the node-link adjacency matrix, whose element on i-th row and a-th column $p_{ia} = 1$ if $a \in \mathcal{A}^+(i)$, –1 if $a \in \mathcal{A}^-(i)$, and 0 otherwise. The size of matrix P is $m \times n$ with $n = |\mathcal{A}|$ and $m = |\mathcal{I}|$. The operation $|\cdot|$ counts the number of elements in its argument. The vector-matrix form of the above equation is

$$PZ = 0, \tag{2.6}$$

where $Z = [Z_a] \in \mathbb{R}_+^n$.

Due to the temporal change in traffic intensity and the presence of congestion shockwaves passing along the paths, the discrepancy between upstream and downstream flow in a time interval may exist, and the flow balance is violated, as

$$\sum\nolimits_{a\in\mathcal{A}^+(i)} Z_a - \sum\nolimits_{a\in\mathcal{A}^-(i)} Z_a = \eta_i, \qquad \forall i \in \mathcal{I}. \tag{2.7}$$

For a well-defined observation interval, η_i is a random variable with a zero mean and a relatively small scale compared to true traffic volume. The flow imbalance ratio is defined as $\tau_i = 2\eta_i/(\sum_{a\epsilon\mathcal{A}^+(i)} Z_a + \sum_{a\epsilon\mathcal{A}^-(i)} Z_a)$. From statistics perspective, it is straightforward to understand that the flow imbalance ratio should approach zero as the time interval of observations extends. It means that flow imbalance is of less concern when the observation interval is long enough. From traffic characteristics, nearly all traffic should be cleared over an observation interval of 24 hours. Therefore, in sensor data studies, link counts are often aggregated to a daily observation to make sure that flow balance law holds to an almost perfect level (for example, in Sun et al. (2016) and Yin et al. (2017)). The disadvantage of cumulating traffic counts by day is that the data sample size might be too small to conduct a proper estimation of sensor error. However, in those studies, it is critical to have nearly zero η_i's because traffic flows in the entire network may be related at the same equation. Clearly, it is much less probable to have link flows on the different edges of a large network conforming flow balance law with marginal discrepancy, though such equations can be derived via linear transformation of P. In statistics, it is easy to comprehend that link flows that are geographically apart with dozens of intermediate nodes in between may have a notably significant discrepancy, since it is equal to the sum of several η_i's which have very likely positive correlation that enhances the imbalance.

In this paper, it is acceptable to divide the entire observation horizon into finer intervals (such as hourly window) and to have flow balance hold in an imperfect but satisfactory level. The first reason is that we use nodal balance law directly at its original form without any affine transformation of Z so that only a single η_i exists in a structural equation that relates neighboring traffic flows. Second, in the estimation method proposed in the subsequent section, η_i can be absorbed into random measurement error and does not affect the identification of systematic error ratios except slight influence on the efficiency of their estimators. This statement will be further illustrated using a numerical experiment in 5.2.4. Unbalanced Flows.

2.3. Structural Equations

We wish to conduct estimation for the parameter vector $\mu = [\mu_a] \in \mathbb{R}^n$ which is critical in evaluating sensor health and correcting traffic counts. In doing so we need to prepare flow balance equations in a form that the chosen estimation principle can be conveniently applied to.

Let us denote the $n \times 1$ nonnegative vector $V = [V_a]$. $V^{(t)}$ and $Z^{(t)}$ are the traffic counts and true flows respectively in observation interval t. The likelihood of μ given the sequence of observed data V^1,…,$V^{(T)}$ is

$$\mathcal{L}(\mu; V^{(1)}, \dots, V^{(T)}) = \int_{PZ^{(1)}=0} \dots \int_{PZ^{(T)}=0} p(V^{(1)}, \dots, V^{(T)}, Z^{(1)} = z^{(1)}, \dots, Z^{(T)} = z^{(T)}; \mu) dz^{(1)} \dots dz^{(T)}, \tag{2.8}$$

where $p(\cdot)$ denotes the joint probability density of traffic counts and true flow. Self-evidently, explicitly handling latent variables Z in the estimation model would require excessive knowledge to characterize the stochasticity of Z. The mutual dependence among link flows Z both spatially and temporally is not merely a result from flow balance but governed by travel demand generation and assignment processes that are interrelated and complicated.

In order to obviate directly dealing with latent Z, we substitute it by observable V and another latent variable U in (2.5). Hence, the coupling of measurement error model and flow balance relation results in a system of structural equations,

$$\sum_{a\in\mathcal{A}^+(i)} f_a^{-1}(V_a - U_a;\mu_a) - \sum_{a\in\mathcal{A}^-(i)} f_a^{-1}(V_a - U_a;\mu_a) = 0, \qquad \forall i \in I. \tag{2.9}$$

To continue the derivation, we have to concentrate on one specific model under the big umbrella (2.3). With the i.i.d. assumption of ϵ_a^s among all s, the equations (2.9) become

$$\sum_{a\in\mathcal{A}^+(i)} \frac{V_a}{1+\mu_a} - \sum_{a\in\mathcal{A}^-(i)} \frac{V_a}{1+\mu_a} = \sum_{a\in\mathcal{A}^+(i)} \frac{U_a}{1+\mu_a} - \sum_{a\in\mathcal{A}^-(i)} \frac{U_a}{1+\mu_a}, \qquad \forall i \in I. \tag{2.10}$$

To simplify this expression, let $\beta = [\beta_a] = [1/(1+\mu_a)]$, so $\beta_a(v_a - u_a) = z_a$, then a concise form of (2.9) is given by

$$P(V \circ \beta = P(U \circ \beta). \tag{2.11}$$

The operator $\circ$ is the Hadamard product[1]. Both sides of (2.11) involve unknown systematic error ratio β. The left hand side contains observables V instead of any latent Z, and the right hand side entails random error $U = [U_a] \in \mathbb{R}^n$.

The system of structural equations (2.11) provides the fundamental relation for GMM principle to estimate systematic error. In this model, μ is strictly greater than –1, because it is meaningless to conduct error estimation for sensors with no counts, which should be a completely broken case. Thus, the parameter space for vector β is $\mathbb{R}^n_{++}$. The case that $\beta_a = 1$ or $\mu_a = 0$ indicates that the sensor on link a does not have any systematic error.

3. Generalized Method of Moment Estimation

Having developed a model combining sensor measurement errors and flow balance law, we now propose an adaptable estimation framework based on GMM principle, which includes both classic moment matching and generalized least square (GLS). The primary concern of this section is whether it is possible to obtain the "correct" estimate of parameter β. There are two

[1] This binary operation takes two matrices/vectors of the same dimensions, and produces another matrix/vector where each element i, j is the product of elements i, j of the original two matrices/vectors.

important issues to be addressed: essentially, parameter identifiability, which is to ensure the resulting method has a unique estimate without ambiguity; furthermore, estimator consistency, which is to ascertain the estimates approaching to the true ones when the data size is sufficiently large.

3.1. Estimation Framework

We define a zero-mean vector-valued stochastic function $g(\beta)$ using relation (2.11). In GMM framework, the estimate of β is found by minimizing a vector norm of $g(\beta)$. Adopting Euclidean distance, we will obtain a minimization problem formulated as

$$\min_{\beta > 0} \mathrm{g}(\beta)^{\top} W \mathrm{g}(\beta) \tag{3.1}$$

where W is a positive-definite weighing matrix, which only affects the rate of estimator's quality improvement against the data sample size. According to GMM theory, the optimal weighting matrix that achieves an efficient estimator of β with minimum variance is the inverse of variance-covariance matrix of random function $g(\beta)$, $\mathrm{Cov}[g(\beta)]$, denoted as Ω. We will further investigate the specification and the updating scheme of W in the next section on statistical inference.

Under this paradigm, a specific statistical estimation method to estimate β is determined by the way that $g(\beta)$ is constructed. For each measurement interval t, we know

$$P\left(V^{(t)} \circ \beta\right) = P\left(U^{(t)} \circ \beta\right), t = 1, \dots, T. \tag{3.2}$$

Then the classic method of moments computes the average from all the observations and solves parameters by matching population moments with their sample analogs, i.e.,

$$g(\beta) = P\left(\frac{1}{T}\sum_{t-1}^{T} V^{(t)} \circ \beta\right), \tag{3.3}$$

then the dimension of $g(\beta)$ is $m \times 1$. Compressing data into its first moment greatly reduces the number of elements in $g(\beta)$ and thereof restricts the amount of information used for estimating β. Oppositely, the GLS method minimizes the sum of squared residuals of $PV^{(t)}\beta$ in all intervals, which utilizes all the observations without any transformation. The generalized moment conditions in this case are now

$$g(\beta) = \left[\left(P\left(V^{(1)} \circ \beta\right)\right)^{\top}, \dots, \left(P\left(V^{(T)} \circ \beta\right)\right)^{\top}\right]^{\top}, \tag{3.4}$$

then the dimension of $g(\beta)$ is $Tm \times 1$. In a more flexible manner, it is possible to aggregate observation data for estimation to reduce the problem scale without losing much structural information. First, traffic counts collected from different time intervals are assigned into K

groups, $T(k), k = 1, .., K$ based on their similarity. Each observation $v^{(t)}$ in the same group does not have to be temporally adjacent. The exact choice of a clustering approach, for instance, K-nearest neighbors, is not critical in this framework. In fact, simply grouping observations based on time-of-day could be a proper choice. Finally, the $g(\beta)$ functions are constructed as

$$\mathrm{g}(\beta) = \left[\left(P\left(\frac{1}{|\mathcal{T}(1)|} \sum_{t \in \mathcal{T}(1)} V^{(t)} \circ \beta \right) \right)^{\mathsf{T}}, \dots, \left(P\left(\frac{1}{|\mathcal{T}(K)|} \sum_{t \in \mathcal{T}(K)} V^{(t)} \circ \beta \right) \right)^{\mathsf{T}} \right]^{\mathsf{T}}. \tag{3.5}$$

with (3.3) and (3.4) being its special cases when $K = 1$ and $K = T$, respectively. The number of elements in $g(\beta)$ is Km then. In the following subsections, the advantages and disadvantages of different grouping strategies will be explained and compared in terms of parameter identification and estimator consistency.

3.2. Parameter Identification

From an algebraic perspective, the minimization problem (3.1) is to solve a homogeneous system of equations with only strictly positive solution permitted as

$$W^{1/2} A\beta = 0, \text{where A} = \begin{bmatrix} A^{[1]} \\ \vdots \\ A^{[K]} \end{bmatrix} \in \mathbb{R}^{Km \times n}, A^{[K]} = P\mathrm{diag}\left(\frac{1}{\mathcal{T}(k)} \sum_{t \in \mathcal{T}(k)} V^{(t)} \right) \in \mathbb{R}^{m \times n}. \tag{3.6}$$

$W \in \mathbb{R}^{Km \times Km}$ should have a block structure where each nontrivial sub-matrix $W^{[k]}$ corresponds to group k located on its diagonal. In the case where the rank of A is less than its number of columns n, the constrained homogeneous system admits infinitely many solutions. Precisely, the solution set is the intersection of strictly positive orthant and the null space of A. Thus, no unique estimate of β can be found by solving (3.1). In the case A has full column rank, since only the trivial solution solves equations (3.6), the Euclidean norm of the estimate using (2.9), $\left\| \hat{\beta} \right\|_2$, will be extremely close to zero. In spite of the fact that the unique estimate theoretically exists, it is of little use to our estimation problem, because then $\mu \to \infty$ and $Z \to 0$ regardless actual values in the dataset. Thus, we are not able to obtain unique and meaningful estimate unless additional information is incorporated.

For a concrete problem, there usually exist a large variety of constraints that can be formulated into problem (3.1) based on knowledge and beliefs towards sensor quality. Here we simply choose a way that is commonly applicable and effective in finding an estimate. When there is a set of sensors recently installed or calibrated in the network, denoted as $\mathcal{A}_1 \subset \mathcal{A}$, these can be treated as free of systematic error, i.e., $\mu_a = 0$. The corresponding constraints are expressed as

$$\beta_a = 1, \forall a \in \forall_1. \tag{3.7}$$

The set of the remaining sensors is denoted $\mathcal{A}_0 = \mathcal{A} - \mathcal{A}_1$. We now create an indicator matrix M_0 for $\mathcal{A}_0$ by removing rows that are not associated with $\mathcal{A}_0$ from an $n \times n$ identity matrix. Similarly, M_1 is made for $\mathcal{A}_1$. For an arbitrary $n \times 1$ vector $x, x_0 = M_0 x$ and $x_1 = M_1 x$. For an arbitrary X with n columns, $X_0 = XM_0^\top$ and $X_1 = XM_1^\top$.

Let β_1 be the subvector of parameters corresponding to those good sensors. β_0 is the subvector that is still unknown. Substituting $\beta_1 = 1$ into the homogeneous system and moving the constant terms to the right hand side, we will acquire a different system of linear equations as follows

$$W^{1/2} A_0 \beta_0 = W^{1/2} b, \text{ with } b = -A_1 \beta_1 \in \mathbb{R}^{Km \times 1}. \tag{3.8}$$

In order to claim that (3.8) is a non-homogeneous system of equations, we only need two simple justifications. First, the good sensors are involved in flow balance relation, so P_1 contains at least one non-zero entry. Then there are traffic counts recorded on those sensors. $V_1^{(t)}, t = 1, \dots, T$ are not all zeros. Now we express $A_1 = [A_1^{[k]^\top}, \dots, A_1^{[k]^\top}]^\top$ with $A_1^{[k]} = P_1 \mathrm{diag}(\sum_{t \in \mathcal{T}(K)} V_1^{(t)}), k = 1, \dots, K$. Given that at least one element in all $V_1^{(t)}, t = 1, \dots, T$ is strictly positive, A_1 is nontrivial. Since $\beta_1 = 1, b$ is not of all zeros. Therefore, without the positiveness constraint which is rarely bounded in practice, by the first order conditions of problem (3.1), the estimate of β_0 is the solution of the non-homogeneous system (3.8) and given as

$$\hat{\beta}_{\mathrm{GMM}} = (A_0^\top W A_0)^{-1} A_0^\top W b, \tag{3.9}$$

if A_0 has a full column rank.

If A_0 is column rank deficient, this linear system is underdetermined and the minimization problem (3.1) admits infinitely many solutions with the same objective values. So let us take a further look at whether this important condition holds for all different aggregation strategies. If none of the elements in vector $\sum_{t \in \mathcal{T}(k)} V_0^{(t)}$ is zero, the rank of each block $A_0^{(k)}$ is equivalent to the rank of P_0, which is bounded above by the number of intermediate nodes m. Typically m is less than the number of links n minus $n_1 = |A_1|$ in a general network, so $A_0^{(k)}$ is not full rank and yields insufficient information to identify β by itself. On account of the variability of $V^{(t)}$ across the observation sets, the stack-up matrix A_0 is possibly full rank. With K groups specified, the rank of A_0 is bounded above by the less value between mK and $n_0 = n - n_1$. The actual rank should be positively correlated with K.

For the extreme strategy $K = 1$, regardless data sample size T, classic moment matching with only the first moments of V is most likely not capable to identify this measurement error model unless $m \geq n_0$. Albeit it is possible to improve identification by incorporating second moment conditions (See 4.1. Estimating Random Errors), those equations involve unknown nuisance parameters and render the optimization problem notably harder to solve.

Besides the column rank of A_0, we need to note that it is also critical to examine the numerical stability of the estimation problem. This relates to the way of selecting group members. If the grouped means are very close, the resulting matrix A_0 will have singular values with very small magnitude. Consequently, the matrix $A_0^{\top} W A_0$ is ill conditioned and the solution to (3.9) is numerically unstable. Therefore, K-mean clusters and simply grouping based on time-of-day are sound choices to have distinct group means.

3.3. Estimator Consistency

The next immediate task is to verify that the unique estimate is statistically consistent, in other words, the estimated values will approach the true ones when the number of observations grows infinitely. Although moment matching method $K = 1$ provides limited information to estimate β_0, it always provides consistent estimator once the model is identified. In contrast, the resultant least square method from $K = T$ is able to provide n_0 linearly independent equations as long as the traffic counts vary enough, but its estimates suffer from "error-in-variable" model and do not converge to true ones even with an infinitely large data sample. This issue generally arises when the correlation between observed values or error terms is significant. In our problem, this is due to the existence of measurement errors in traffic counts, such as $\mathrm{Cov}[V_a^{(t)}, U_a^{(t)}] \neq 0$.

As $V^{(t)} = Z^{(t)}\mathrm{diag}(\beta)^{-1} + U^{(t)},\ t = 1, \ldots, T$, we can expand

$$A = C\mathrm{diag}(\beta)^{-1} + D, \tag{3.10}$$

where

$$C = \begin{bmatrix} C^{[1]} \\ \vdots \\ C^{[K]} \end{bmatrix}, C^{[k]} = P\mathrm{diag}\left(\sum_{t \in \mathcal{T}(k)} Z^{(t)} \right)$$

and

$$D = \begin{bmatrix} D^{[1]} \\ \vdots \\ D^{[K]} \end{bmatrix}, D^{[k]} = P\mathrm{diag}\left(\sum_{t \in \mathcal{T}(k)} U^{(t)} \right).$$

Also let $b^{[k]} = A_1^{[k]}\beta_1, k = 1, \ldots, K$. Now let us focus on the case ($K = T$) least square estimation. Because $\mathrm{Cov}\left[\mathrm{Z}^{(t)}, U^{(t)}\right] = 0$ and

$$P_0 V_0^{(t)} = P_0 Z_0^{(t)}\mathrm{diag}(\beta_0)^{-1} + P_0 U_0^{(t)} \text{ and } P_1 V_1^{(t)} = P_1 Z_1^{(t)} + P_1 U_1^{(t)} = P_0 Z_0^{(t)} + P_0 U_0^{(t)}, \tag{3.11}$$

we obtain the formulas for the following statistics among T observations

$$E_T\left[A_0^{(t)^\top} W^{(t)} b^{(t)}\right] = \mathrm{diag}(\beta)^{-1} E_T\left[C_0^{(t)^\top} W^{(t)} C_0^{(t)}\right], \quad (3.12)$$
$$E_T\left[A_0^{(t)^\top} W^{(t)} A_0^{(t)}\right] = \mathrm{diag}(\beta)^{-2}\left(E_T\left[C_0^{(t)^\top} W^{(t)} C_0^{(t)}\right] + E_T\left[D_0^{(t)^\top} W^{(t)} D_0^{(t)}\right]\right),$$

where E_T is the expectation across all time intervals when $T \to \infty$. Therefore, by Slutsky's theorem, when the sample size increases infinitely, the least square estimate of β_0 should converge almost surely to a vector that is distinct from β_0 such as

$$\hat{\beta}_{LS} \xrightarrow{p} \beta\left(E_T\left[C_0^{(t)^\top} W^{(t)} C_0^{(t)}\right] + E_T\left[D_0^{(t)^\top} W^{(t)} D_0^{(t)}\right]\right)^{-1} E_T\left[C_0^{(t)^\top} W^{(t)} C_0^{(t)}\right] \neq \beta_0. \quad (3.13)$$

The exact correction approach for least square estimates requires parameters for both systematic and random errors. Let $\sigma_0 = M_0\sigma$. Since the second moments of the error term on interval t can be expressed as

$$E\left[P_0 U_0^{(t)}\left(P_0 U_0^{(t)}\right)^\top \middle| Z^{(t)}\right] = P_0 \mathrm{diag}(\mathrm{Z}_0^{(t)})\mathrm{diag}(\sigma_0)^2 P_0^\top, \quad (3.14)$$

the extra term in (3.13) is the average of those moments across all time intervals,

$$E_T\left[D_0^{(t)^\top} W^{(t)} D_0^{(t)}\right] = \frac{1}{T} P_0 \mathrm{diag}\left(\sum_{t=1}^{T} Z_0^{(t)}\right)\mathrm{diag}(\sigma_0)^2\, P_0^\top \xleftarrow{a.s.} \frac{1}{T} P_0 \mathrm{diag}\left(\sum_{t=1}^{T} V_0^{(t)} \circ \beta_0\right)\mathrm{diag}(\sigma_0)^2 P_0^\top. \quad (3.15)$$

Therefore, the corrected least square estimate is given by

$$\hat{\beta}_{\mathrm{CRLS}} = \left(A_0^\top W A_0 - P_0 \mathrm{diag}\left(\sum_{t=1}^{T} V_0^{(t)} \circ \beta_0\right)\mathrm{diag}(\sigma_0)^2 P_0^\top\right)^{-1} A_0^\top W b. \quad (3.16)$$

A comprehensive overview on linear error-in-variable models and a variety of remedial means can be found in Gillard (2010). However, most of them require either additional data (instrumental variable) or information (maximum likelihood). Among all practical approaches that do not rely on knowledge about β or σ, the most competing one for our specific problem is total least square (TLS). It typically involves a separate numerical procedure (See Golub and Van Loan, 1980 for details) other than using a simple quadratic minimization problem. Oppositely, our novel approach is naturally embedded in GMM estimation methods and can be dealt within the same optimization framework. In fact, it is simply achieved by aggregating data to maximize the variation of group means.

On one hand, when data is aggregated by K groups, the second order statistics E_K of group means converges to that of pure traffic counts E_T

$$E_K\left[C_0^{[k]^\top} W^{[k]} C_0^{[k]}\right] \xrightarrow{a.s.} E_T\left[C_0^{(t)^\top} W^{(t)} C_0^{(t)}\right]. \quad (3.17)$$

Here $E_k[\cdot]$ is the expectation across all K groups as $K \to \infty$ given the growth of K is slower than T. On the other hand,

$$E_K\left[D_0^{[k]^\top} W^{[k]} D_0^{[k]}\right] \xrightarrow{a.s.} E_T\left[D_0^{(t)^\top} W^{(t)} D_0^{(t)}\right]. \tag{3.18}$$

The estimate with aggregated data is approaching to β_0 as K/T diminishes, since

$$\hat{\beta}_{\text{GMM}} \xrightarrow{p} \beta\left(E_T\left[C_0^{(t)^\top} W^{(t)} C_0^{(t)}\right] + \frac{K}{T} E_T\left[D_0^{(t)^\top} W^{(t)} D_0^{(t)}\right]\right)^{-1} E_T\left[C_0^{(t)^\top} W^{(t)} C_0^{(t)}\right]. \tag{3.19}$$

4. Statistical Inference

As the previous section focuses merely on attaining a unique and consistent estimate of systematic error parameters, we now delve into the way to improve estimation efficiency against sample size, conduct hypothesis tests to infer biased sensors and ultimately reconstruct traffic flows using estimated parameters. For those purposes, we first have to estimate nuisance parameter σ for random measurement error scale. Based on that, we construct optimal weighting matrix and derive the large sample properties for β estimator. Finally, a maximum likelihood estimation of $Z^{(t)}$ in each interval t is proposed together with an updating algorithm summarizing the efficient estimation of both β and σ.

4.1. Estimating Random Errors

The estimation of σ resides in the same GMM framework as for β except that the second order conditions are in use instead. According to the structural equations (2.10),

$$E\left[\left(\sum_{a\in\mathcal{A}(i)} p_{ia}\beta_a V_a^{(t)}\right)\left(\sum_{a\in\mathcal{A}(j)} p_{ja}\beta_a V_a^{(t)}\right)\middle|Z^{(t)}\right] = E\left[\left(\sum_{a\in\mathcal{A}(i)} p_{ia}\beta_a U_a^{(t)}\right)\left(\sum_{a\in\mathcal{A}(j)} p_{ja}\beta_a U_a^{(t)}\right)\middle|Z^{(t)}\right]. \tag{4.1}$$

After the terms on the right hand side is rearranged, we obtain

$$\sum_{a\in\mathcal{A}(i)}\sum_{a'\in\mathcal{A}(j)} p_{ia}p_{ja'}\beta_a\beta_{a'} E\left[U_a^{(t)}U_{a'}^{(t)}\middle|Z^{(t)}\right] = \sum_{a'\in\mathcal{A}(i)\cap\mathcal{A}(j)} p_{ia}p_{ja}\beta_a^2 \operatorname{Var}\left[U_a^{(t)}\middle|Z^{(t)}\right], \tag{4.2}$$

because of the mutual independence of random error generation process in each sensor. We substitute the formula for $Var\left[U_a^{(t)}\middle|Z^{(t)}\right] = \sigma_a^2 Z_a^{(t)}$ and express the second moment condition in a matrix form

$$E\left[P\operatorname{diag}(\beta\circ V^{(t)})^2 P^\top\middle|Z^{(t)}\right] = P\operatorname{diag}(\beta\circ\sigma)^2\operatorname{diag}(Z^{(t)})P^\top. \tag{4.3}$$

Because $E[V^{(t)}\circ\beta|Z^{(t)}] = Z^{(t)}$, the moment condition becomes

$$E\left[P\operatorname{diag}(\beta)\operatorname{diag}\left(V^{(t)}\right)^2 P^\top - P\operatorname{diag}(\beta\circ\sigma)^2\operatorname{diag}(\beta\circ V^{(t)})P^\top\right] = 0. \tag{4.4}$$

We avoid requiring unknown $Z^{(t)}$ in this combined condition. In the estimation of σ with K groups, we would like all elements in the following vector function to simultaneously become zero,

$$h(\sigma,\beta)=\begin{bmatrix}\mathrm{vech}\left(P diag(\beta)\left(\frac{1}{|\mathcal{T}(1)|}\sum_{t\in\mathcal{T}(1)}\mathrm{diag}\left(V^{(t)}\right)^2\right)P^\top-P\mathrm{diag}(\beta\circ\sigma)^2\mathrm{diag}(\beta\circ\frac{1}{|\mathcal{T}(1)|}\sum_{t\in\mathcal{T}(1)}V^{(t)})P^\top\right)\\ \vdots\\ \mathrm{vech}\left(P\mathrm{diag}(\beta)\left(\frac{1}{|\mathcal{T}(K)|}\sum_{t\in\mathcal{T}(K)}\mathrm{diag}\left(V^{(t)}\right)^2\right)P^\top-P\mathrm{diag}(\beta\circ\sigma)^2\mathrm{diag}(\beta\circ\frac{1}{|\mathcal{T}(K)|}\sum_{t\in\mathcal{T}(K)}V^{(t)})P^\top\right)\end{bmatrix}. \tag{4.5}$$

Here $\mathrm{vech}(\cdot)$, the half vectorization of an $m\times m$ square matrix, returns an $\frac{m(m+1)}{2}\times 1$ vector containing all the elements of the lower triangular portion. The nuisance parameter σ becomes the only unknown in this relation after β is estimated. It is clear that $h(\sigma;\beta)$ is linear in σ. The GMM estimate of σ is found by solving a quadratic problem that minimizes $\left\|h\left(\sigma;\hat{\beta}\right)\right\|_2^2$. Although the total number of entries in h is $K\frac{m(m+1)}{2}$ there are much less valid equations that can be used to identify σ because the coefficients of σ_a^2's in some equations are all zeros. From the relation (4.2), we know both sides of equations are simply zero if nodes i and j are not directly connected by one link. Therefore, the number of valid equations that associate two different nodes is equal to the number of non-leaf links. Because the number of leaf links is equivalent to that of nodes with degree one $(\mathcal{N}-m)$, we should have $n-(|\mathcal{N}|-m)$ equations for non-trivial covariance of node relation. We know that m equations are given for the variance of nodal relation. In sum, after removing all useless equations, h would have $K(n-|\mathcal{N}|+2m)$ entries.

4.2. Efficient Estimator and Infer Sensor Health

Next we develop a general formula for the optimal choice of weighting matrix W. For any two elements i and j in $g(\beta)$, if i and j are two nodes that belong to the same aggregation group k, then

$$E\left[g_i(\beta)g_j(\beta)\right]=\frac{1}{|\mathcal{T}(k)|^2}\sum_{t\in\mathcal{T}(k)}\sum_{a\in\mathcal{A}(i)\cap\mathcal{A}(j)}p_{ia}p_{ja}\sigma_a^2\beta_a^3V_a^{(t)}, \tag{4.6}$$

by assuming random error $U^{(t)}$ from different intervals are independent. If i and j are not from the same group, $E\left[g_i(\beta)g_j(\beta)\right]$ is simply zero. The corresponding matrix form of all elements covariance Ω is then made of K blocks:

$$\Omega=\begin{vmatrix}\frac{1}{|\mathcal{T}(1)|^2}\sum_{t\in\mathcal{T}(1)}P\mathrm{diag}(V^{(t)})\mathrm{diag}(\sigma)^2\mathrm{diag}(\beta)^3P^\top & \dots & 0\\ \vdots & \ddots & \dots\\ 0 & \dots & \frac{1}{|\mathcal{T}(K)|^2}\sum_{t\in\mathcal{T}(K)}P\mathrm{diag}(V^{(t)})\mathrm{diag}(\sigma)^2\mathrm{diag}(\beta)^3P^\top\end{vmatrix}. \tag{4.7}$$

Thus, in method of moments estimation, the covariance matrix is just one single block, while in GLS, it is a block diagonal matrix of T non-trivial blocks. Knowing β_1 or not does not make any difference to the formula of $\Omega(\beta)$.

One important characteristic of statistical inference of systematic error is to provide evidence for the costly decision of replacing/recalibrating installed sensors. For a particular sensor a, the null hypothesis, H_0, is that the sensor *a* does not have any systematic measurement error, i.e., $\beta_a = 1$. The alternative hypothesis is that $\beta_a \neq 1$. Thus, a marginal two-sided location test should come in handy. First of all, according to the GMM theory, the estimator β_0 converges in distribution as T arises infinitely

$$\sqrt{K}(\hat{\beta}_0 - \beta_0) \xrightarrow{d} (0, (A_0^\top W A_0)^{-1} A_0^\top W \Omega W A_0 (A_0^\top W A_0)^{-1}). \tag{4.8}$$

In the case that $W \xrightarrow{p} \Omega^{-1}$, the formula collapses to a simpler expression,

$$\sqrt{K}(\hat{\beta} - \beta) \xrightarrow{d} (0, \Sigma), \text{ where } \Sigma = (A_0^\top \Omega^{-1} A_0)^{-1}. \tag{4.9}$$

and the variance-covariance matrix of estimator using Ω^{-1} is proven to be the smallest among all results using any possible positive definite matrices W. Therefore, with the most efficient estimator, the standard error of β estimates is denoted by $\hat{\beta}_a$ is then

$$\text{se}(\hat{\beta}_a) = \sqrt{\Sigma_{aa}/K}, \tag{4.10}$$

where Σ_{aa} is the diagonal entry of Σ corresponding to β_a. Hence, the test statistics is simply

$$\frac{\hat{\beta}_a - 1}{\text{se}(\hat{\beta}_a)}. \tag{4.11}$$

Then it is compared with the critical values of a standard normal distribution (asymptotic) at any chosen level of significance to infer whether β_a is statistically significantly different from one.

4.3. Algorithm: Estimation and Recovery

The algorithm to implement the proposed estimation method for β_0. is outlined as follows

Step 0. Split the observations into K groups; set initial weighting matrix to be $W = I$.

Step 1. Find $\hat{\beta}_0^{\text{old}}$ using (3.9).

Step 2. Find $\hat{\sigma} = \arg\min_{\sigma \geq 0} ||h(\sigma; \hat{\beta}_0^{\text{old}})||_2^2$.

Step 3. Construct Ω using $\hat{\beta}_0^{\text{old}}$ and $\hat{\sigma}$; update $W = \Omega^{-1}$.

Step 4. Find $\hat{\beta}_0^{\text{new}}$ using (3.9).

Step 5. Check if $||\hat{\beta}_0^{\text{new}} - \hat{\beta}_0^{\text{old}})|| \leq tol$. If not, let $\hat{\beta}_0^{\text{new}} = \hat{\beta}_0^{\text{old}}$ and go to step 3. Otherwise, $\hat{\beta}_0 = \hat{\beta}_0^{\text{old}}$ and stop.

It is noteworthy that in Step 1 and 4, we basically ignore the positiveness constraints and apply the analytical formula directly, because the estimate is not supposed to get close to those boundary unless sensors are completely malfunctioning.

This algorithm can be casted as a typical iteratively reweighed unconstrained least square. The convergence of such algorithms has been proved and discussed in depth in classical work, such as Osborne (1985) and state-of-art research, for example Daubechies et al. (2010).

We now apply maximum likelihood estimation to find the remedied flows based on observed counts. The likelihood of observations on interval t conditioning on the true traffic flows is given by

$$\mathcal{L}\left(Z^{(t)};\mu,\sigma\middle|V^{(t)}\right) = \mathbb{I}\left(PZ^{(t)}=0\right)\prod_{a\in\mathcal{A}}\frac{1}{\sqrt{2\pi\sigma_a^2 Z_a^{(t)}}}\exp\left(-\frac{\left(V_a^{(t)}-(1+\mu_a)Z_a^{(t)}\right)^2}{\sigma_a^2 Z_a^{(t)}}\right). \tag{4.12}$$

Then traffic counts can be corrected by maximizing the loglikelihood which is expressed as

$$\ell\left(Z^{(t)};\mu,\sigma\middle|V^{(t)}\right) = \mathbb{I}\left(PZ^{(t)}=0\right)\left(-\frac{n}{2}\ln 2\pi - \sum_{a\in\mathcal{A}}\left(\ln\sigma_a + \frac{1}{2}\ln Z_a^{(t)} + \frac{\left(V_a^{(t)}-(1+\mu_a)Z_a^{(t)}\right)^2}{\sigma_a^2 Z_a^{(t)}}\right)\right). \tag{4.13}$$

Dropping the constant terms, we have the following nonlinear constrained optimization problem to tackle with,

$$\hat{Z}_{\text{MLE}}^{(t)} = \arg\min_{Z\geq 0}\sum_{a\in\mathcal{A}}\frac{1}{2}\ln Z_a + \frac{\left(V_a^{(t)}-\left(1+\hat{\mu}_a\right)Z_a^{(t)}\right)^2}{\hat{\sigma}_a^2 Z_a^{(t)}} \quad \text{s.t. } PZ=0, \tag{4.14}$$

Although the problem is highly nonlinear and appears hard to solve in nature, fortunately it is for one time interval and involves only n variables. The number of sensors in a large size regional transportation network rarely exceeds a thousand. It is not considered as a computationally challenging task given the current development of optimization techniques. We have employed a gradient descent based algorithm to solve the problem in all the numerical experiments.

For real-time online applications using streaming data, an alternative least square based Z estimation formulation is also stated as,

$$\hat{Z}_{\mathrm{LS}}^{(t)} = \arg\min_{Z \geq 0} \sum_{a \in \mathcal{A}} \left(V_a^{(t)} - \left(1 + \hat{\mu}_a\right) Z_a^{(t)}\right)^2 \quad \text{s.t. } PZ = 0, \tag{4.15}$$

which can be handled by simple least square solvers.

5. Numerical Examples

5.1. An Illustrative Example using a Freeway Corridor

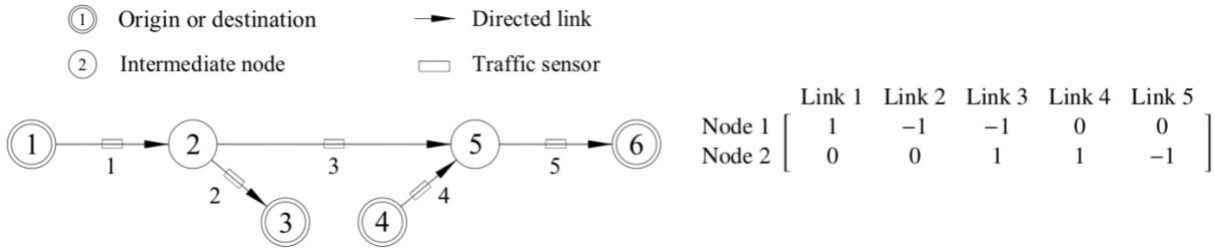


Figure 2. Network 1: A Segment of Freeway Corridor

The purview of the first example is to demonstrate the process of utilizing the proposed method to identify malfunctioning sensors and correct erroneous data. In lieu of the display convenience of estimation results, Test Network 1 is a freeway segment consisting of five directed links, including one on-ramp, one off-ramp and three mainline links, as shown in Figure 2. Out of six nodes in this network graph, four of them are origin or destination nodes, where traffic flows enter or exit; and the other two are intermediate ones, where flow balance law is supposed to hold, so $m = 2$. Each link is equipped with a traffic loop detector that counts all passing vehicles. Therefore, P is a 2×5 matrix as illustrated next to the network graph.

Table 1. True Parameters of Network

	Link 1	Link 2	Link 3	Link 4	Link 5
μ	.150	-.150	-.350	.000*	-.200
β	.869	1.176	1.538	1.000*	1.250
σ	.300	.200	.500	.500*	.300

The true systematic and random error parameters for five sensors are given in Table 1. The sensor on link 4 is recently calibrated so that both μ_4 and σ_4 are known. We now simulate 100 samples of traffic data to conduct a Monte Carlo experiment. Each sample consists of 365×24 hourly traffic counts. Origin-destination demand in each hour is a normal variable with parameters specified only for that interval of days. Means for weekends and holidays are discounted on the basis of that of weekdays. Figure 3 presents the true traffic flows of mainline corridor and ramp respectively in a sample year. Then the observed flows for each hour are

generated with true hourly volumes and the previously stated sensor measurement error model with the listed parameters.

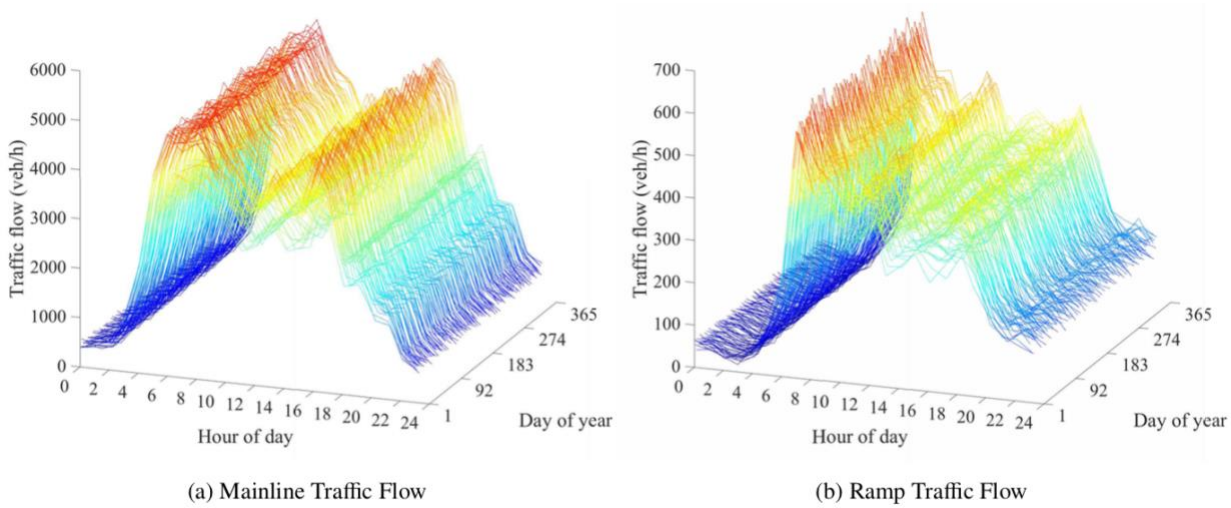


Figure 3. Daily Traffic Profile of Network 1

In data aggregation, the total 8760 hourly observations of traffic flow were grouped by the hours of a day, so $K = 24$. In Table 2, we present the mean and standard deviation in parentheses of estimated β using the proposed GMM method in comparison with GLS, a straightforward approach but with inconsistent estimator, and TLS, a generic means to handle error-in-variable model issue. All the estimates are found without the knowledge of the random error ratios. The GLS estimates that are corrected using true σ, crGLS, are also shown at last as a benchmark to assess estimation performances. Evidenced by the sample mean and sample standard errors, the GMM estimates without knowing σ's are much more accurate and precise compared to ungrouped GLS and TLS. Its precision is only slightly worse than that of crGLS since the latter uses true values of parameters σ.

Table 2. Means (and Standard Deviations) of β Estimates

	μ_1	μ_2	μ_3	μ_5
True Value	.150	-.150	-.350	-.200
GMM	.151(.005)	-.149(.003)	-.349(.003)	-.199(.300)
GLS	.274(.033)	-.100(.018)	-.276(.019)	-.118(.021)
TLS	.180(.013)	-.159(.012)	-.331(.007)	-.179(.008)
crGLS	.151(.003)	-.150(.002)	-.350(.002)	-.199(.002)

Concerning the GMM estimation results with different sample sizes, we conduct the same experiment using data that span one month, three months, half year and one year,

respectively. In Figure 4, the solid lines denote the finite sample distribution of β estimates normally fitted using repeated simulation experiment results, while the dashed lines are asymptotic distribution constructed using all true parameters for the corresponding sample sizes. With only one month data, the peak (mean) of estimate differs from the true parameter due to a small sample size. As the number of observations grows, expectedly those two distributions tend to collide. The shrinkage rate of estimates' standard deviations is about $\sqrt{T}$.

In order to demonstrate the inferential procedure after obtaining estimates of β, we pick the fiftieth sample among a total of one hundred based on the order of their β estimates accuracy, measured by the average of relative mean squared errors. First, the generalized moment dispersion matrix Ω is computed using estimated σ as shown in Table 3 and illustrated by showing its first two blocks that correspond to the first two groups $k = 1, 2$. Then the estimate variance-covariance matrix Σ is found and used to find standard errors. Note they are slightly different than those of asymptotic distribution shown in Figure 4, because they are computed using true parameters instead of one particular sample estimate. Finally, the absolute values of test statistics are much larger than critical values at a significance level of .01. Therefore, it is statistically significant to reject the null hypotheses that those sensors do not have systematic errors.

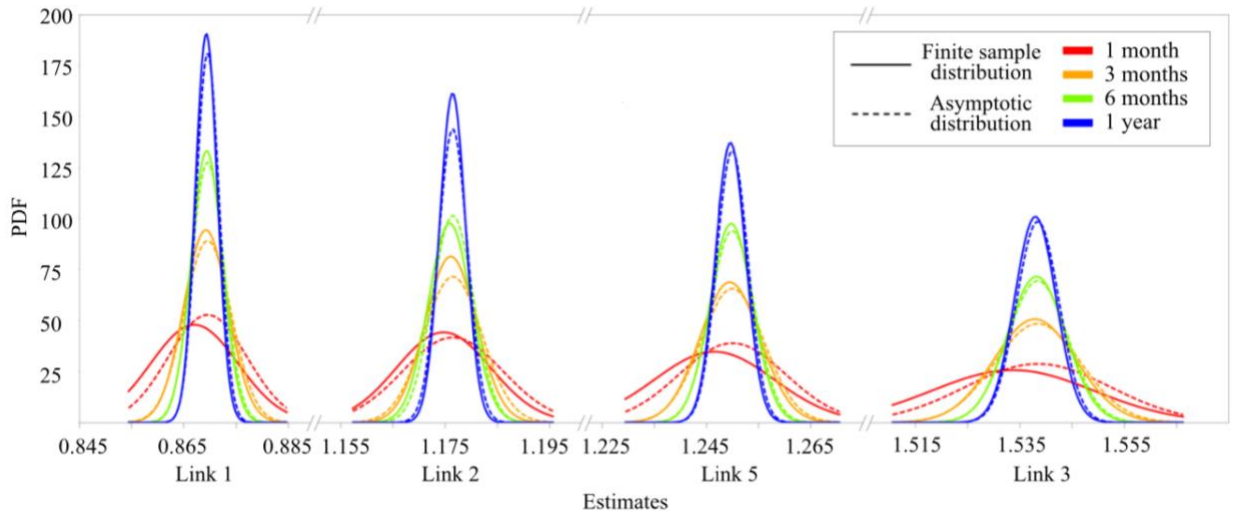


Figure 4. Distribution of β Estimates over Sample Size T

Group, Node	1,1	1,2	2,1	2,2	⋯
1, 1	.328	−.251	0	0	⋯
1, 2	−.251	.404	0	0	⋯
2, 1	0	0	.496	−.399	0
2, 2	0	0	−.399	.554	0
⋮	⋮	⋮	0	0	⋱

Illustration of Ω Matrix of Network 1

Table 3. Statistical Inference Results of Network 1

	Link 1	Link 2	Link 3	Link 5
Estimate β	. 869	1.173	1.538	1.250
Estimate σ	. 295	.218	.494	.299
Standard Error	.00199	.00313	. 00368	.00273
Test Statistic	−65.87	55.33	146.31	91.58

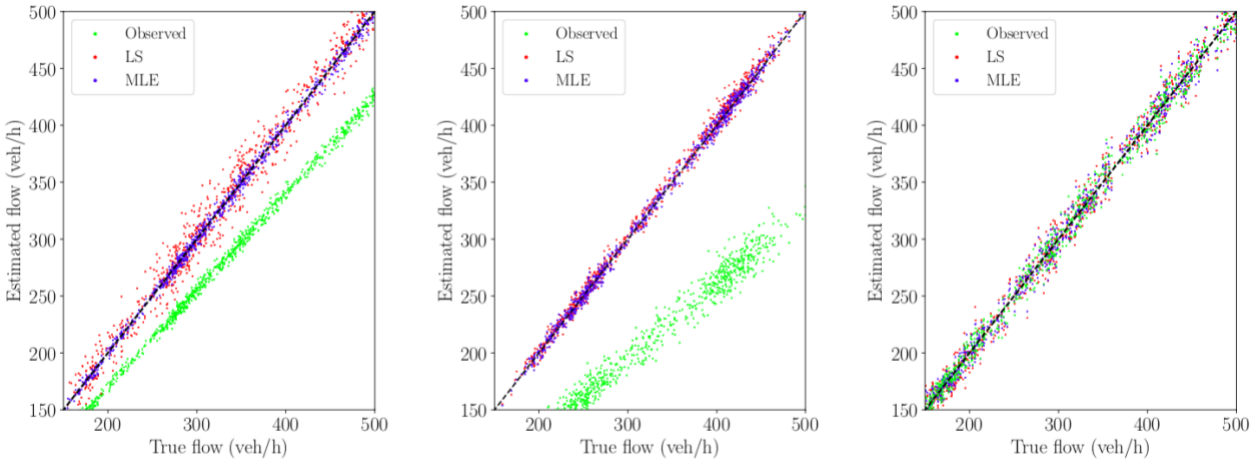
* The critical value of a two-sided Wald test at the a significance level of .01 is ±2.58.

Next we compare the observed hourly counts, corrected flows using LS or MLE in Figure 5 against true traffic flows. We are particularly interested in comparing the performances of correcting flows on mainline versus ramp sensors as well as uncalibrated malfunctioning versus calibrated sensors. For the sake of clarity, we randomly select ten percent of sample points to show on the scatter plots. From all three graphs, it is manifest that MLE corrected flows with relatively accurate estimates of σ are more reliable than those of LS. Although the existing random errors of the mainline sensor are higher than that of ramp sensors as shown by green dots, the mean squared errors of MLE estimates for mainline sensor is actually smaller as shown by blue dots. In short, the correction results on the mainline sensor appear better than that on ramp sensors in this example. This is mainly due to the dominant magnitude of mainline flows. For calibrated ramp sensor 4, the MLE corrected flows have a similar error scale with the original observed counts. It indicates that MLE approach does not have a significant improvement on eliminating random error for this particular sensor. This could be ascribed to the fact that, unlike mainline link that connects two intermediate nodes, ramp link is only associated with one nodal balance equation.

5.2. Numerical Tests using a General Network

We performed a series of numerical tests using a general network in Figure 6. It consists of 6 origin/destination (shown with double circles), 19 intermediate nodes and 50 directed links. Instead of having a fixed structure like in unidirectional freeway corridor, this general network

is bidirectional with asymmetric flows. One hundred samples of road traffic are simulated by randomizing flows on 50 different paths jointly and available for download.[2]



(a) Sensor 2: Ramp and Uncalibrated (b) Sensor 3: Mainline and Uncalibrated (c) Sensor 4: Ramp and Calibrated

Figure 5. Corrected Traffic Flow of Network 1

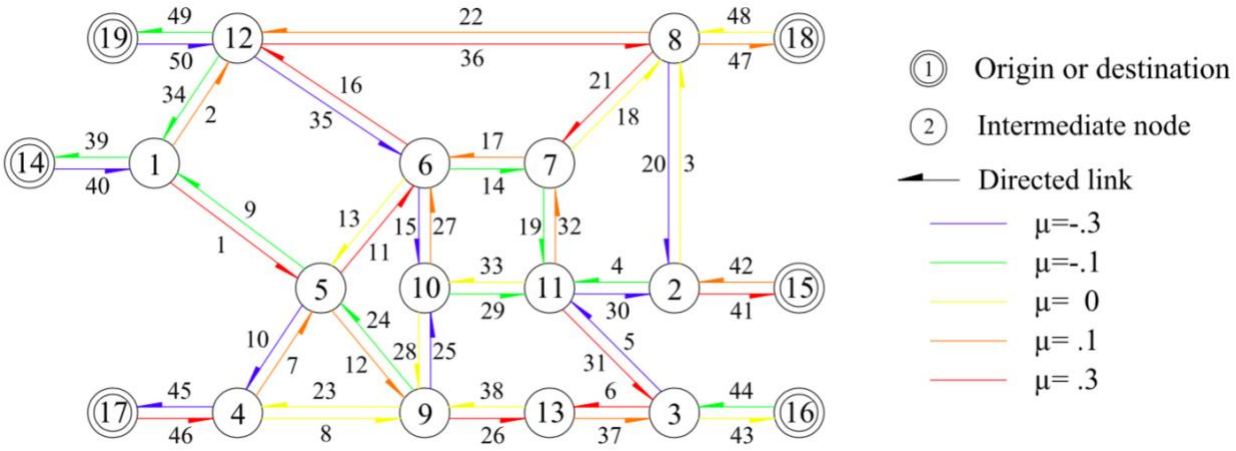


Figure 6. Modified Nguyen-Dupuis Network

Sensors are deployed on all the links. In order to analyze method's performance for sensors of different levels of systematic measurement errors and to mitigate the effect caused by link location in the network (e.g., connecting to one or two intermediate nodes), we divide sensors into five groups: (1) $\mu = -.3$ (severely under-counting), (2) $\mu = -.1$ (mildly under-counting), (3) $\mu = 0$ (accurately counting), (4) $\mu = .1$ (mildly over-counting), and (5) $\mu = .3$ (severely over-counting). Each group is assigned with ten sensors and illustrated using different colors in

[2] https://github.com/yudiaspen/sensor-bias-estimation/nguyen-depuis

Figure 6. The calibrated sensors are on link 3, 8 and 13. The random error parameter σ are given in Table 4.

There are three types of measures used to assess the estimation quality. From each data sample, estimation error of μ for sensor a is calculated as,

$$\text{estimation error}_a^{\text{smpl}} = \hat{\mu}_a^{smpl} - \mu, a \in \mathcal{A}, \text{smpl} = 1, \dots, S, \tag{5.1}$$

Table 4. Random Error Parameters

Sensor Group 1	1	6	11	16	21	26	31	36	41	46
True σ	.20	.05	.20	.30	.35	.40	.05	.15	.40	.20
Sensor Group 2	2	7	12	17	22	27	32	37	42	47
True σ	.30	.10	.30	.20	.40	.05	.20	.30	.30	.40
Sensor Group 3	3	8	13	18	23	28	33	38	43	48
True σ	.05	.15	.10	.25	.15	.05	.40	.35	.15	.15
Sensor Group 4	4	9	14	19	24	29	34	39	44	49
True σ	.15	.20	.40	.10	.30	.10	.25	.05	.35	.15
Sensor Group 5	5	10	15	20	25	30	35	40	45	50
True σ	.10	.25	.05	.10	.40	.40	.30	.35	.05	.10

where S is the number of samples. Average estimation bias and standard error among all sensors are expressed as

$$\begin{aligned} \text{average bias} &= \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\left|\frac{1}{S}\sum_{\text{smpl}=1}^{S}\hat{\mu}_a^{\text{smpl}} - \mu,\right|, \\ \text{average standard error} &= \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\sqrt{\frac{1}{S-1}\sum_{\text{smpl}=1}^{S}\left(\hat{\mu}_a^{\text{smpl}} - \frac{1}{S}\sum_{\text{smpl}=1}^{S}\hat{\mu}_a^{\text{smpl}} - \mu\right)^2}. \end{aligned} \tag{5.2}$$

5.2.1. Aggregation Group Size

The first experiment compares those quality measures using different group sizes for aggregating observations. In the last example, the aggregation strategy is based on the hour of day. This matches the data generation setup that each hourly traffic is sampled from a different population. For the current network, without exploiting this feature, we adopt K-means clustering technique, a more generic way to partition observations. The clustering approach adopted in this paper is based on Euclidean distance between V's. Figure 7 demonstrates the results of K-means clustering for $K = 24$. The color of a cell represents the probability of each hour (column) observations fall in a cluster (row): a warmer color indicates a higher probability

and vice versa. Most of clusters tend to scatter over several time periods during daytime, while other clusters concentrate on several night hours, since traffic flows at that time are considerably lower than that in daytime.

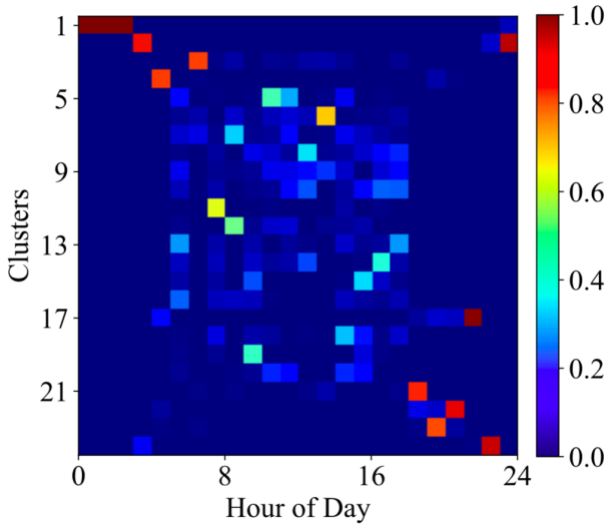


Figure 7. The Distribution of Clusters: $K = 24$

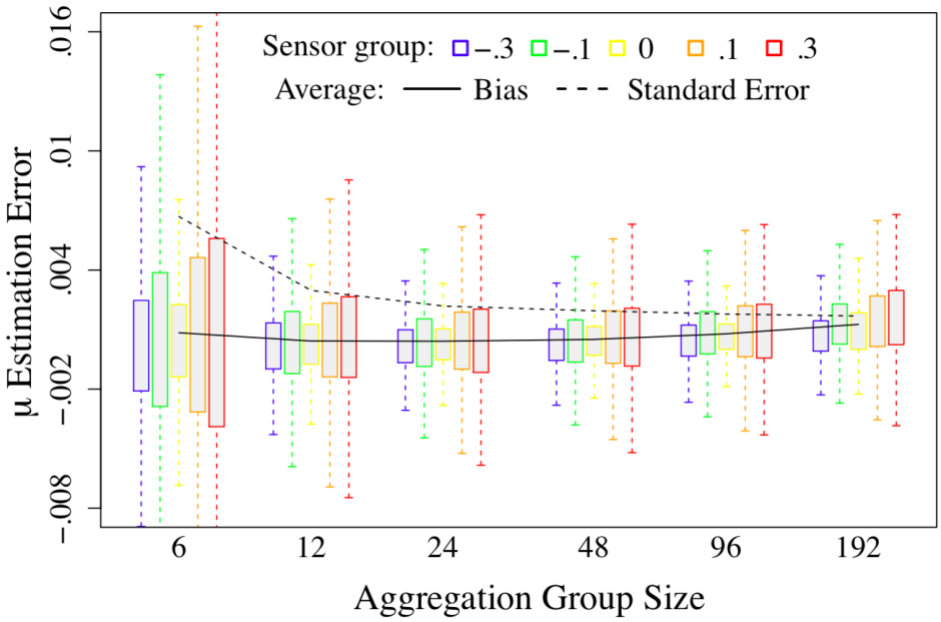


Figure 8. The Effect of Aggregation Group Size

We start from $K = 6$ in Figure 8, since estimation problem with fewer group does not allow a numerically stable solution due to lack of information. On one hand, as more independent equations are supplied, the average standard error drops rapidly as the group size doubles from 6 to 12. The effect of incorporating new equations gradually vanishes after K is greater than 12.

On the other hand, having smaller groups enlarges the random error in $C^{[k]}$ and leads to a slight increase in bias from $K = 24$ to $K = 192$. The box plots of individual estimation errors echo the observations made on the average measures. They shrink in size and lean towards the positive direction as K rises.

5.2.2. Random Error Scale

In this experiment, we are interested in examining the effect of random error scale on the estimates of systematic error ratio. The magnitude of random error is varied by multiplying the original value of σ_a given in Table 4 with a scalar, then

$$\sigma_a^{\text{test}} = \Delta\sigma \cdot \sigma_a. \tag{5.3}$$

We consider six scenarios $\Delta\sigma$ = 0, .4, .8, 1.2, 1.6, and 2. Instead of directly showing $\Delta\sigma$ on the horizontal axis of the plot in Figure 9, we employ a more straightforward quantity to present the scale of random error, which is computed using

$$\text{Relative Random Error} = \sum_{\text{smpl}=1}^{S} \sum_{a\in\mathcal{A}} \sum_{t=1}^{T} \left| U_a^{(t),\text{smpl}} \right| \Big/ \sum_{\text{smpl}=1}^{S} \sum_{a\in\mathcal{A}} \sum_{t=1}^{T} Z_a^{(t),\text{smpl}}. \tag{5.4}$$

and linearly related to $\Delta\sigma$.

In Figure 9, average standard error has approximately a linear growth as the relative random error rises, because average standard error$^2 \propto \text{tr}(\Sigma) \propto \text{tr}(\Omega) \propto \Delta\sigma^2$. The operator tr() denotes the trace of a matrix. The curve of average bias is much more convex: the "attenuation bias" caused by random measurement error in $C^{[k]}$ becomes a significant problem after relative random error exceeds 8%. It is known that in this case β is underestimated, so $\mu = 1/\beta - 1$ is overestimated as illustrated by the boxes of estimation errors grouped by sensor μ's.

5.2.3. Calibrated Sensors

In this paper, having a calibrated sensor indicates that there is no systematic errors and known random error scale. It is manifest that the more calibrated sensors are there before estimation, the better the estimates should be. Therefore, we concern about the dependence of estimate quality on the number of calibrated sensor. On one hand, the method should be able to find a more reliable estimate of systematic error as more sensors are calibrated. On the other hand, it should reach a satisfactory level of performance without requiring too many calibrated sensors, which would otherwise compromise the purpose of estimation. From Figure 10, we can tell that the proposed method excels on both aspects: the bias and standard error decreases as the number of calibrated sensors increases; despite a substantial improvement from one sensor to two, the marginal gain from more calibrated sensors diminishes as the number of calibrated sensors goes beyond two.

5.2.4. Unbalanced Flows

It is not very likely that hourly link flows that connect to the same physical road traffic network node has significant imbalance since the effect of shockwaves caused by queuing and dequeuing between links are probably averaged out over such a long period of time. However, since our estimation model assumes perfect balance, it is still meaningful to examine the robustness of the proposed approach against the different levels of flow balance law violation. The nodal relation is now expressed as in (2.7) and the true flow is set to be

$$Z_a^{(t)} = Z_{B,a}^{(t)} + \Delta Z \sqrt{Z_{B,a}^{(t)}} \mathcal{Z}, a \in \mathcal{A}, \tag{5.5}$$

Where $\bar{Z}_{B,a}$ is the adjusted flow on link a that obey balance law precisely, ΔZ is a disturbance parameter, and $\mathcal{Z}$ is a standard normally distributed scalar.

Figure 11 demonstrates that as ΔZ increases from 0 to 1 and flow imbalance ratio rises to 10% correspondingly, estimation error of μ only gradually increases to .1 on average. This is because upon long observation interval (one hour), the aggregation of data further mitigates the impact of flow imbalance. The use of K mean clusters also helps obviating potential structural imbalance caused by queuing and dequeuing at certain link in a particular hour of day.

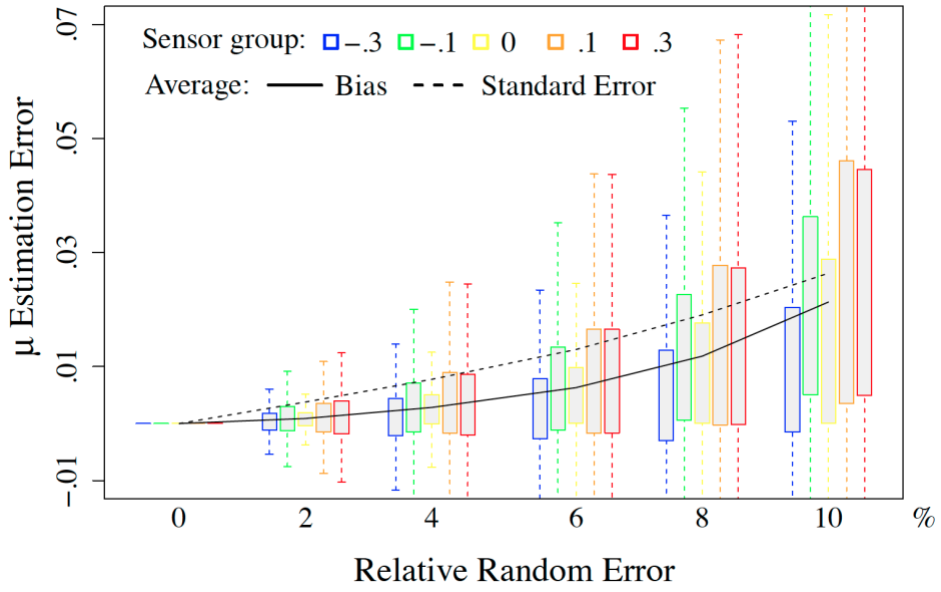


Figure 9. The Effect of Random Error Scale

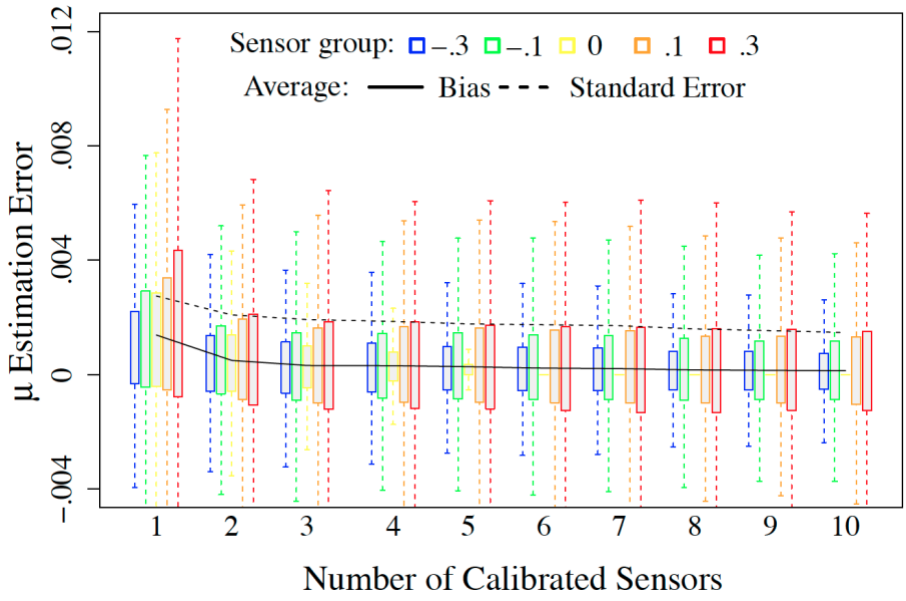


Figure 10. The Effect of Calibrated Sensors

The situation is different for correcting flows. In Figure 12, we plot the scatter points for correction quality of a data sample in six scenarios ($\Delta Z = 0, .2, \dots, 1$) against the difference between true and balanced flow, represented by mean squared error (MSE), that is

$$MSE(Z, Z_B) = \frac{1}{|\mathcal{A}|}\frac{1}{T}\sum_{a\in\mathcal{A}}\sum_{t=1}^{T}\left(Z_a^{(t)}, Z_{B,a}^{(t)}\right)^2, \tag{5.6}$$

We use $\hat{Z}_B$ to denote the corrected flow when there is no flow imbalance. The shaded areas cover 95% points of their own colors, respectively. The dotted lines represent the mean values of the MSE. Since corrected flows using imbalanced flow $\hat{Z}$ still follow balance law strictly, the correction error (blue) between that and unbalanced true flow is increasing and reaches around 200 when $\Delta Z = 1$. But it is worth noting that such error is significantly lower than $MSE(Z, Z_B)$. In fact, among balanced flows, $\hat{Z}$ is a better prediction of Z compared to original balanced flows Z_B except in the case where flow balance is perfectly held.

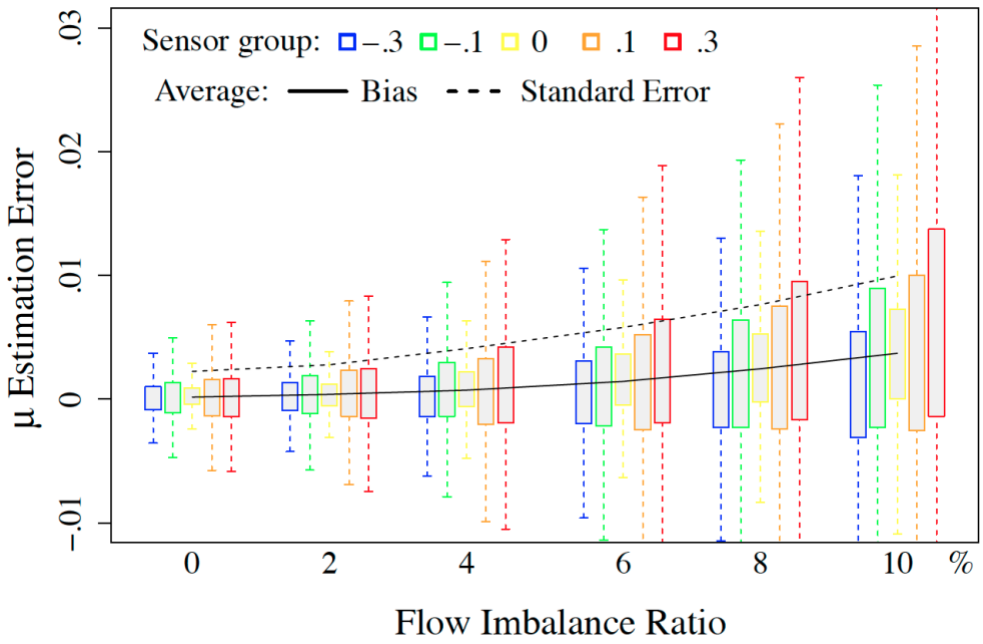


Figure 11. The Effect of Flow Imbalance on Estimating μ

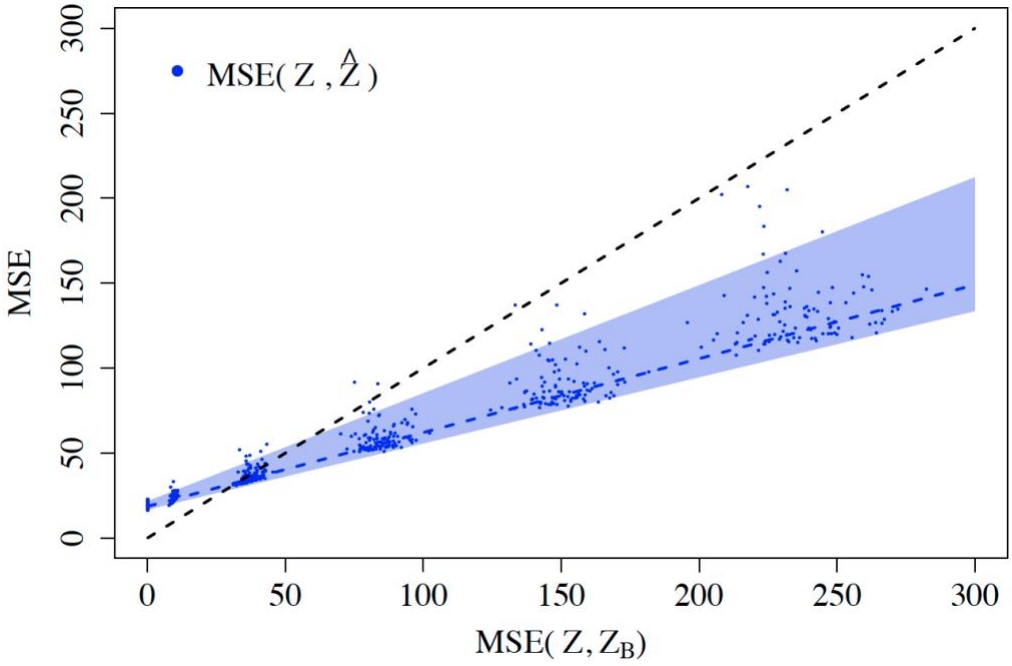


Figure 12. The Effect of Flow Imbalance on Correcting Z

5.2.5. Stochastic Error Generation

As explained in 2.1. Measurement Errors, the relative strong assumption we impose to obtain a linear measurement error model is that the error generation mechanism is consistent in all

time intervals and does not depend on the flow variables. We now examine how our model performs when such constant error generation assumption does not hold true.

In this test, we suppose that systematic error μ to be a random vector that is normally distributed

$$\mu_a = \bar{\mu}_a(1 + \Delta\mu\mathcal{Z}), a \in \mathcal{A}, \tag{5.7}$$

where $\bar{\mu}_a$ is a constant and has the same meaning with μ in the fixed linear model, $\Delta\mu$ is a multiplier that controls the variation of all μ_a's and $\mathcal{Z}$ is a standard normal variable. In Figure 13, the value of $\Delta\mu$ is given from 0 to .3 with a step of .05. The shaded region is the 95% probabilistic range of true systematic error ratios resulted from stochastic error generation, while the monochromatic region is the 95% probabilistic range of estimated systematic error ratios. For those severely malfunctioning sensors, the range of $\hat{\mu}$ is much smaller than μ, indicating a significant effect of variance reduction. For the healthy sensors, μ does not vary and the range of $\hat{\mu}$ is relatively small. For sensors that are mildly functioning, those two regions are approximately the same and the one for $\hat{\mu}$ is slightly tilted up due to estimation bias accounted for this additional randomness. Although error generation is stochastic now and our model is indeed misspecified, as shown in violin plots of Figure 14, the relative correction error of $\mathcal{Z}$ is marginal and only reaches 5% when $\Delta\mu = .3$. In this figure, each diagram is made of $T \times |\mathcal{A}|$= 438000 points from a single data sample. Colored shape and boxes insides cover 95% and 50% of them respectively.

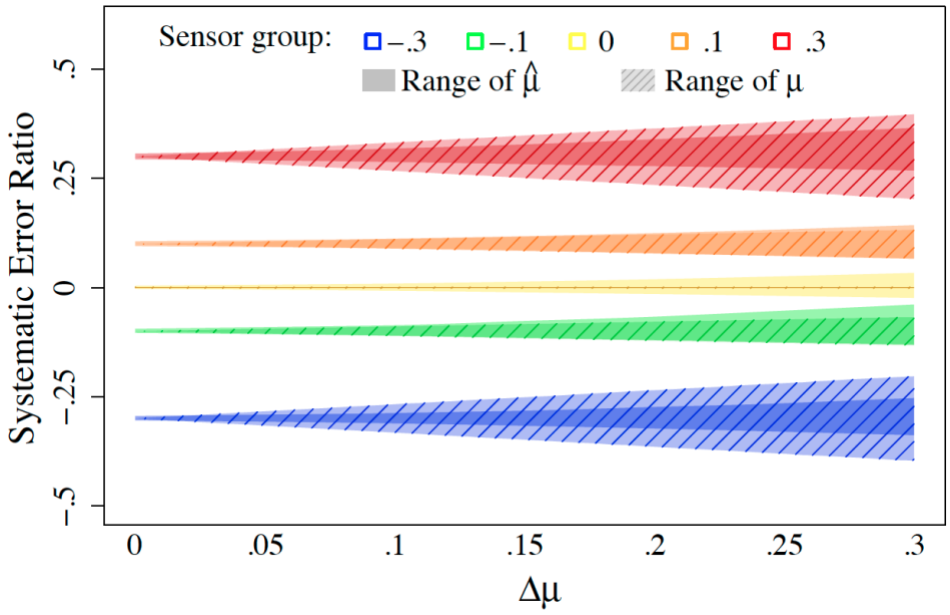


Figure 13. The Effect of Stochastic Error on Estimating μ

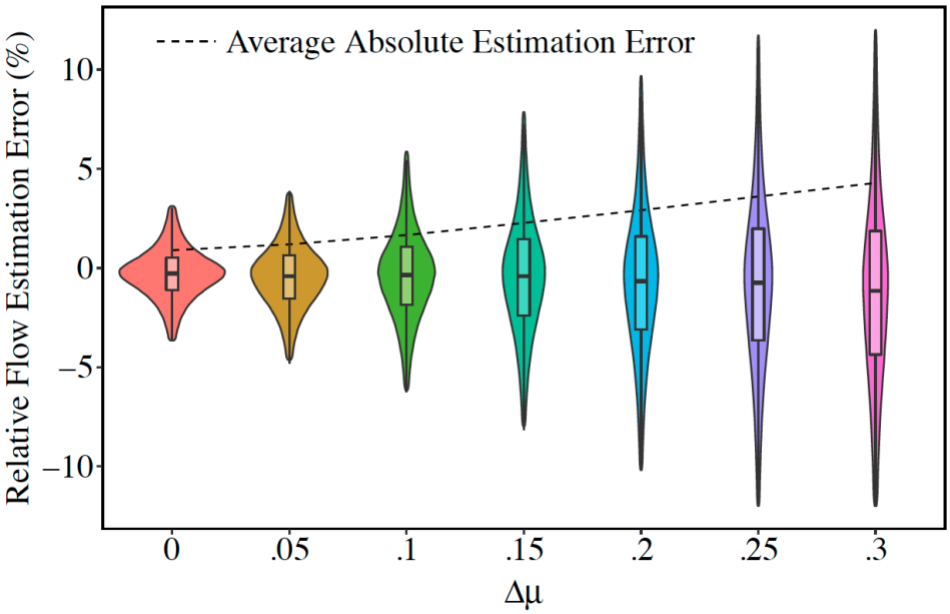


Figure 14. The Effect of Stochastic Error on Correcting Z

5.3. A Large-Scale Case

The North Orange County freeway network shown in Figure 15 is to demonstrate the scalability of this proposed method. On the OpenStreet map, the graph constituted by blue links is the example network. It is consisted of 494 nodes (92 origin/destination nodes and 402 intermediates nodes) and 674 links (362 mainline segments, 56 interfreeway ramps, 128 on ramps, and 128 off ramps). The connectivity data as well as simulated traffic flows and observation are available online[3].

Out of the 674 sensors, there are 274 healthy ones with $\mu\_a = 0$, but only 5 of them are recently calibrated, so the number of known elements in μ is only 5. Systematic error ratios μ_a for the other 400 problematic sensors is randomly drawn from a uniform distribution between –.5 and .5. Random error ratios σ_a for all the sensors are randomly drawn from a uniform distribution between .05 and .45. We still have the same 365×24 hourly data over a year. K-means clustering is used to group observations with $K = 24$. Figure 16 shows that the estimates (blue) of μ for 669 sensors of 100 simulated data samples are exceptionally good with a very narrow 95% range (red) along the diagonal line which indicates perfect estimation. To ensure that the scatter plot for Z is readable, we only present $\hat{Z}$ in one day from just a single sample (16176 data points) in Figure 17, which clearly demonstrates the correction benefits via comparing $\hat{Z}$ (blue) with observed V (green).

[3] https://github.com/yudiaspen/sensor-bias-estimation/north-orange

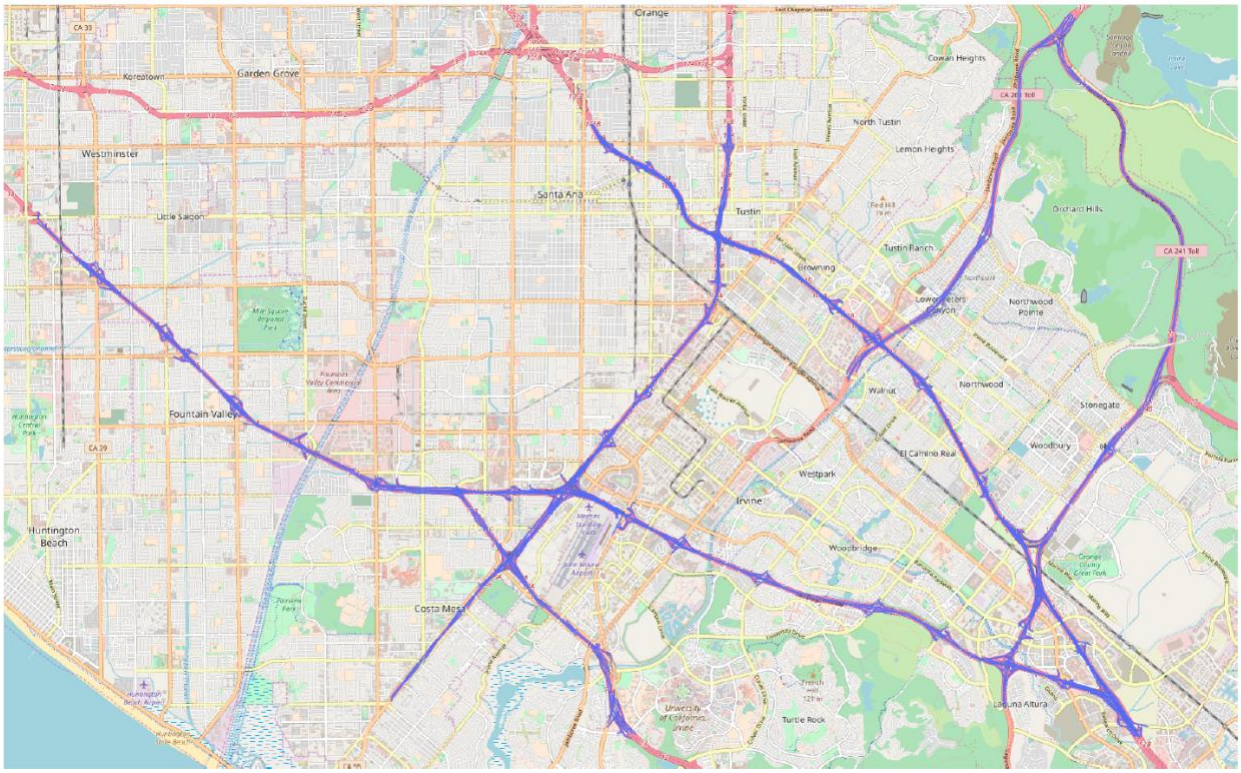


Figure 15. North Orange County Freeway Network

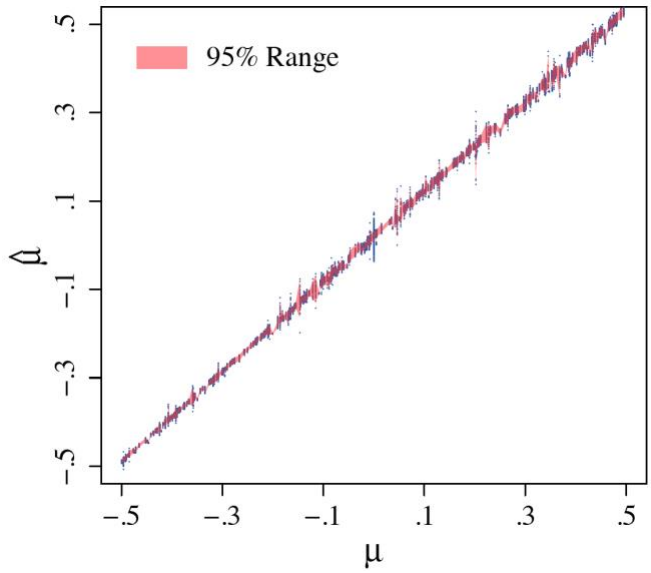


Figure 16. μ Estimation Result in North Orange County Network

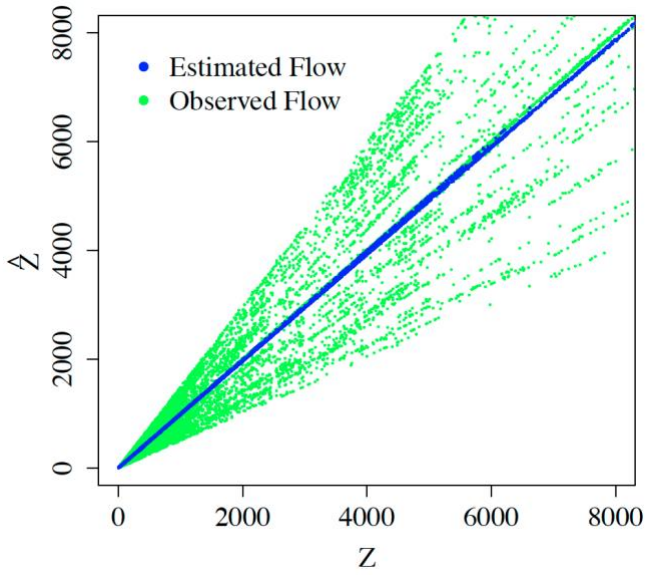


Figure 17. Z Correction Result in North Orange County Network

6. Conclusion

6.1. Contributions

In this project, we have developed a GMM-based statistical model to identify sensor measurement errors in a network context. We translate nodal flow balance law into structural equations, whose first moments are employed to estimate the systematic error ratio of sensors. The proposed framework allows a flexible data aggregation strategy, for which the traditional MOM and GLS are extreme cases. With such strategy, it is possible, without knowing random error ratios, to improve parameter identification by separating observations to more groups or to amend estimator consistency by clustering observations to fewer groups. Then we leverage the second generalized moments to obtain the estimates of random error ratios. It results in a simple quadratic minimization problem with systematic error ratios estimate known-a-priori. There are multiple uses of such nuisance parameters: first, to construct the optimal weighting matrix in order to refine estimator precision with a fixed sample size; second, to infer sensor health by conducting Wald tests; third, to derive MLE estimates for true traffic flow given observed counts.

The major contribution of our work is two-fold. First, the proposed method is capable of evaluating the level of data issue and correcting traffic flow data in addition to identifying malfunctioning sensors, while most previous sensor health studies concerned only the latter. Second, it utilizes network structure of traffic monitoring system, while many previous studies that focused on spatial relation gave attention only to those immediately neighboring sensors on a corridor. Compared to the works in Sun et al. (2016) and Yin et al. (2017), which also exploited the network feature, our method lessens their requirement of flow balance on the entire network, which may take several hours to establish. Instead, the way of flow balance equations (2.5) being used in our method, only concerns the adjacent sensors at one time and requires much less time to establish. Thus, it is possible for users to choose much shorter time interval and obtain larger sample within a fixed total observation time. It is also interesting to notice that the network flow balance equations are also widely considered in the studies of other related estimation problems, for instance, link flow inference and path or O-D flow reconstruction (Cascetta, 1984; Hazelton, 2000). In those problems, such knowledge is used to infer unknown variables based on unbiased observations. However, in this problem, even the bias and random ratios of the observed data are both unknown, thus we creatively construct relation between unknown variables, true flows based on such knowledge.

6.2. Future Extensions

The estimation method is somewhat exemplary in the sense that it provides a conservative statistical approach to a novel problem. It only considers the most well-examined data type, traffic counts as well as probably most commonly accepted measurement error and network models. In practice, there are multiple types of sensor data available, such as flow, density, and speed. Also, other than having proportional measurement errors, an error model that can exactly capture the error generation mechanisms of different sensor issues could probably result in a better fit of real data. Besides flow balance law based on network graph, other useful

transportation domain knowledge including speed-density relationship and macroscopic traffic flow models can certainly provide additional information, which should be incorporated to an error identification model in the future. Finally, it is also convenient to formulate common beliefs on sensor health, such as fewest malfunctioning sensors and least total systematic errors, using regularization techniques. In light of the highly adaptable nature of the proposed framework, we foresee no obstacle in extending the existing approach using supplementary data and knowledge types and alternative model specifications.

More interesting and important research opportunities are available when we are open to discuss technical details of constructing the network graph. The absence of sensors in certain links creates a situation that requires non-adjacent links to form a nodal flow balance relation. It must be handled with caution in order to avoid unnecessary bias introduced by relating too distant sensors. A promising way to do so is to build a new sensor network graph focusing on the spatial relation of sensors. Another issue is about sensor aggregation. In this paper, we consider a detector station as a sensor. However, it will be practically more useful to monitor the health of an individual detection unit in each lane of a multi-lane roadway, so we can narrow down to the one that needs calibration. In fact, this challenge can be handled as a natural extension of this proposed framework by splitting a road based link into multiple lane based links and augmenting a network graph to a multi-graph (multiple arc connecting two adjacent nodes). The resulting mathematical model is expected to be larger but only in a linear growth rate. Even with the same amount of available data, we may still be able to identify the model with only some small loss in estimation reliability.

6.3. Encountered Data Issues for Caltrans to Note

Throughout this project, while using PeMS traffic data, we have encountered a serious sensor data issue. The coordinates of off-ramp sensors and on-ramp sensors are wrong: All the ramp sensors that should be located on the corresponding lanes, have their coordinates on the main freeway. In most cases, their location coincides with the corresponding mainline sensors. This creates a problem in assigning the ramp sensor to the correct ramp, especially if there are multiple ramps in the vicinity of that sensor. For example, as shown in Figure 18, off ramp sensor 1123135 should be located on the corresponding ramp, but its location coordinates map it on the freeway. Similarly, the onramp sensor 1108604 is also erroneously mapped.

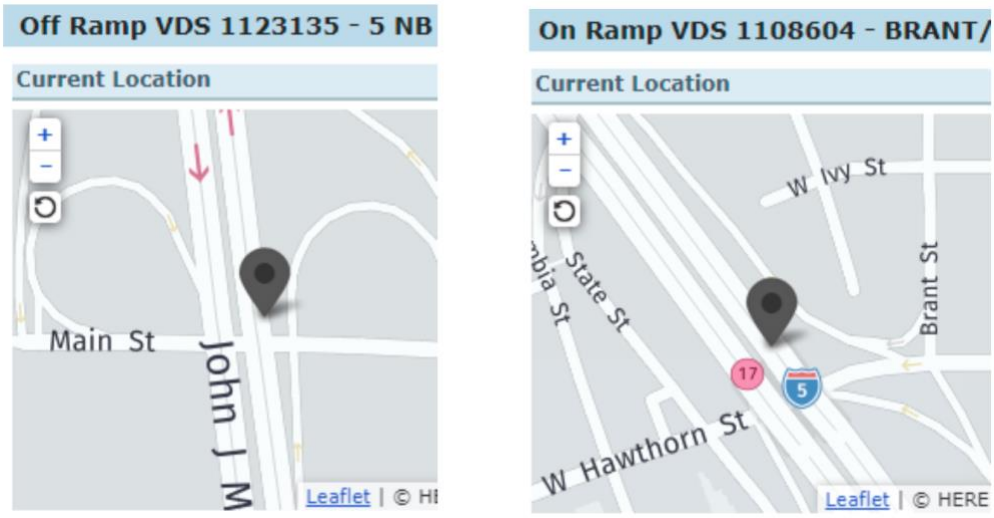


Figure 18. Example of Sensors with Incorrect Location Coordinates

References

Cascetta, E., 1984. Estimation of trip matrices from traffc counts and survey data: a generalized least squares estimator. Transportation Research Part B: Methodological 18, 289–299.

Chen, C., Kwon, J., Rice, J., Skabardonis, A., Varaiya, P., 2003. Detecting errors and imputing missing data for single-loop surveillance systems. Transportation Research Record: Journal of the Transportation Research Board, 160–167.

Dailey, D.J., 1993. Improved error detection for inductive loop sensors. Technical Report.

Daubechies, I., DeVore, R., Fornasier, M., Güntürk, C.S., 2010. Iteratively reweighted least squares minimization for sparse recovery. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences 63, 1–38.

Duan, Y., Lv, Y., Liu, Y.L., Wang, F.Y., 2016. An efficient realization of deep learning for traffic data imputation. Transportation research part C: emerging technologies 72, 168–181.

Dunn, G., 1989. Design and analysis of reliability studies: The statistical evaluation of measurement errors. Edward Arnold Publishers.

Gillard, J., 2010. An overview of linear structural models in errors in variables regression. REVSTAT–Statistical Journal 8, 57–80.

Golub, G.H., Van Loan, C.F., 1980. An analysis of the total least squares problem. SIAM journal on numerical analysis 17, 883–893.

Hazelton, M.L., 2000. Estimation of origin–destination matrices from link flows on uncongested networks. Transportation Research Part B: Methodological 34, 549–566.

Hu, P., Goeltz, R., Schmoyer, R., 2001. Proof of Concept of ITS as An Alternative Data Resource: A Demonstration Project of Florida and New York Data. Technical Report. ORNL Oak Ridge National Laboratory (US).

Klein, L.A., Mills, M.K., Gibson, D.R., 2006. Traffic Detector Handbook: -Volume II. Technical Report.

Kwon, J., Chen, C., Varaiya, P., 2004. Statistical methods for detecting spatial configuration errors in traffic surveillance sensors. Transportation Research Record: Journal of the Transportation Research Board, 124–132.

Li, Y., Li, Z., Li, L., 2014. Missing traffic data: comparison of imputation methods. IET Intelligent Transport Systems 8, 51–57.

Nihan, N.L., 1997. Aid to determining freeway metering rates and detecting loop errors. Journal of Transportation Engineering 123, 454–458.

Osborne, M.R., 1985. Finite algorithms in optimization and data analysis. Wiley New York.

Qu, L., Li, L., Zhang, Y., Hu, J., 2009. Ppca-based missing data imputation for traffic flow volume: A systematical approach. IEEE Transactions on intelligent transportation systems 10, 512–522.

Rajagopal, R., Varaiya, P., 2007. Health of California's loop detector system. California PATH Program, Institute of Transportation Studies, University of California at Berkeley.

Sun, Z., Jin, W.L., Ng, M., 2016. Network sensor health problem. Transportation Research Part C: Emerging Technologies 68, 300–310.

Tang, J., Zhang, G., Wang, Y., Wang, H., Liu, F., 2015. A hybrid approach to integrate fuzzy c-means based imputation method with genetic algorithm for missing traffic volume data estimation. Transportation Research Part C: Emerging Technologies 51, 29–40.

Turner, S., 2007. Quality control procedures for archived operations traffic data: Synthesis of practice and recommendations. Final Report. Texas Transportation Institute.

Turner, S., Margiotta, R.A., Lomax, T., 2004. Monitoring urban freeways in 2003: current conditions and trends from archived operations data. Technical Report.

Turochy, R., Smith, B., 2000. New procedure for detector data screening in traffic management systems. Transportation Research Record: Journal of the Transportation Research Board, 127–131.

Vanajakshi, L., Rilett, L., 2004. Loop detector data diagnostics based on conservation-of-vehicles principle. Transportation Research Record: Journal of the Transportation Research Board, 162–169.

Yin, P., Sun, Z., Jin, W.L., Xin, J., 2017. *l*1-minimization method for link flow correction. Transportation Research Part B: Methodological 104, 398–408.

Data Management

Products of Research

The raw data used in this research are open source from OpenStreet Map and PeMS data. The script in R programming has been uploaded to NCST data portal for sharing with the scientific community for research and teaching purposes.

Data Format and Content

The computer script is in R programming, and can be viewed in a text editor.

Data Access and Sharing

The raw data used in this research are open source for the public to download without restriction. The script in R programming has been uploaded to NCST data portal for sharing with the scientific community for research and teaching purposes. Dataset DOI: https://doi.org/10.25338/B8TP5Q

Reuse and Redistribution

The data can be reused for research and teaching purposes. It should not be used for commercial purpose. Usage of the data requires shall be acknowledged with a proper citation for the dataset.