Reducing the dimension of online calibration in Dynamic Traffic Assignment systems

Arun Prakash A., Corresponding Author
Massachusetts Institute of Technology
77 Massachusetts Avenue, 1-180, MA 02139, USA
Email: arunakkin@gmail.com

Ravi Seshadri
Singapore-MIT Alliance for Research and Technology (SMART)
1 Create Way, #09-01, Singapore 138602, Singapore
Email: ravi@smart.mit.edu

Constantinos Antoniou
Technical University of Munich
80333 Munich, Arcisstr. 21, Germany
Email: c.antoniou@tum.de

Francisco C. Pereira
Technical University of Denmark
Building 116, room 123A, 2800 Kgs. Lyngby, Netherlands
Email: camara@dtu.dk

Moshe E. Ben-Akiva
Massachusetts Institute of Technology
77 Massachusetts Avenue, 1-181, MA 02139, USA
Email: mba@mit.edu

Word count: 5197 words + 2250 (7 figures + 2 tables) = 7447

TRR Paper number: 17-04138

Submission Date: 15 March 2017

ABSTRACT

Effective real-time traffic management strategies often require Dynamic Traffic Assignment systems that are calibrated online. But the computationally intensive nature of online calibration limits their application to smaller networks. This paper presents a principal component based dimensionality reduction of the online calibration problem, which overcomes this limitation. To demonstrate this approach, we formulate the origin-destination flow estimation problem in terms of their principal components. The efficacy of the procedure is tested using real data on Singapore Expressway network in an open loop framework. We observe a reduction in the problem dimension by a factor of 50 with only 2% loss in estimation accuracy. Further, the computational times reduced by an order of 100. Interestingly, the procedure led to better predictions as the principal components capture the structural spatial relationships. This work has the potential to make the online calibration problem more scalable.

INTRODUCTION

Dynamic Traffic Assignment (DTA) systems are increasingly being used in practice to aid in policy decision-making and traffic control management. DTA systems are calibrated to better represent the real world. Calibration of DTA systems involves adjusting the inputs and estimated model parameters to replicate the real-world measurements. Traditionally, calibration is classified into two categories: offline calibration *(1)* and online calibration *(2)*. Offline calibration involves estimating an 'average day' and is essential in making medium term policy decisions. Online calibration, shown in Figure 1, involves adjusting the parameters from the offline calibration using measurements in real-time to better estimate and predict the traffic conditions. This is essential for the deployment of effective real-time traffic control strategies, such as adaptive tolling, incident management, and consistent information provision. In this paper, the online calibration variables are only the origin-destination flows. However, the proposed procedure can be extended to include other calibration variables.

The motivation of this study is the increasing need for computationally-efficient approaches to large-scale online calibration. It is essential not only in the interest of scalability, but also for the effectiveness of policies. For example, in the context of system-level adaptive tolling, the network needs to be representative of real-world route choices, which typically involve large choice-sets.

The objectives of this study are the following: 1) To reduce the dimension of the origin-destination flows using principal component analysis; 2) To formulate the online calibration problem in terms of the principal components; and 3) To test the proposed procedure using a case study with real data.

The contributions of this study are, first, the online calibration problem is formulated in terms of principal components. Second, the case study conducted on the real-world Singapore expressway network demonstrates that using principal components reduces the dimensionality of the problem by 50 times with little loss in accuracy in estimation of origin-destination flows. Thirdly, principal component based calibration is shown to perform better than the conventional calibration in traffic prediction as it captures the spatial relationships and avoids over-fitting, while estimating the flows.

LITERATURE REVIEW

As discussed, calibration of DTA systems is classified into two categories: offline calibration and online calibration. The offline calibration problem for the DTA Systems has been extensively studied over the past two decades *(1,3)*. In this section, we focus on the online calibration problem.

The online calibration problem involves fine-tuning of the historical parameter values —that are the result of offline calibration— to better represent the current traffic flow conditions in real-time. The literature can be broadly classified into two categories, depending on whether the calibration variables are either origin-destination flows or supply-side parameters. Some studies have also simultaneously calibrated supply and demand parameters *(4)*.

Initial studies on origin-destination flow calibration involved heuristics. Peeta and Bulusu *(5)* used a least squares approach to minimize the discrepancy between simulated and observed flows. The least squares problem was combined with an analytical DTA formulation. The problem was formulated as a Genalized Least Squares (GLS) problem in *(6)* where the observed consistency in

the current time interval dictates the adjustments in the next time interval. The GLS porblem is an extension of the traditional least squares problem where each term in the objective function is given a different weight corresponsing to its `importance'. Ashok and Ben-Akiva *(7)* proposed a Kalman Filter approach to correct the historical estimates of OD flows based on current information, and showed the equivalence between the GLS approach and Kalman Filter. They also introduced the idea of estimating the deviations in parameters —as opposed to parameters themselves— to explicitly capture the historical information. Zhou and Mahmassani *(8)* also used a Kalman filter, where the transition equation is a polynomial trend filter designed to capture historical trends and structural deviations.

In calibrating the supply-side parameters, Zhou and Mahmassani presented a dynamic programming approach to adjust the flow propagation where the simulator was approximated at a macroscopic level *(9)*. Antoniou et al. *(10)* used extended, limiting, and unscented Kalman Filters to calibrate the parameters of speed-density relationships. The same authors extended the procedure to simultaneously calibrate both supply and demand parameters *(4)*, proposing as a practical online calibration algorithm the Limiting Extended Kalman Filter, as it drastically reduces the computational complexity of the Extended Kalman Filter, without sacrificing too much of its performance. Hashemi and Abdelghany *(11)* proposed a simultaneous calibration approach, where the supply side parameters are calibrated using a feedback controller and demands are calibrated using a least squares approach.

In recent literature, the focus has been on incorporating the new data sources into online calibration. Automatic vehicle identification data was used in *(12)*. Data from bluetooth devices was used in *(13)*. Travel times from GPS data are also being used.

To tackle the high dimensionality of the problem, Frederix et al. *(14)* proposed a network decomposition. Further, dimensionality reduction of the calibration problem was studied in *(15,16)*, where principal component analysis was used. This study differs from the above in three aspects. First, we demonstrate the applicability to online calibration —where the focus is on both estimation and prediction— as opposed to offline calibration. Second, we use real-world data to test the efficacy of the principal component based formulation. Third, we use a generalized least squares approach as opposed to the Kalman Filter.

In summary, the research in online calibration has been on focused two aspects: developing efficient algorithms and incorporating different types of measurements. However, the application of online calibration to large networks is still computationally challenging. An alternative approach for efficiency is to reduce the dimensionality of the problem, on which there is limited work. Although the temporal relationships between the origin-destination demands are incorporated into the online calibration, the spatial relationships are seldom modeled in practice due to difficulty in estimating them. This study tries to address these gaps.

ONLINE CALIBRATION: PROBLEM FORMULATION

The online calibration problem is briefly formulated in the following paragraphs. The reader is referred to *(4,17)* for a more detailed discussion. Consider an analysis period T which is divided into equal intervals $h = 1,2,3 \dots n$ of size t. The transportation network is represented by $G(N, L, S)$, where N represents the set of nodes, L represents the set of links, and S represents the set of segments. The network has n_N nodes, n_L links, and n_S segments. The segments are sections

of road with homogeneous geometry; a link comprises one or more segments. The set of OD pairs are represented by K and are n_K in number. Further, n_s of the n_S segments are assumed to be equipped with surveillance sensors.

Let $\boldsymbol{\pi}_h$represent the parameter vector in time interval h; it contains the OD flow variables, along with behavioral and supply parameters. Similarly, let $\boldsymbol{\pi}_h^a$ represent the *a priori* estimate of the parameters in interval h. The direct measurement equation in the parameters is given by

$$\boldsymbol{\pi}_h^a = \boldsymbol{\pi}_h + \boldsymbol{\eta}_h \tag{1}$$

where $\boldsymbol{\eta}_h$ is the vector of random errors. Also, let $\boldsymbol{\pi}_h^H$ represent the historical values of the parameters in interval h. The historical values $\boldsymbol{\pi}_h^H$ are generally obtained through offline calibration *(1)*. From *(4)*, the *a priori* estimate $\boldsymbol{\pi}_h^a$ can be given by

$$\boldsymbol{\pi}_h^a = \boldsymbol{\pi}_h^H + \sum_{i=h-q}^{h-1} \mathbf{G}_i^h \left(\boldsymbol{\pi}_i - \boldsymbol{\pi}_i^H\right) \tag{2}$$

where $\mathbf{G}_i^h$ is a matrix relating the parameter estimates of interval i to the estimates of interval h. In equation (2), q is the degree of the autoregressive process in the *deviations*. As mentioned in *(4)*, the above equation models the temporal relationship in the deviations in parameters and captures the structural information in the trip patterns through the historical OD flows. The idea of modeling and estimating deviations instead of actual parameters was proposed in *(17)*.

Let $\mathbf{M}_h$ denote the vector of measurements in interval h. The indirect measurement equation denoting the relationship between measurements and parameters is given by

$$\mathbf{M}_h - \mathbf{M}_h^H = \boldsymbol{S}(\boldsymbol{\pi}_h) - \boldsymbol{S}\left(\boldsymbol{\pi}_h^H\right) + \varsigma_h \tag{3}$$

where $\boldsymbol{S}()$ represents the simulation model and ς_h is the error term representing the validity of the measurements.

By defining the deviations in parameters and measurements as,

$$\boldsymbol{\Delta\pi}_h = \boldsymbol{\pi}_h - \boldsymbol{\pi}_h^H \tag{4a}$$

$$\boldsymbol{\Delta\pi}_h^a = \boldsymbol{\pi}_h^a - \boldsymbol{\pi}_h^H \tag{4b}$$

$$\boldsymbol{\Delta M}_h = \mathbf{M}_h - \mathbf{M}_h^H \tag{4c}$$

the direct and indirect measurements in equations (1) and (3) can be rewritten as

$$\boldsymbol{\Delta\pi}_h = \boldsymbol{\pi}_h^a + \boldsymbol{\eta}_h = \sum_{i=h-q}^{h-1} \mathbf{G}_i^h(\boldsymbol{\Delta\pi}_i) + \boldsymbol{\eta}_h \tag{5a}$$

$$\boldsymbol{\Delta M}_h = \boldsymbol{S}(\boldsymbol{\pi}_h) - \boldsymbol{S}\left(\boldsymbol{\pi}_h^H\right) + \varsigma_h \tag{5b}$$

OD flow estimation problem

The dimensionality reduction proposed in this research can be applied to any set of parameters. To demonstrate its applicability, we choose the OD flow estimation problem which is a special case of the general online calibration problem (5). The parameters are the OD flow variables and the

measurements are the sensor-flow counts.

Let $\mathbf{x}_h$ represent the OD flow vector in time interval h, whose length is n_K. Similarly, let $\mathbf{x}_h^a$ represent the *a priori* estimate of OD flows in interval h. The direct measurement equation in OD flows is given by

$$\mathbf{x}_h^a = \mathbf{x}_h + \mathbf{u}_h \tag{6}$$

where $\mathbf{u}_h$ is the vector of random errors. Also, let $\mathbf{x}_h^H$ represent the historical value of OD flows in interval h. The *a priori* estimate $\mathbf{x}_h^a$ can be given by

$$\mathbf{x}_h^a = \mathbf{x}_h^H + \sum_{i=h-q}^{h-1} \mathbf{F}_i^h \left(\mathbf{x}_i - \mathbf{x}_i^H\right) \tag{7}$$

where $\mathbf{F}_i^h$ is an $n_K \times n_K$ matrix relating the OD estimates of interval i to the OD estimates of interval h.

Let $\mathbf{y}_h$ denote the vector of sensor-flow counts in interval h, whose length is n_s. The indirect measurement equation denoting the relationship between sensor-flow counts and OD flows is given by

$$\mathbf{y}_h = \mathbf{y}_h^H + \sum_{i=h-p}^{h} \mathbf{A}_i^h \left(\mathbf{x}_i - \mathbf{x}_i^H\right) + \mathbf{v}_h \tag{8}$$

where $\mathbf{y}_h^H = \sum_{i=h-q}^{h} \mathbf{A}_i^h \mathbf{x}_i^H$. The matrix $\mathbf{A}_i^h$ —the assignment matrix— is an $n_s \times n_s$ matrix relating the OD flows in interval i to the sensor-flow counts in interval h. In the context of DTA systems, the assignment matrix $\mathbf{A}_i^h$ is obtained from the simulator. In the above equation, p represents the maximum number of time-intervals taken to travel between any OD pair of the network. The equation (8) assigns the proportions of OD flows in intervals $h, h-1, \ldots h-p$ to the corresponding sensor flows in interval h.

The direct and indirect measurements in equations (6) and (8) can be rewritten explicitly in terms of deviations as

$$\Delta\mathbf{x}_h = \Delta\mathbf{x}_h^a + \mathbf{u}_h = \sum_{i=h-q}^{h-1} \mathbf{F}_i^h \Delta\mathbf{x}_i + \mathbf{u}_h \tag{9a}$$

$$\Delta\mathbf{y}_h = \sum_{i=h-p}^{h} \mathbf{A}_i^h \Delta\mathbf{x}_i + \mathbf{v}_h \tag{9b}$$

Let the covariance matrix of $\mathbf{u}_h$ be given by $\mathbf{\Omega}_{\mathbf{u}_h}$ and that of $\mathbf{v}_h$be given by $\mathbf{\Omega}_{\mathbf{v}_h}$.The system of equations in (9) —given the above assumptions— are solved using the GLS approach, which can be formulated as an optimization problem as

$$\mathbf{Min}.\ (\Delta\mathbf{x}_h - \Delta\mathbf{x}_h^a)^T \mathbf{\Omega}_{\mathbf{u}_h}{}^{-1} (\Delta\mathbf{x}_h - \Delta\mathbf{x}_h^a) + \left(\Delta\mathbf{y}_h - \sum_{i=h-p}^{h} \mathbf{A}_i^h \Delta\mathbf{x}_i\right)^T \mathbf{\Omega}_{\mathbf{v}_h}{}^{-1} \left(\Delta\mathbf{y}_h - \sum_{i=h-p}^{h} \mathbf{A}_i^h \Delta\mathbf{x}_i\right) \tag{10a}$$

subject to

$$\mathbf{x}_h \geq \mathbf{0} \tag{10b}$$

Note that, in the context of online systems, the above problem is solved to obtain an estimate of *only* the OD flow estimated in interval h, $\mathbf{x}_h$. The OD flow estimates of the earlier intervals $h-$

$1, h-2, \ldots$ are not re-estimated as that would involve *rolling back* the simulator in real-time, which is avoided due to computational constraints.

To predict the OD flow vectors in the subsequent intervals, the *a priori* estimates —calculated using equation in (7)— are used. For example, the prediction for the time interval $h+1$ is calculated as

$$\mathbf{x}_{h+1}^{a} = \mathbf{x}_{h+1}^{H} + \sum_{i=h+1-q}^{h} \mathbf{F}_i^h \left(\mathbf{x}_i - \mathbf{x}_i^H\right) \tag{11}$$

The matrices $\mathbf{F}_i^h$ and $\mathbf{\Omega}_{\mathbf{u}_h}$ are generally provided *a priori* and are based on the knowledge of the network. However, determining them is not trivial as only limited *a priori* knowledge is generally available. In practice, the matrix $\mathbf{F}_i^h$ is assumed to be diagonal, i.e., an OD flow in current interval depends only on its values in the previous intervals. Similarly, a diagonal $\mathbf{\Omega}_{\mathbf{u}_h}$ is generally assumed which implies that the OD flows are not correlated with each other. These restrictive assumptions can deteriorate the accuracy of the traffic estimates and predictions. The principal component based calibration presented in the next section overcomes these practical issues.

PRINCIPAL COMPONENTS BASED CALIBRATION

This section discusses an approach to reduce the dimensionality of the online calibration problem: using principal components analysis. The approach reduces the dimensionality of the problem by changing the decision variables and effectively reducing the degrees of freedom of the OD flows.

Through the principal component analysis of the OD flows, the systematic variations in OD flows are captured in lower dimensions. Subsequently, the GLS problem in (10) is solved with principal components as the decision variables. Finally, the OD flows are constructed back from the principal component solution for the supply simulator in the online DTA system.

Construction of principal components

In this section we discuss the interpretation and construction of principal components of the OD flow vector. The principal components are the linear combinations of the OD flow vector that explain the majority of the *variability* of the OD flows. The co-efficients for the linear combinations are called the principal component directions. The first principal component displays the highest sample variance in the data. The second principal component has the highest sample variance subject to being independent of the first principal component, and so on. Consequently, the resulting principal components have a desirable property of being independent of each other, making the calibration problem formulation straightforward. As the OD flow vectors are generally highly correlated and sparse, the principal component analysis is expected to be ideal for the current application: to reduce and decouple the dimensions of the OD flow vector.

To construct principal components of the OD flow vector, multiple OD flow estimates are required. A straightforward approach is to use the estimated OD flows from the day-to-day offline (or online) calibration. In the offline calibration procedure, the time-dependent OD flow vectors are simultaneously estimated for the time-period of interest. Please refer to *(1)* for formulations and algorithms of the offline calibration problem. After estimating the OD flows over multiple days in the time-period of interest using offline calibration, a data matrix $\mathbf{X}$ can be constructed. The size of the data matrix $\mathbf{X}$ is $n_p \times n_K$. The variable n_p represents the number of data points,

i.e., each row of the data matrix is the estimated OD flow vector in any of the intervals. The variable n_K, as defined earlier, represents the number of OD pairs. The principal component directions can then be determined by performing Singular Valued Decomposition (SVD) on the centered data matrix $\widetilde{\mathbf{X}}$.

$$\widetilde{\mathbf{X}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}} \tag{12}$$

where $\boldsymbol{\Sigma}$ is a $n_p \times n_K$ rectangular-diagonal matrix with positive values called singular values; $\mathbf{U}$ is a $n_p \times n_p$ matrix with orthogonal column vectors called the left singular vectors; and, $\mathbf{V}$ is a $n_K \times n_K$ matrix with orthogonal column vectors called the right singular vectors. The columns of the matrix $\mathbf{V}$ are the principal component directions. Alternatively, the columns of $\mathbf{V}$ can also be interpreted as the eigenvalues of the matrix $\frac{1}{n_p}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$, which is the sample covariance matrix. Note that the principal components as calculated above capture only the structural *spatial* relationship between the OD flows and not the temporal relationships.

Let the individual principal component directions of the OD flow vector $\mathbf{x}$ be represented by $\mathbf{v}_1, \mathbf{v}_2 \ldots \mathbf{v}_{n_K}$. Here, $\mathbf{v}_1$ represents the principal component direction with the largest sample variance; $\mathbf{v}_2$ represents the principal component direction with the largest sample variance subject to being orthogonal to $\mathbf{v}_1$, and so on. Assume that only first n_d directions explain a majority of the variance in the OD flow vector. The first n_d principal component directions can be represented using a $n_K \times n_d$ matrix as

$$\mathbf{V} = [\mathbf{v_1} \quad \mathbf{v_2} \ldots \mathbf{v}_{\boldsymbol{n_d-1}} \quad \mathbf{v}_{\boldsymbol{n_d}}] \tag{13}$$

Then the $n_d \times 1$ principal component vector $\mathbf{z} = [z_1 \quad z_2 \ldots z_{n_d-1} \quad z_{n_d}]^T$ of the OD flow vector $\mathbf{x}$ can be written as

$$\mathbf{z} = \mathbf{V}^{\boldsymbol{T}}\mathbf{x} \tag{14}$$

and the OD flow vector $\mathbf{x}$ can be approximately constructed back as

$$\mathbf{x} \approx \mathbf{V}\mathbf{z} \tag{15}$$

PC-GLS problem formulation

As the online calibration GLS problem is in terms of the deviations we do the same for principal component based GLS (PC-GLS). The principal component deviations are written as

$$\boldsymbol{\Delta}\mathbf{z}_{\boldsymbol{h}} = \mathbf{z}_{\boldsymbol{h}} - \mathbf{z}_{\boldsymbol{h}}^{\boldsymbol{H}} \tag{16a}$$

$$\boldsymbol{\Delta}\mathbf{z}_{\boldsymbol{h}}^{\boldsymbol{a}} = \mathbf{z}_{\boldsymbol{h}}^{\boldsymbol{a}} - \mathbf{z}_{\boldsymbol{h}}^{\boldsymbol{H}} \tag{16b}$$

where $\mathbf{z}_{\boldsymbol{h}}^{\boldsymbol{H}} = \mathbf{V}^{\boldsymbol{T}}\mathbf{x}_{\boldsymbol{h}}^{\boldsymbol{H}}$ and $\mathbf{z}_{\boldsymbol{h}}^{\boldsymbol{a}} = \mathbf{V}^{\boldsymbol{T}}\mathbf{x}_{\boldsymbol{h}}^{\boldsymbol{a}}$. The direct and indirect measurement equations in terms of the principal components can be written as

$$\boldsymbol{\Delta}\mathbf{z}_{\boldsymbol{h}} = \boldsymbol{\Delta}\mathbf{z}_{\boldsymbol{h}}^{\boldsymbol{a}} + \mathbf{w}_{\boldsymbol{h}} \tag{17a}$$

$$\boldsymbol{\Delta}\mathbf{y}_{\boldsymbol{h}} = \sum_{\boldsymbol{i=h-p}}^{\boldsymbol{h}} \mathbf{A}_{\boldsymbol{i}}^{\boldsymbol{h}}\mathbf{V}\boldsymbol{\Delta}\mathbf{z}_{\boldsymbol{i}} + \mathbf{v}_{\boldsymbol{h}} \tag{17b}$$

which are obtained by substituting equations (14) and (15) in the direct and indirect equations in (9a) and (9b). From equations (15) and (9a), $\mathbf{w}_h$ can be written as

$$\mathbf{w}_h = \mathbf{V}^T \mathbf{u}_h \tag{18}$$

The covariance matrix of $\mathbf{w}_h$, $\mathbf{\Omega}_{\mathbf{w}_h}$ is given by

$$\mathbf{\Omega}_{\mathbf{w}_h} = \mathbf{V}^T \mathbf{\Omega}_{\mathbf{u}_h} \mathbf{V} \tag{19}$$

Following the above results, the system of equations in (17) can be solved using GLS, which can be written as the following optimization problem

$$\mathbf{Min}.\ (\mathbf{\Delta z}_h - \mathbf{\Delta z}_h^a)^T \mathbf{\Omega}_{\mathbf{w}_h}{}^{-1} (\mathbf{\Delta z}_h - \mathbf{\Delta z}_h^a) + \left(\mathbf{\Delta y}_h - \sum_{i=h-p}^{h} \mathbf{A}_i^h \mathbf{V} \mathbf{\Delta x}_i\right)^T \mathbf{\Omega}_{\mathbf{v}_h}{}^{-1} \left(\mathbf{\Delta y}_h - \sum_{i=h-p}^{h} \mathbf{A}_i^h \mathbf{V} \mathbf{\Delta x}_i\right) \tag{20a}$$

subject to

$$\mathbf{V z}_h \geq \mathbf{0} \tag{20b}$$

The constraint in (20b) ensures that the derived OD flows from the estimated principal components satisfy the non-negativity constraint. As in the standard GLS in (10), only the principal components in interval h, $\mathbf{z}_h$, are estimated in problem (20); the estimates of the earlier intervals are not re-estimated. The OD flows are constructed back from the estimates of principal components as follows

$$\mathbf{x}_h = \mathbf{V z}_h \tag{21}$$

The prediction of OD flows for the subsequent intervals is calculated as before using equation (11).

From the standpoint of practice, the spatial relationships between the OD flows can be modelled easily using the PC-GLS. The decision variables in PC-GLS are the principal components which are combination of OD flows. Intuitively, it means that the matrix relating the OD flows in one time interval to another is now non-diagonal. It implies that both the spatial and temporal structural dependencies between the OD flows are captured. Further, in implementation, the covariance matrix of the principal components $\mathbf{\Omega}_{\mathbf{w}_h}$ can be assumed to be diagonal as —by definition— the principal components are independent of each other in the data sample.

CASE STUDY ON SINGAPORE EXPRESSWAY NETWORK

In this section, we apply the principal component based online calibration presented in Section 5 to the Singapore Expressway Network. Firstly, the simulation setup and data is explained followed by the estimation of the inputs to the online calibration. Finally, the results are discussed.

Overview

The case study is conducted on the Singapore Expressway Network and the real-time DTA system used is DynaMIT-R *(18)*. The road network is depicted in Figure 2, which has 939 nodes, 1157 links, and 3906 segments. The network specification also contains information about segment lengths, segment curvatures, speed limits, lanes specifications, lane-connections, and dynamic tolling gantries which are replicated from the real-world.

The network has 4121 OD pairs, whose locations and historical values were determined by an earlier work through offline calibration *(19)*. The work also determined the supply-side parameters, which include the modified Greenshield's speed-density equation parameters and the segment capacities. The network has 357 sensors, each of which is associated with a segment; these video camera based sensors count the vehicular flow for a period of 5 minutes. The Land Transport Authority (LTA) of Singapore provided the measured sensor counts after preprocessing the raw-data.

For the current case study, the simulation time-period was taken from 06:00 hrs to 12:00 hrs which includes the morning peak period and also the peak to off-peak transition. The estimation interval was 5 minutes and the prediction interval —to estimate future traffic states and provide guidance— was 15 minutes. Thus, we have 72 estimation intervals with a total of $72 \times 4121 = 102312$ variables.

For this study, sensor count data of 30 weekdays in August-September 2015 was used. The 30 day data was divided into a training set of the first 25 days and the calibration procedures were tested on the last 5 days.

Estimating the inputs to the calibration

To solve the standard GLS online calibration problem in (10), estimates of the following variables are required: covariance matrix $\mathbf{\Omega}_{\mathrm{u}}$, covariance matrix $\mathbf{\Omega}_{\mathrm{v}}$, and the autoregressive matrix $\mathbf{F}_i^h$ for $i = h, h-1, \ldots h-q$. Additionally, to solve the principal component based GLS problem in (20), the principal component directions of the OD flow vector, $\mathbf{V}$, also need to be determined.

The covariance matrix $\mathbf{\Omega}_{\mathrm{v}}$, which measures the variance in sensor measurements was determined based on sensor credibility. The covariance matrix $\mathbf{\Omega}_{\mathrm{u}}$ was determined using the procedure outlined in *(2)*. Specifically, an Ordinary Least Squares (OLS) problem —where $\mathbf{\Omega}_{\mathrm{u}}$ is an identity matrix in problem (10)— was solved for the first 10 days of the training data using the corresponding sensor counts. From the estimates determined by the OLS problem, the covariance matrix of $\mathbf{u}$, $\mathbf{\Omega}_{\mathrm{u}}$, was estimated. To determine the autoregressive equation (7), first, the standard GLS problem in (10) was solved to get the estimates of OD flow using the previously determined estimates of $\mathbf{\Omega}_{\mathrm{v}}$ and $\mathbf{\Omega}_{\mathrm{u}}$. From these estimates of OD flow the autoregressive equation in (7) was determined for each OD pair assuming that the OD pairs are independent of each other. Note that when no autoregressive relation could be established a random walk was assumed.

The data from days 11 to 25 were used to determine the principal component directions of the OD flow vector. Specifically, the online calibration was carried out to estimate the time-dependent OD flow vectors in the simulation period for the 10 days. As discussed in Section 5.1, a data matrix was constructed —with $n_p = 72 \times 10 = 720$ data points— on which the principal component analysis was carried out. The results of the principal component analysis are presented in Figure 3

From Figure 3, very few principal components explain a majority of the variance in the OD flows. Specifically, only 75 principal components explain about 95% of the variance. It appears that most of the variation in the data can be captured using only a few components facilitating the reduction in the online calibration problem dimension. For the purpose of this experiment, $n_d = 75$ principal components were chosen. The dimensionality of the problem has been reduced by a

factor of 54.

Calibration results

The results from the online calibration on the test set of the final 5 days are discussed in this section. The performance measures adopted were the Normalized Root Mean Squared (RMSN) Errors and Mean Absolute Percentage Errors (MAPE) which are defined as

$$\mathbf{RMSN} = \frac{\sqrt{n\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}}{\sum_{i=1}^{n} y_i} \tag{22a}$$

$$\mathbf{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{y_i} \tag{22b}$$

where $\boldsymbol{y_i}$ represents actual measurement and $\boldsymbol{\hat{y}_i}$ represents simulated measurement. The sensor count RMSNs and MAPEs for each of the days of GLS (10) and PC-GLS (20) are presented in Table 1. The errors in sensor counts estimated without online calibration —which represents the historical values— are also presented. The percentage difference in errors of between GLS and PC-GLS is also presented in the table which is calculated as

$$\mathbf{100} \times \frac{err_{pc-gls}-err_{gls}}{err_{gls}} \tag{23}$$

where err represents either MAPE or RMSN.

In the context of estimation, the GLS on average exhibits an RMSN of 0.268 and MAPE of 18.81%. It improves over the historical by 75% in RMSN and 50% in MAPE. The PC-GLS on average exhibits an RMSN of 0.272 and MAPE of 19.95%: the error values are close to those of the GLS with percentage difference of 1.5% and 6%. This is also observed in each of the individual 5 days. It appears that the PC-GLS can indeed perform as well as GLS in the estimation interval with added computational benefits from dimensionality reduction. The time-dependent sensor count RMSN values in estimation for historical, GLS, and PC-GLS are presented in Figure 4. These trends for the 5 test days also show that the PC-GLS consistently performs nearly as well as GLS in estimating the sensor counts.

The results from the prediction are represented in three steps, which represent the three 5 minute intervals in the complete 15 minute prediction interval. In the context of prediction, from Table 1, the GLS on average exhibits an RMSN of 0.311, 0.350, and 0.372 for the three steps and MAPE of 20.49%, 22.33%, and 23.41% for the three steps. The PC-GLS on average exhibits an RMSN of 0.276, 0.289, and 0.299 for the three steps and MAPE of 20.27%, 21.93%, and 22.92% for the three steps. The PC-GLS does better than standard GLS in prediction by about 16% in RMSNs and 1.7% in MAPEs. These improvements can be explained using two conjunctures. Firstly, GLS —which has a higher number of decision variables compared to PC-GLS— appears to overfit in the current interval leading to poorer predictions. Secondly, the OD flow predictions from the PC-GLS take into consideration the structural OD flow variations/patterns; the GLS does not consider such variations. Therefore, the PC-GLS has better predictions compared to standard GLS. The time-dependent sensor count RMSNs in prediction are presented in Figure 5. These trends show that PC-GLS consistently performs better than the GLS in predicting the sensor counts.

The scatter plots representing the estimated/predicted sensor counts versus actual sensor counts are presented in Figures 6 and 7 for the second test day. The complete day's sensor count values are represented in a single plot as a heat map. As expected, the the estimation results are better than predictions for both the GLS and PC-GLS procedures. From Figures 6a and 6b the PC-GLS seems to underestimate the sensor counts at higher values, probably because of the imposed dimensionality constraint. From Figures 6c, 6d, and 7 the better predictions of PC-GLS compared to GLS can be ascertained.

Finally, the comparison of the computational times of the GLS and PC-GLS procedures are presented in Table 2. The computational times presented are the average values over the 72 estimation intervals for a given day. From the results, the PC-GLS is substantially more computationally efficient than the standard GLS procedure for all the five days: the times are shorter by the order of 100. This demonstrates that the proposed approach has a practical impact of reducing not only the problem dimension but also the computational times.

CONCLUSIONS

This paper presented an approach to reduce the dimensionality of the online calibration problem through principal component analysis. The construction of principal components of the OD flow vector was discussed so as to capture the structural, spatial relationship between the OD flows. The online calibration problem was formulated in the principal components and was solved using the traditional Generalized Least Squared approach. Finally, a case study with real world data on a large network was presented to assess the performance of the principal component based online calibration. The rigorous performance evaluation over 5 test days led to the following insights

1. As only 75 principal components of 4121 OD pairs were used in PC-GLS, the dimensionality of the problem was reduced by 54 times.
2. The proposed PC-GLS procedure —because of the dimentionality reduction— was also faster by an order of 100 compared to the traditional GLS procedure.
3. In the context of estimation, the PC-GLS performed nearly as well as GLS with an average percentage difference in RMSN of 1.5% and in MAPE of 6%.
4. In the context of prediction, PC-GLS did better than GLS by 16% in RMSN and by 1.7% in MAPE on average. We believe that PC-GLS performed better due to the following reasons: (i) GLS, which has a large number of variables, tends to overfit in the estimation interval leading to poor predictions, and (ii) the PC-GLS captures the structural relations which are also realized in its predictions.

Finally, exploring other dimensionality reduction techniques, like clustering, and testing the performance of other algorithms, like the extended Kalman Filter, under dimensionality reduction offer valuable scope for further investigations.

ACKNOWLEDGEMENT

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore, under its CREATE program, Singapore-MIT Alliance for Research and Technology (SMART) Future Urban Mobility (FM) IRG.

REFERENCES

1. Balakrishna, R., *Off-line calibration for dynamic traffic assignment models*. Ph.D. thesis, Massachusetts Institute of Technology, 2006.
2. Antoniou, C., *On-line calibration for dynamic traffic assignment*. Ph.D. thesis, Massachusetts Institute of Technology, 2004.
3. Balakrishna, R., M. Ben-Akiva, and H. Koutsopoulos, Offline Calibration of Dynamic Traffic Assignment: Simultaneous Demand-and-Supply Estimation. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2003, 2007, pp. 50–58.
4. Antoniou, C., M. Ben-Akiva, and H. N. Koutsopoulos, Nonlinear Kalman filtering algorithms for on-line calibration of dynamic traffic assignment models. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 8, No. 4, 2007, pp. 661–670.
5. Peeta, S. and S. Bulusu, Generalized singular value decomposition approach for consistent on-line dynamic traffic assignment. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1667, 1999, pp. 77–87.
6. Zhou, X. and H. Mahmassani, Online consistency checking and origin-destination demand updating: Recursive approaches with real-time dynamic traffic assignment operator. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1923, 2005, pp. 218–226.
7. Ashok, K. and M. E. Ben-Akiva, Alternative approaches for real-time estimation and prediction of time-dependent origin–destination flows. *Transportation Science*, Vol. 34, No. 1, 2000, pp. 21–36.
8. Zhou, X. and H. S. Mahmassani, A structural state space model for real-time traffic origin-destination demand estimation and prediction in a day-to-day learning framework. *Transportation Research Part B: Methodological*, Vol. 41, No. 8, 2007, pp. 823–840.
9. Zhou, X. and H. Mahmassani, Dynamic programming approach for online freeway flow propagation adjustment. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1802, 2002, pp. 263–270.
10. Antoniou, C., M. Ben-Akiva, and H. Koutsopoulos, Online calibration of traffic prediction models. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1934, 2005, pp. 235–245.
11. Hashemi, H. and K. Abdelghany, Integrated Method for Online Calibration of Real-Time Traffic Network Management Systems. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2528, 2015, pp. 106–115.
12. Antoniou, C., M. Ben-Akiva, and H. Koutsopoulos, Incorporating automated vehicle identification data into origin-destination estimation. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1882, 2004, pp. 37–44.
13. Barceló, J., L. Montero, L. Marqués, and C. Carmona, Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2175, 2010, pp.19–27.
14. Frederix, R., F. Viti, and C. M. Tampère, A hierarchical approach for dynamic origin-destination matrix estimation on large-scale congested networks. In *14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2011, pp. 1543–1548.

15. Djukic, T., J. Van Lint, and S. Hoogendoorn, Application of principal component analysis topredict dynamic origin-destination matrices. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2283, 2012, pp. 81–89.
16. Djukic, T., *Dynamic OD demand estimation and prediction for dynamic traffic management*. Ph.D. thesis, TU Delft, Delft University of Technology, 2014.
17. Ashok, K. and M. E. Ben-Akiva, Estimation and prediction of time-dependent origin destination flows with a stochastic mapping to path flows and link flows. *Transportation Science*, Vol. 36, No. 2, 2002, pp. 184–198.
18. Ben-Akiva, M., H. N. Koutsopoulos, C. Antoniou, and R. Balakrishna, Traffic simulation with DynaMIT. In *Fundamentals of traffic simulation*, Springer, 2010, pp. 363–398.
19. Lu, L., Y. Xu, C. Antoniou, and M. Ben-Akiva, An enhanced SPSA algorithm for the calibration of dynamic traffic assignment models. *Transportation Research Part C: Emerging Technologies*, Vol. 51, 2015, pp. 149–166.

LIST OF TABLES

LIST OF FIGURES

TABLE 1 Aggregate values of RMSNs and MAPEs of sensor-flow counts (5 min) for historical, GLS, and PC-GLS

Days	Method	RMSN Estimation	RMSN Prediction			MAPE Estimation	MAPE Prediction		
			step1	step2	step3		step1	step2	step3
1	Hist	0.431	0.43	0.429	0.429	26.59	26.75	26.91	26.88
	GLS	0.275	0.311	0.353	0.373	16.75	18.44	20.7	21.64
	PC–GLS	0.28	0.282	0.287	0.295	17.71	18.03	18.85	19.97
	% difference	-1.84	9.49	18.65	20.81	-5.73	2.22	8.94	7.72
2	Hist	0.437	0.436	0.435	0.435	29.99	30.15	30.12	29.93
	GLS	0.266	0.32	0.359	0.379	22.01	23.49	24.28	25.02
	PC–GLS	0.271	0.275	0.305	0.318	23.82	23.32	24.54	24.99
	% difference	-2.01	13.93	15.11	16.25	-8.22	0.72	-1.07	0.12
3	Hist	0.439	0.438	0.436	0.437	27.9	28.01	28.1	28.04
	GLS	0.276	0.308	0.357	0.385	17.13	19.9	21.97	23.7
	PC–GLS	0.277	0.278	0.289	0.3	17.67	19.09	21.76	23
	% difference	-0.32	9.7	19.12	22.18	-3.15	4.07	0.96	2.95
4	Hist	0.422	0.421	0.419	0.419	27.15	27.21	27.34	27.22
	GLS	0.265	0.305	0.342	0.365	17.96	19.33	21.68	22.87
	PC–GLS	0.267	0.276	0.284	0.293	18.85	19.27	21.03	22.11
	% difference	-1.01	9.65	16.96	19.86	-4.96	0.31	3	3.32
5	Hist	0.408	0.407	0.406	0.406	28.75	28.85	29.01	28.85
	GLS	0.26	0.309	0.338	0.356	20.18	21.26	23.01	23.8
	PC–GLS	0.266	0.268	0.28	0.291	21.67	21.66	23.46	24.52
	% difference	-2.33	13.28	16.98	18.41	-7.38	-1.88	-1.96	-3.03
Average	Hist	0.427	0.426	0.425	0.425	28.08	28.19	28.3	28.19
	GLS	0.268	0.311	0.35	0.372	18.81	20.49	22.33	23.41
	PC–GLS	0.272	0.276	0.289	0.299	19.95	20.27	21.93	22.92
	% difference	-1.49	11.23	17.36	19.52	-6.06	1.07	1.79	2.09

TABLE 2 Average computational times for GLS and PC-GLS problems for each of the 5 test days

Days	Computation Times (sec)	
	GLS	PC-GLS
1	112.9	1.18
2	137.8	1.21
3	86.3	1.14
4	97.9	1.13
5	102	1.14

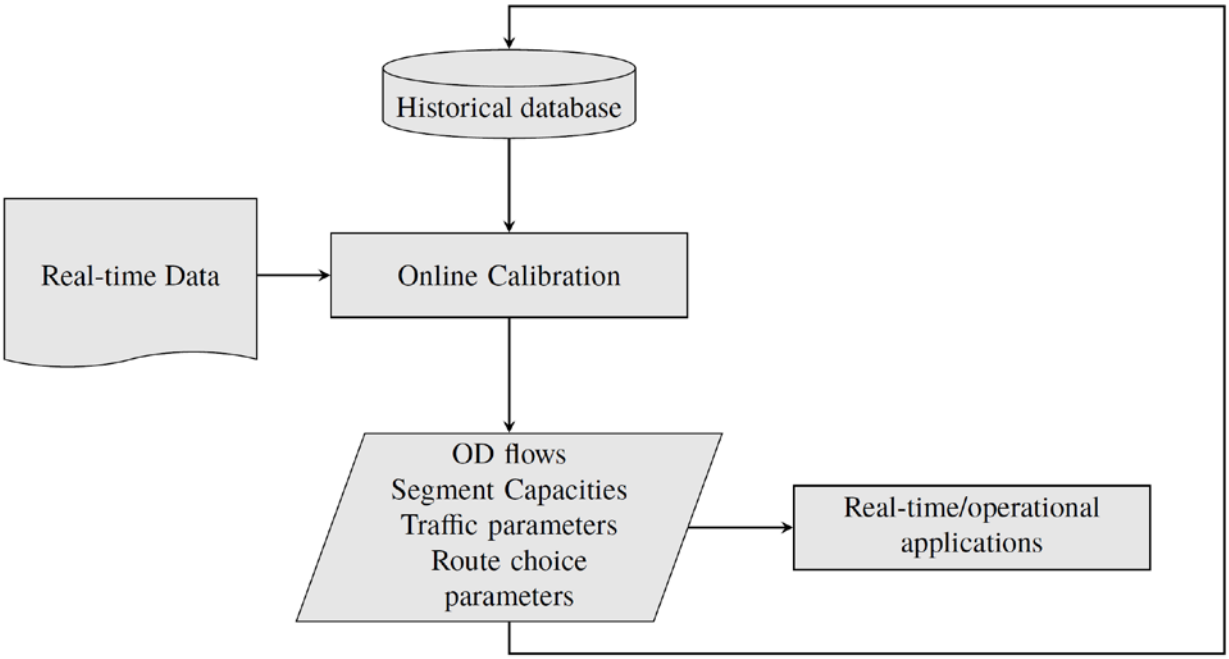
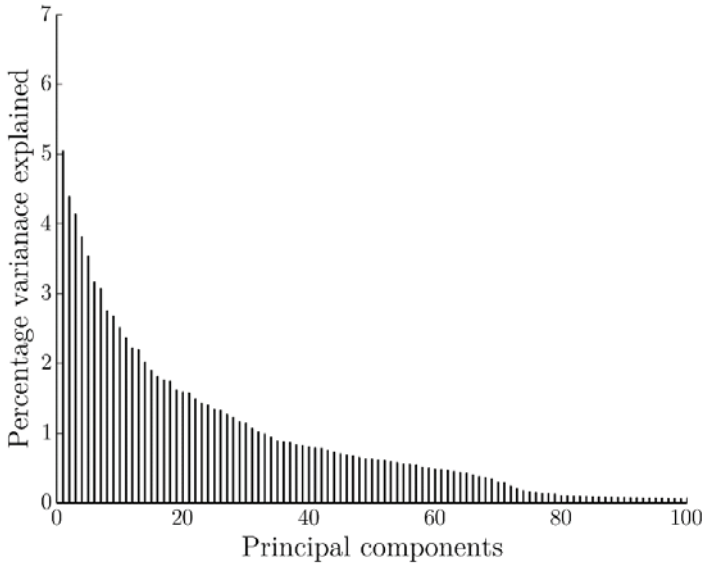
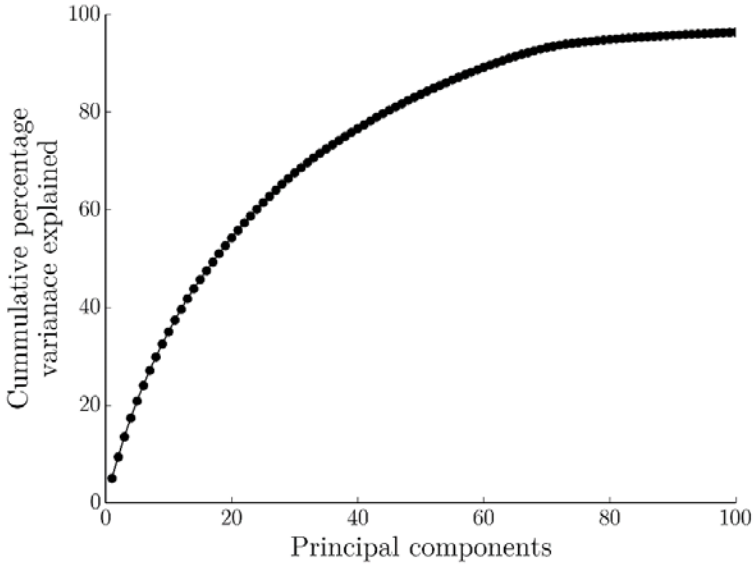


FIGURE 1 Flowchart of Online Calibration Process

FIGURE 2 Singapore Expressway Network

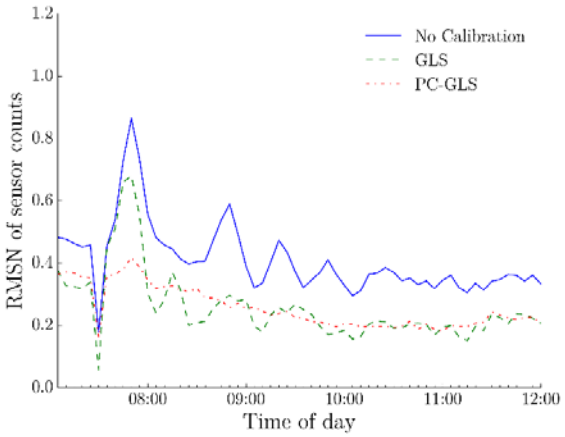


(a) Variance Explained

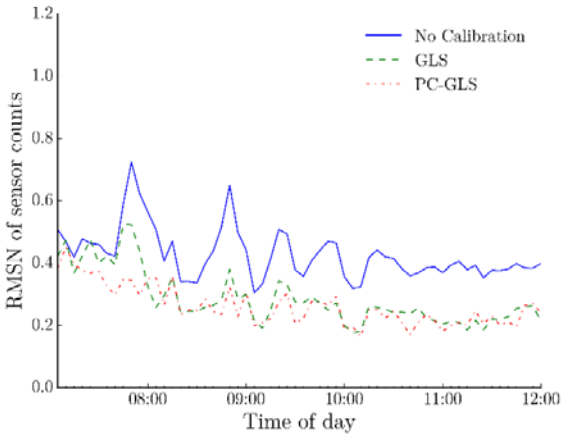


(b) Cumulative Variance Explained

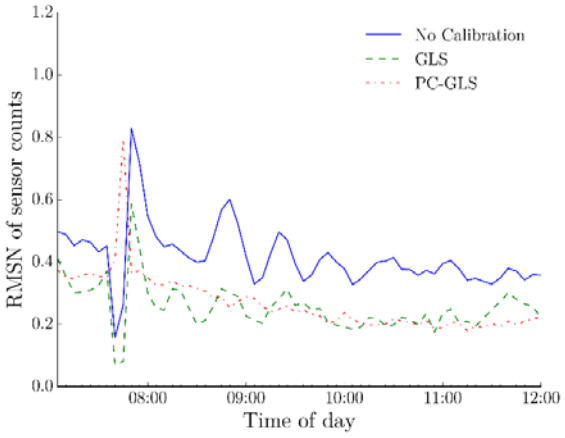
FIGURE 3 Principal Component Analysis of OD flow vector using the estimated values from day 11 to day 25



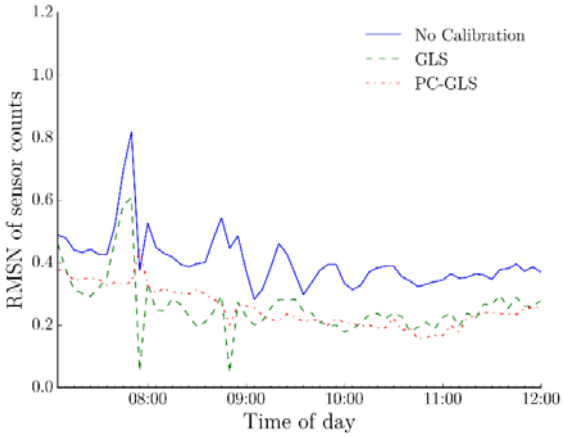
(a) Day 1



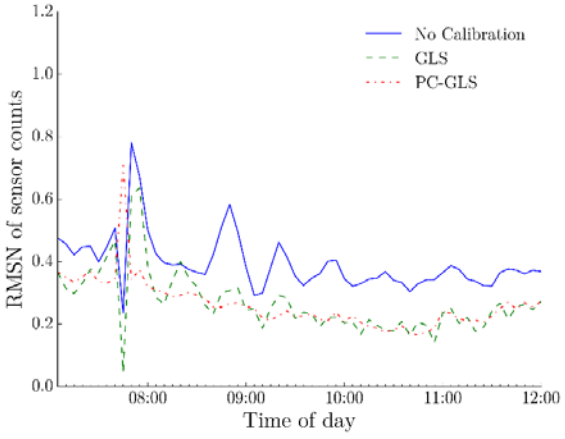
(b) Day 2



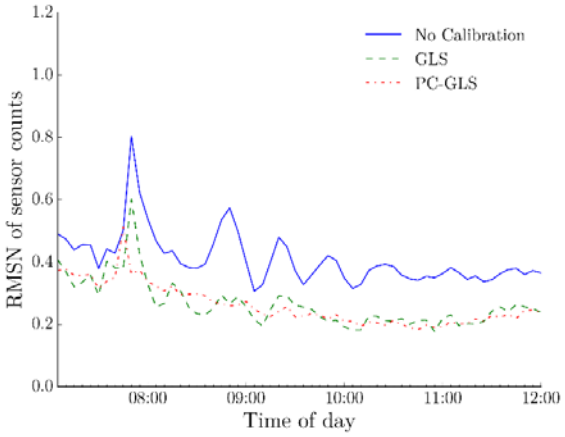
(c) Day 3



(d) Day 4

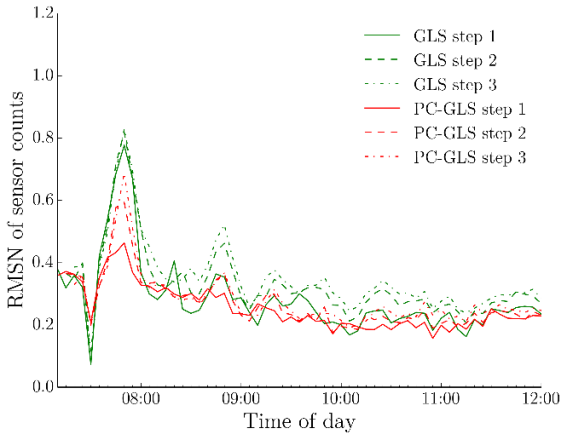


(e) Day 5

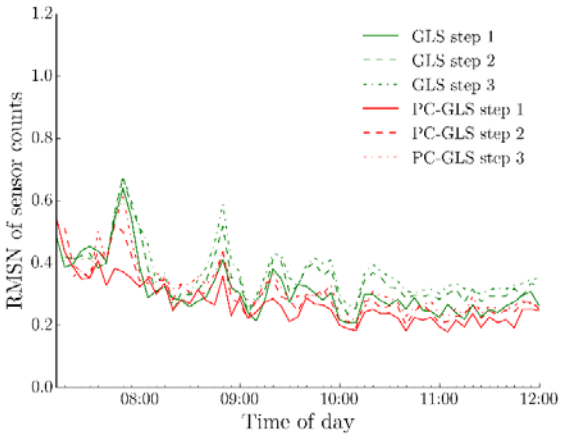


(f) Average across days

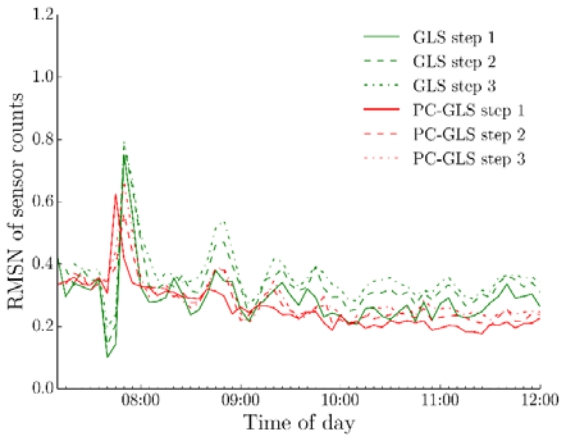
FIGURE 4 Plots of sensor count RMSNs for GLS and PC–GLS with respect to time–of–day in the estimation interval



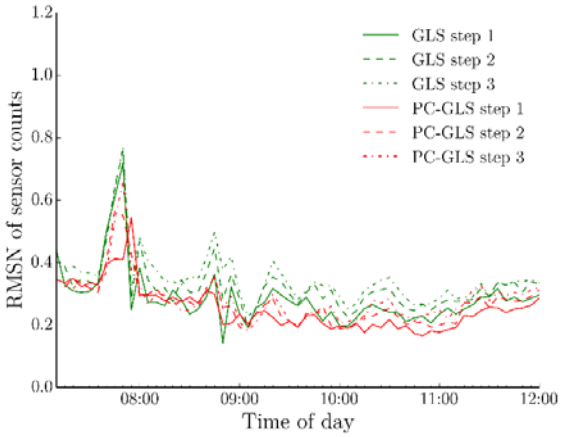
(a) Day 1



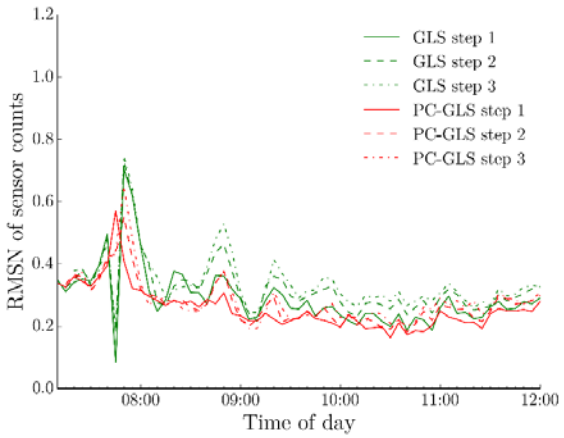
(b) Day 2



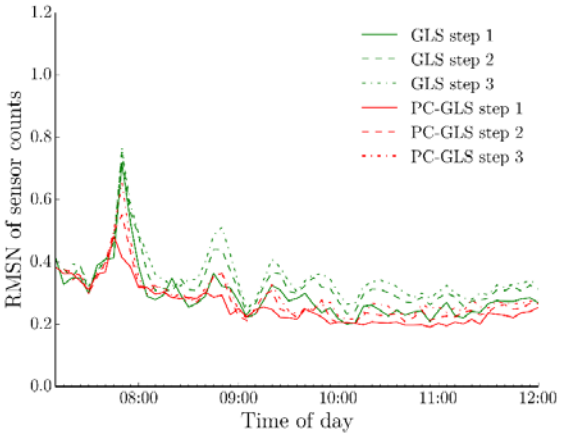
(c) Day 3



(d) Day 4

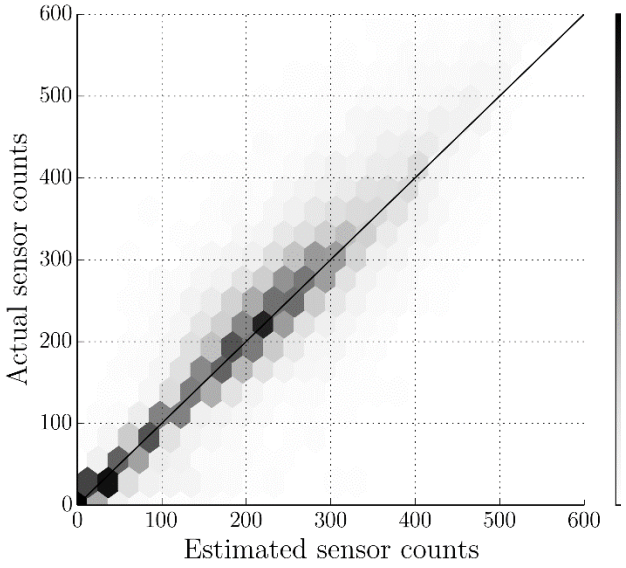


(e) Day 5

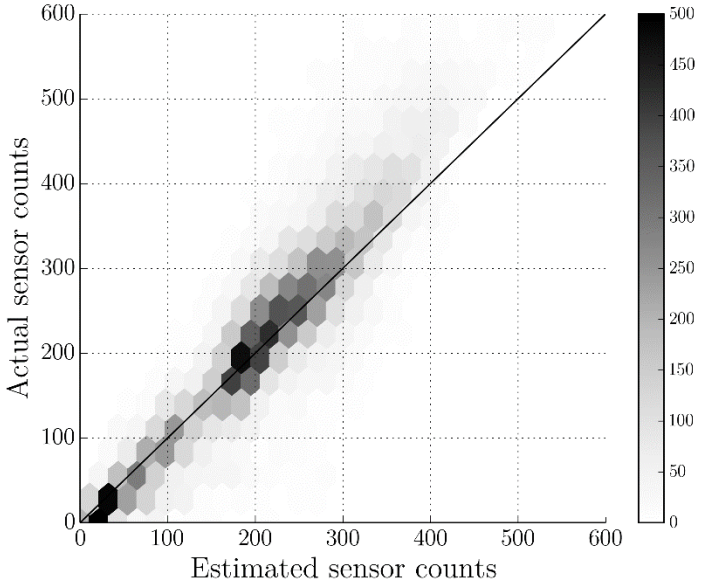


(f) Average across days

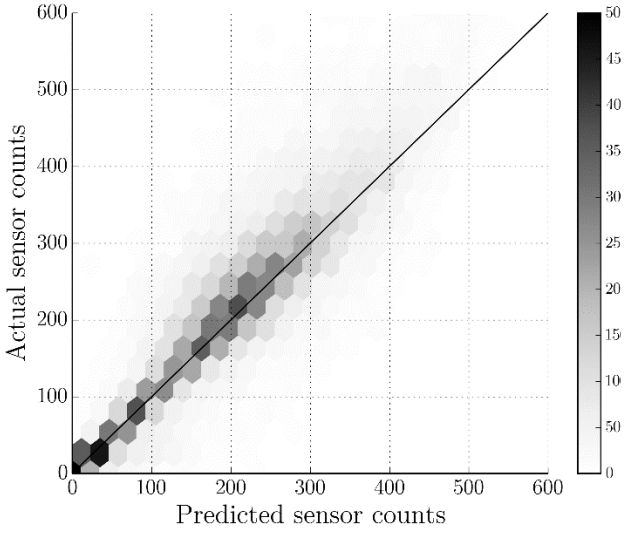
FIGURE 5 Plots of Sensor Count RMSNs for GLS and PC–GLS with Respect to Time–of–day in the Prediction Interval



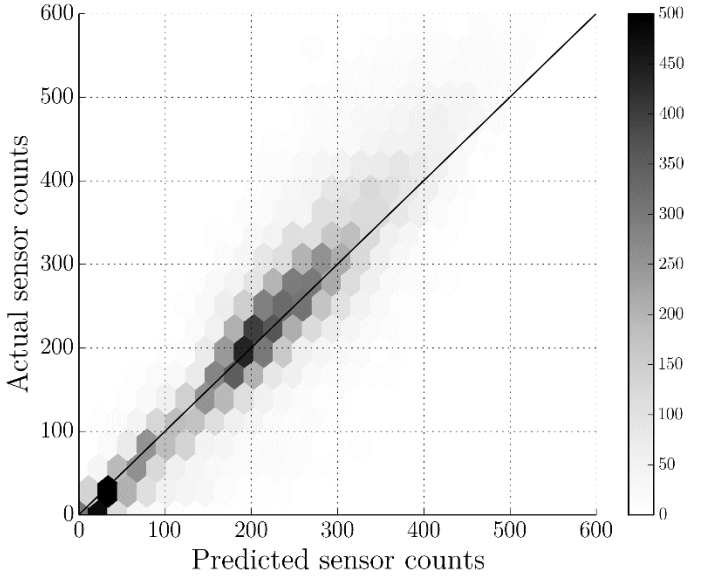
(e) GLS Estimation



(g) PC-GLS Estimation

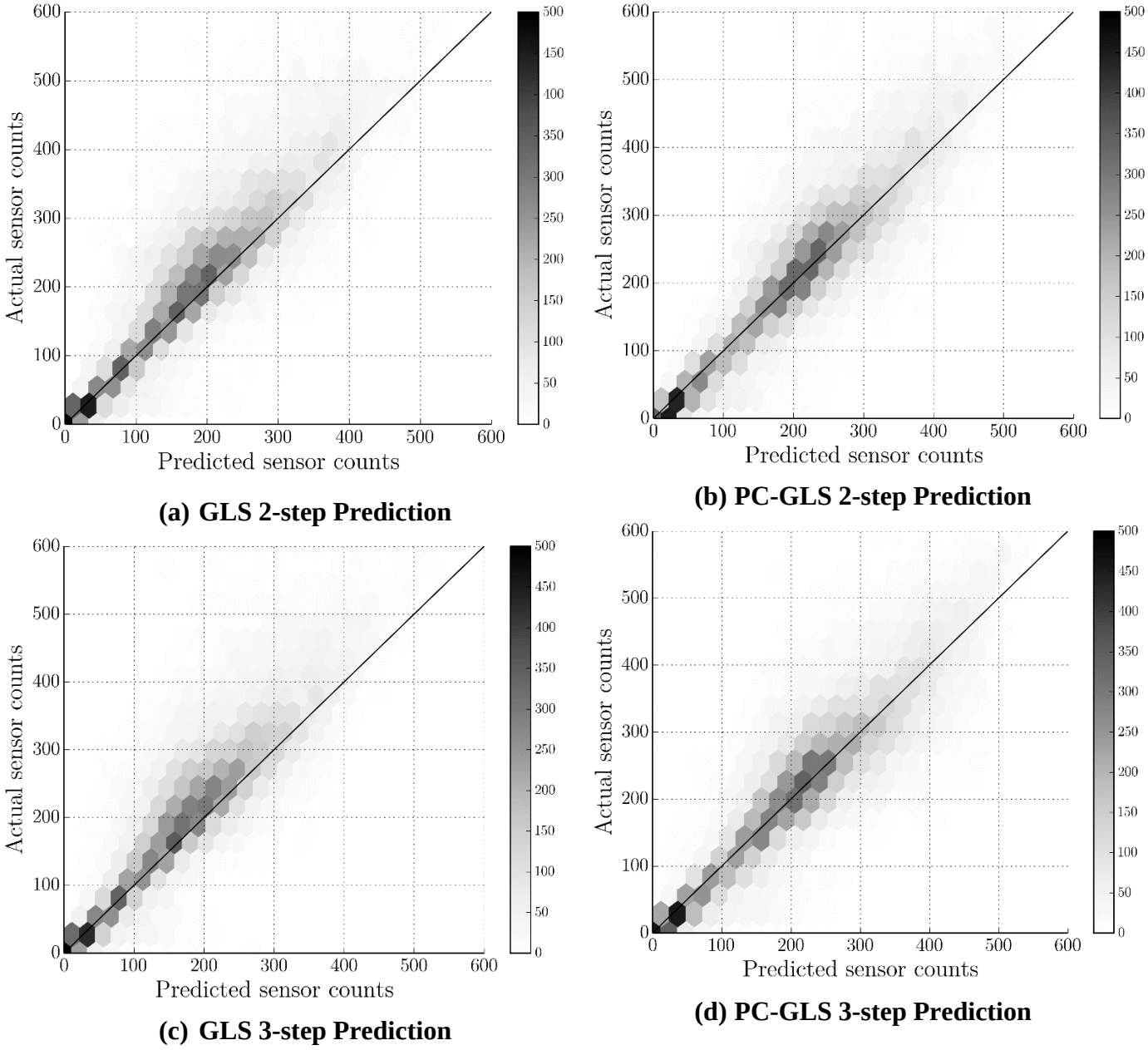


(f) GLS 1-step Prediction



(h) PC-GLS 1-step Prediction

FIGURE 6 Comparison of estimation and 1 step predictions of GLS and PC–GLS procedures through the scatter plots of 5 minute estimated/predicted vs. actual sensor counts in day 2. The darker the cell, higher the number of points in it.



(a) GLS 2-step Prediction

(b) PC-GLS 2-step Prediction

(c) GLS 3-step Prediction

(d) PC-GLS 3-step Prediction

FIGURE 7 Comparison of 2 step and 3 step predictions of GLS and PC–GLS procedures through the scatter plots of 5 minute predicted vs. actual sensor counts in day 2. The darker the cell, higher the number of points in it.