PC-SPSA: Exploration and assessment of different historical data-set generation methods for enhanced DTA model calibration

Moeid Qurashi[1, *], Qing-Long Lu[1], Guido Cantelmo[1], and Constantinos Antoniou[1]

[1]Department of Civil, Geo and Environmental Engineering, Technical University of Munich University, Germany
[*]Corresponding author: moeid.qurashi@tum.de

1 Introduction

Dynamic Traffic Assignment (*DTA*) model calibration has long been a relevant research topic within the transport community due to its complexity and non-linearity, which increases exponentially with the sizes of the network. DTA calibration leverages traffic measures, such as link counts and speeds, to estimate the most likely distribution of parameters (route choice parameters, link capacity, Origin-Destination demand flows – ODs) able to simulate the observed traffic conditions. The main problem comes from the fact that the number of parameters to be calibrated usually far exceeds the number of available traffic data (Marzano, Papola, & Simonelli, 2009). As a consequence, the problem becomes highly undetermined, prone to over-fitting, and, in many cases, unbounded in terms of error (Yang, Iida, & Sasaki, 1991; Cantelmo, Viti, Cipriani, & Nigro, 2018). Recently, a novel procedure has been widely adopted for solving both the on-line and the off-line problem, the Principal Component Analysis (*PCA*) (Djukic, Van Lint, & Hoogendoorn, 2012; Prakash, Seshadri, Antoniou, Pereira, & Ben-Akiva, 2018; Qurashi, Ma, Chaniotakis, & Antoniou, 2019). PCA is a technique that, given a series of observations, brings out strong patterns and correlations within data. Simply stated, if a certain number of observations about the historical OD demand is available, PCA will return its correlations (or principal components), meaning a reduced number of variables that can explain the variance of the unknown variable (the OD demand). While powerful and intuitive, the PCA requires a data-set to extrapolate these patterns. This last point leads to two strong research questions:

1. Is it possible to use PCA without a historical database?
2. PCA has been widely tested in controlled scenarios, where demand matrices are artificially created through a certain probability function. Would PCA still provide reasonable estimations if wrong patterns are used?

Among the several studies that deployed PCA, an algorithm named 'PC-SPSA' (Qurashi et al., 2019) has been proposed that significantly improved upon previously defined SPSA (Spall, 1998; Balakrishna, Antoniou, Ben-Akiva, Koutsopoulos, & Wen, 2007; Cantelmo, Cipriani, Gemma, & Nigro, 2014; Antoniou, Azevedo, Lu, Pereira, & Ben-Akiva, 2015; Lu, Xu, Antoniou, & Ben-Akiva, 2015). PC-SPSA, with the application of PCA, reduces the problem dimensions within the variance present among the historical estimates limiting SPSA search space for faster and efficient calibration. Within this research, we aim to understand and explore possible generation methods of historical estimates with different correlations among time-dependent OD flows in three different dimensions:

- **Spatial correlation among OD pairs:** We analyze how the principal components (PC) change when different assumptions on the spatial structure of the demand are performed.
- **Temporal correlation among OD pairs:** We analyze temporal correlations of the OD flows and fluctuation of the demand from a one-time interval to the other.

- **Day to day correlation among OD pairs:** Mobility demand is correlated to the demand for activities. As such, it follows a structure. Day to day variations are likely to occur and can be considered within the model.

These probability functions will be used for generating a series of scenarios and assess how the model performs when erroneous assumptions are done. Additionally, these distributions have the advantage of capturing (up to a certain extent) user behaviors. This means that alternative data sources (travel surveys, mobile phone network data) may be used to capture some of the parameters of these distributions and generate a more correct set of data for the PCA.

The remainder of the document is structured as follows. Section 2 briefly describes PCA implementation, SPSA and PC-SPSA. Then, Section 2.4 introduces the new probability functions that will be adopted to produce different historical data–sets for the PCA. Finally, Sections 3-4 introduce our case study and results, while the conclusion is discussed in Section 5.

2 Methodology

In this section, we introduce our methodology for DTA calibration. The following table reports the most relevant notations.

Table 1: List of Symbols

D	Historical data matrix
x	Current/prior OD estimate
D_{ij}^t, x_{ij}^t	OD pair between zone i to j at time interval t
n_{ij}, n_t, n_d	Number of OD pairs, time intervals and historical days
δ_{rand}	Randomly generated number
δ_{od}	Normally distributed random vector of size equal to OD vector with $\mu = 0$ and $\sigma = 0.333$
δ_t	Normally distributed random vector of size equal to total time intervals with $\mu = 0$ and $\sigma = 0.333$
δ_d	Normally distributed random vector of size equal to total historical estimated (days) with $\mu = 0.5$ and $\sigma = 0.08325$
R_d, R_{od}, R_t	Factor/weight coefficients for days, OD and time interval based randomness values

2.1 PCA implementation

PCA is implemented as per (Qurashi et al., 2019). Singular value decomposition (SVD) is applied to the historical data matrix D to evaluate its principal components (PCs) as:

$$D = U\Sigma V^T \tag{1}$$

Where columns of the $n_{ij} \times n_{ij}$ unitary matrix V present orthogonal PCs, with their corresponding PC-scores stored in the rectangular-diagonal matrix Σ with dimension $n_d \times n_{ij}$. U is a $n_p \times n_p$ unitary matrix with orthogonal vectors .V is reduced to $\hat{V}$, where only the first few significant PCs n_v are retained:

$$\hat{V} = [\, v_1 \; v_2 \; v_3 \; ... \; v_{n_v} \,] \tag{2}$$

The new matrix $\hat{V}$ is then used to reduce our starting OD flows vector x into PC scores z of dimension $[n_v \times 1]$, as:

$$z = \hat{V}^T x \tag{3}$$

Furthermore, the OD vector can be approximated as:

$$x \approx \hat{V} z \tag{4}$$

2.2 Simultaneous perturbation stochastic approximation (SPSA)

As defined by (Spall, 1998), SPSA randomly perturbs its set of estimation variables θ (equation 5) by a perturbation coefficient c_k and Δ (± 1 bernoulli distribution random vector) to evaluate a random numerical gradient g (equation 6) and later minimize the estimation variables θ by the evaluated gradient and a minimization coefficient a_k (equation 7). The function $f(\theta)$ in equation 6 captures the error associated to a set of parameters θ.

$$\theta^{\pm} = \theta_k \pm \theta_k \times c_k \Delta \tag{5}$$

$$g' = \frac{f(\theta^+) - f(\theta^-)}{2c_k} \left[\Delta_1 \ \Delta_2 \ . \ . \ \Delta_h\right]^T \tag{6}$$

$$\theta_{k+1} = \theta_k - a_k {g'}_k(\theta_k) \tag{7}$$

2.3 PC-SPSA

Within PC-SPSA, PC-scores vector z resulted from the implementation of PCA on the OD flows vector x are calibrated instead of the OD flows vector x itself, using a modified SPSA algorithm settings (as per (Qurashi et al., 2019)). The two modifications include: 1) Replacing the estimation variables θ from OD flows vector x to its PC-scores z, 2) Modified steps of perturbation and minimization from equation 5 and 7 to equation 8 and 9.

$$\text{Perturbation:} \qquad z^{\pm} = z_k \pm z_k \times c_k \Delta \tag{8}$$

$$\text{Minimization:} \qquad z_{k+1} = z_k - z_k \times a_k g' \tag{9}$$

2.4 Historical matrix estimation

Using PCA along with SPSA limits it's search space within the patterns/correlations captured by the estimated PCs. Hence, the relevance of historical estimates with the targeted estimate is crucial, as if, the patterns of the target solution are not present within the variance of historical estimates, PC-SPSA will not be able to provide a good quality solution. To exploit the three dimensions of correlation covering possible user behaviors, historical estimates are generated using 6 different methods.

1. Spatial correlation:

$$D = \sum_{d=1}^{n_d} \sum_{t=1}^{n_t} D^t = \sum_{d=1}^{n_d} \sum_{t=1}^{n_t} (1 + R_{od}\delta_{od}) \times x^t \tag{10}$$

2. Temporal correlation:

$$D = \sum_{d=1}^{n_d} \sum_{ij=1}^{n_{ij}} D_{ij} = \sum_{d=1}^{n_d} \sum_{ij=1}^{n_{ij}} (1 + R_t\delta_t) \times x_{ij} \tag{11}$$

3. Spatial and temporal correlation:

$$D = \sum_{d=1}^{n_d} D = \sum_{d=1}^{n_d} (1 + R_{od}\delta_{od\times t}) \times x \tag{12}$$

4. Spatial and day-to-day correlation:

$$D = \sum_{t=1}^{n_t} D^t = \sum_{t=1}^{n_t} (1 + R_{od}\delta_{od\times d}) \times x^t \tag{13}$$

5. Temporal and day-to-day correlation:

$$D = \sum_{ij=1}^{n_{ij}} D_{ij} = \sum_{ij=1}^{n_{ij}} (1 + R_{od}\delta_{t\times d}) \times x_{ij} \tag{14}$$

6. Spatial, temporal and day-to-day correlation:

$$D = \sum_{d=1}^{n_d} D = \sum_{d=1}^{n_d} (1 + R_{od}\delta_{d}\delta_{od\times t}) \times x \tag{15}$$

Each of the above-mentioned estimation methods is generated through a certain correlation setting between the three identified correlation dimensions and result in a historical data matrix D providing the required PC-directions to PC-SPSA for estimation of the target solution.

3 Experimental setup

3.1 Network

The urban network of Munich city center (shown in figure 1) is used to set up the calibration case study in an open-source traffic simulator SUMO (Lopez et al., 2018). The network consists of 2605 edge links with 564 detector locations and the demand of the morning peak between 7 to 10 am is represented in 15 min intervals with an OD matrix of [61 × 61] or 3721 OD pairs. The simulations are set up in the mesoscopic resolution with trip-based (one-shot) stochastic user route choice assignment.

3.2 Calibration scenario

As per the guidelines from (Antoniou et al., 2016), to set up the scenarios for calibration, a network state with its demand and counts is considered as a true network state. Further, the scenarios and historical estimates are created considering the true network state demand as x. The most appropriate and probable scenario that capture the user behavior is generated using the probability functions for spatial and temporal correlations, i.e:

$$X = (R_d + R_{od}\delta_{od\times t}) \times x \tag{16}$$

4 Results

To explore and compare the proposed set of historical data-set generation methods, we choose the fourth scenario (equation 16) creation technique i.e. with spatial-temporal correlation (considering it to be the most correlated/probable scenario to replicate user behavior). Results with the remaining functions will be shown during the hEART2020 conference. The scenario is then calibrated using 6 historical data matrices with PC-SPSA (methods showed in equations 10,-15).

Performance evaluation of such calibration techniques requires three major performance indicators, given as:

1. Best goodness-of-fit between calibrated and target OD matrix [Figure 2 (left)].
2. Best goodness-of-fit between observed and measured counts [Figure 2 (right)].
3. Best convergence performance over the required number of iterations for different time intervals [Figure 3].

In reference to the first two performance indicators, method 6 (i.e. Equation 15, spatial, temporal and day to day correlation) and method 3 (i.e. Equation 12 spatial, temporal correlation) show the best performance for both, getting the best RMSN between observed and calibrated traffic counts and also better quality of the calibrated OD matrix in comparison to the target OD matrix (figure 2).

Figure 1: Network of Munich city center

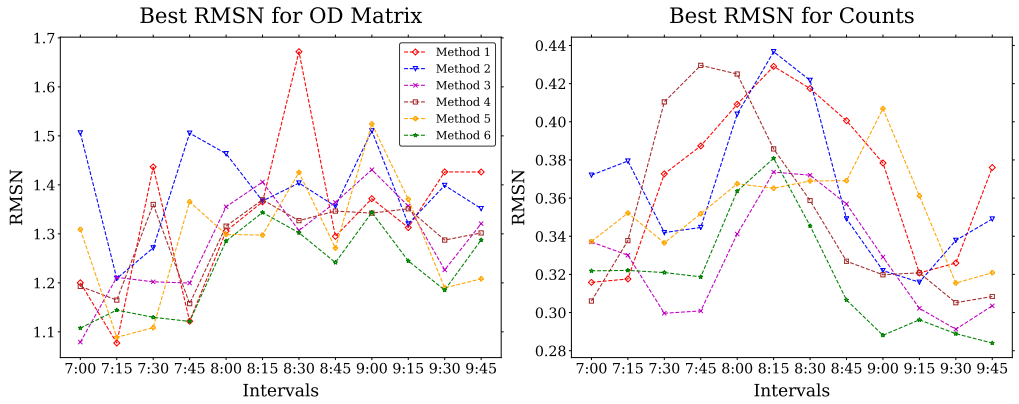

Figure 2: Comparison between all generation methods

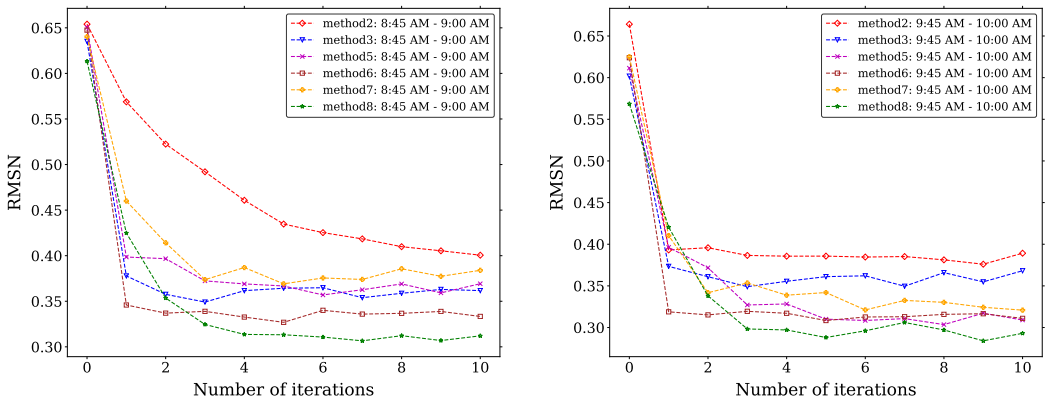

Figure 3: Comparison between generation methods for specific intervals

From figure 2, it's also evident that method 1 and 2 (generating the historical data-set with a single correlation either spatial or temporal, equation 10-11) are almost worst in performing consistently and are also almost worst in getting either the quality of the solution or the traffic counts RMSN.

The performance of method 3 and 6 evidently shows that the combination of spatial and temporal correlation is crucial for scenarios created with similar technique but method 6, which performs the best, also adds a day to day based correlation within the historical estimate providing somewhere more search space or variance for PC-SPSA to find a better solution and improve the overall calibration performance.

Figure 3 also depicts the convergence performance for all 6 historical data-set for two specific time intervals. The convergence plots also confirm method 6 based historical data-set to be the best against the third performance indicator.

5 Conclusion

This paper explores multiple historical data-set estimation methods which are crucial for the calibration performance for principal component analysis (PCA) based algorithms. We first propose multiple sets of historical data-set generation methods with probable calibration scenarios (which replicate more realistic changes within the structure of the demand) and later explore the performance of all the proposed historical data-sets with PC-SPSA to understand the importance of different historical data-set generation parameters. As per the current results, more correlatedly generated historical estimates (i.e. method 3 and 6) outperform other simplified techniques but it will be further interesting to explore and analyze their performance calibrating other different sets of scenarios.

Next steps, to be shown in hEART2020 conference, will include, the exploration of all the proposed methods on the possible demand scenarios to identify the best most generically well-performing data-set generation technique, and later validating that technique on a larger network of Munich city (with a network of 8689 links, 706 detector location and demand of OD matrix [73 $\times$ 73] or 5329 OD pairs) with different demand scenarios and also other network information e.g. travel times etc.

Finally, results proposed in this study are still based on synthetic experiments. This is a limitation, as we aim to test PCA based algorithms when historical data sets are not available (or information is not reliable). To do so, we will use real traffic data from Munich to generate a benchmark scenario that is assumption free - e.g. the "true" network state is derived from real data and not from syntetic functions. This will allow us to validate our probability function against real data in an assumption free scenario.

Acknowledgments

This research has been supported by the German Research Foundation - Deutsche Forschungsgemeinschaft (DFG) [Project number 392047120, Research grants in collaboration with China].

References

Antoniou, C., Azevedo, C. L., Lu, L., Pereira, F., & Ben-Akiva, M. (2015). W-spsa in practice: Approximation of weight matrices and calibration of traffic simulation models. *Transportation Research Part C: Emerging Technologies*, *59*, 129–146.

Antoniou, C., Barceló, J., Breen, M., Bullejos, M., Casas, J., Cipriani, E., ... van Lint, H. (2016). Towards a generic benchmarking platform for origin-destination flows estimation/updating algorithms: Design, demonstration and validation. *Transportation Research Part C: Emerging Technologies*, *66*, 79–98.

Balakrishna, R., Antoniou, C., Ben-Akiva, M., Koutsopoulos, H. N., & Wen, Y. (2007). Calibration of microscopic traffic simulation models: Methods and application. *Transportation Research Record*, *1999*(1), 198–207.

Cantelmo, G., Cipriani, E., Gemma, A., & Nigro, M. (2014). An adaptive bi-level gradient procedure for the estimation of dynamic traffic demand. *IEEE Transactions on Intelligent Transportation Systems*, *15*(3), 1348–1361.

Cantelmo, G., Viti, F., Cipriani, E., & Nigro, M. (2018). A utility-based dynamic demand estimation model that explicitly accounts for activity scheduling and duration. *Transportation Research Part A: Policy and Practice*, *114*, 303–320.

Djukic, T., Van Lint, J., & Hoogendoorn, S. (2012). Application of principal component analysis to predict dynamic origin–destination matrices. *Transportation research record*, *2283*(1), 81–89.

Lopez, P. A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., ... Wießner, E. (2018). Microscopic traffic simulation using sumo. In *The 21st ieee international conference on intelligent transportation systems.* IEEE. Retrieved from <https://elib.dlr.de/124092/

Lu, L., Xu, Y., Antoniou, C., & Ben-Akiva, M. (2015). An enhanced spsa algorithm for the calibration of dynamic traffic assignment models. *Transportation Research Part C: Emerging Technologies*, *51*, 149–166.

Marzano, V., Papola, A., & Simonelli, F. (2009). Limits and perspectives of effective o–d matrix correction using traffic counts. *Transportation Research Part C: Emerging Technologies*, *17*(2), 120–132.

Prakash, A. A., Seshadri, R., Antoniou, C., Pereira, F. C., & Ben-Akiva, M. (2018). Improving scalability of generic online calibration for real-time dynamic traffic assignment systems. *Transportation Research Record*, *2672*(48), 79–92.

Qurashi, M., Ma, T., Chaniotakis, E., & Antoniou, C. (2019). Pc-spsa: Employing dimensionality reduction to limit spsa search noise in dta model calibration. *IEEE Transactions on Intelligent Transportation Systems*.

Spall, J. C. (1998). Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Transactions on aerospace and electronic systems*, *34*(3), 817–823.

Yang, H., Iida, Y., & Sasaki, T. (1991). An analysis of the reliability of an origin-destination trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological*, *25*(5), 351–363.