Australasian Transport Research Forum 2017 Proceedings
27 – 29 November 2017, Auckland, New Zealand
Publication website: http://www.atrf.info

Modified bi-level optimization framework for dynamic OD demand estimation in the congested networks

Tamara Djukic[1], David Masip[1], Martijn Breen[1], Jordi Casas[1]

[1]TSS-Transport Simulation Systems, Ronda Universitat 22B, Barcelona 08007, Spain

Email for correspondence: tamara.djukic@aimsun.com

Abstract

The common approach usually adopted in dynamic OD demand estimation and prediction consists in solving an optimization problem in which the distance between observed and simulated traffic conditions is minimized by assuming the relationship between OD flows and traffic observations independent of traffic conditions in the network. This approach has a sever shortcoming as it does not take into account the impact of demand flows variation on traffic observations in congested networks.

This paper presents a modified bi-level optimization framework to solve the high-dimensionality of nonlinear OD estimation problem by adding the recursive step to overcome the divergence of the optimization function performance. The recursive step involves evaluation of the marginal effects for the subset of significant OD pairs whose variation leads to high changes in link-flow proportions and traffic flows. In this way, we reduce the number of optimization function evaluations that allows the modeller to control the trade-off between simplicity of the model and the level of realism for large-scale, congested networks. Several state-of-the-art, bi-level optimization solution approaches, are used in the performance assessment study. All the approaches are implemented in a mesoscopic traffic simulation tool, Aimsun, to perform the traffic network loading on the large-scale network: the Vitoria urban network with 3249 OD pairs, 389 detectors, and 600km road network. The results demonstrate the applicability of the proposed solution approach to efficiently obtain dynamic OD demand estimates for large-scale, congested networks and that, computationally, this approach outperforms existing methods.

1. Introduction

Transport authorities and practitioners have long been concerned about the unavailability of reliable dynamic OD demand estimates which limits the potential for dynamic traffic assignment (DTA) deployments to analyse and alleviate traffic congestion as part of the Intelligent Transportation Systems (ITS). In congested networks, changes in the demand affect travel times. In turn, travel times affect the route choices and travel times from trip origin nodes to traffic observation detectors that determine the link-flow proportions or the relationship between OD flows and traffic observations. Modelling of non-linear relationship between traffic observations and OD flows, and its dependency on variations in OD flows has been identified by many researchers as a key challenge in the estimation and prediction of a high-quality OD matrices. For example, the dependence of link-flow proportions on the demand flows in assignment matrix should be explicitly included in the DTA process. Finding derivatives of link-flow proportions and traffic observation with respect to demand flows can be cumbersome task, often judged not feasible in terms of computation time.

From a modelling point of view, the most distinguishing difference between the OD demand estimation approaches, is how the relationship between state variables (e.g., OD flows, OD proportions) and any available traffic data (e.g., link traffic counts, speeds) is defined, calculated and re-calculated throughout the estimation process. An accurate description of this relationship leads to an accurate description of traffic state reality in the network, but to more complexity as well. In the past decades, a rich body of literature, stressed the need of relaxing the fixed relationship assumptions in mapping demand flows to traffic observations through estimation process when the congestion occurs in the network. Researchers have been devoted to development of the methods to capture the impact of demand variation on traffic observations, that can be categorized in analytical derivation, simulation-, numerical- and heuristic-based approximation methods.

Typically, to calculate the weights between OD flows and link traffic counts (usually measured by loop detectors in the form of sensor counts), dynamic link-flow proportions, expressed in the assignment matrix form, are commonly used. Theoretically, these link-flow proportions can be analytically derived using network topology, path choice set, current route choice model and equilibrium travel times (Ashok and Ben-Akiva 2002). However, it is recognised that the complexity of the problem at hand can quickly lead to intractable situations (Ashok and Ben-Akiva 2002). Further sophisticated analytical derivations are required to capture the relationship between parameters with less direct impact and non-linear relationship.

The simulation-based approximation of the relationship between demand flows and traffic observations uses traffic simulator to uncover this relationship without the direct derivation of the assignment matrix. The most studied assignment matrix-free method is the Simultaneous Perturbation Stochastic Approximation (SPSA) method (Balakrishna and Ben-Akiva 2007; Ben-Akiva, Koutsopoulos and Mukundan 1994; Cipriani et al. 2011; Cipriani, Gemma and Nigro 2013; Cantelmo et al. 2014; Antoniou et al. 2015) which allows one to approximate a descent gradient direction with significantly lower computational resources than through the explicit calculation. Antoniou et al. (2015) proposed the Weighted-SPSA (W-SPSA) algorithm to overcome the deteriorated performance of the gradient calculated by SPSA algorithm. Although, the main advantage of such method is that the complex relationship between demand flows and traffic observations, not only traffic counts, is estimated by simulation model, the high number of simulation runs for large-scale networks where the DTA is computationally intensive still remains to be resolved. For example, due to stochasticity of the simulation model, for each perturbation of the SPSA or W-SPSA where gradient needs to be determined, the DTA has to be replicated R times, leading to 2R runs (Antoniou et al. 2015).

Recent studies by (Toledo and Kolechkina 2013; Frederix, Viti and Tampére 2013; Shafiei et al. 2017) rely on linear approximation of the assignment matrix with non-separable response in every iteration, which relaxes the assumption of constant link-flow proportions and explicitly accounts the congestion effects. This definition requires the computation of the marginal effects of demand flow change on the link-flow proportions at the current solution of each iteration. It is possible to use the finite differences approach to numerically approximate the Jacobian matrix by using traffic simulator, but it would require in every iteration of the gradient solution to perturb each element in the OD demand vector, one at a time, leading to 2RDK runs, where D the number of OD pairs in the network and K the number of time intervals for the simulation period. To overcome computational overhead, authors proposed heuristic-based approaches. Toledo and Kolechkina (2013) neglected the effect of changes in one OD pair over the other OD pairs in the assignment matrix, Frederix, Viti and Tampére (2013) implemented space decomposition of the network in the congested and non-congested sub-networks, where derivatives were computed only for congested area. Shafiei et al. (2017) reduces computation costs through iterations progress and computing derivatives on OD pairs whose flows have higher tendency to vary during dynamic OD demand estimation process. However, all these approaches rely on strong heuristic assumptions such as ignoring the effect of OD demand changes outside of congested area or have been tested on relatively small and medium sized networks. Further research is necessary to develop solution approach for

nonlinear OD estimation problem that will guaranty its reliability and computational efficiency in the large-scale networks.

Here we extend the previous work by proposing the modified bi-level optimization solution approach to estimate dynamic OD demand for large-scale networks that accounts non-linear relationship between traffic observations and OD flows. Non-linear relationship has been computed for the subset of the OD pairs when performance of the objective function has been deteriorated. Reducing the problem dimensionality through selection of the most significant OD pairs replaces the conventional approach of computing derivatives for all OD pairs, whether through all the iteration steps or in the initial step. The importance of this approach lies in the possibility to capture the most important effects of congestion by relaxing assumption of constant link-flow proportions without loss of accuracy and considerable decreasing the model dimensionality and computational complexity. In this study, mesoscopic traffic simulation model in Aimsun is employed as a DTA traffic simulation tool which realistically captures the congestion phenomena that is more adequate in developing dynamic OD estimation algorithms.

The paper is organized as follows. In the first part of the paper, we summarize the main challenges in defining the non-linear relationship between traffic observations and OD demand in dynamic OD demand estimation problem. In the second part of the paper, we present the modified bi-level optimization framework of the OD demand estimation model with additional recursive step to overcome the divergence of the optimization function performance. Next, we explore the properties of reducing the number of optimization function evaluations by defining the subset of significant OD pairs whose variation leads to high changes in link-flow proportions and traffic flows. In the third part of this paper, we demonstrate the performance of the proposed OD estimation model on a large-scale network, Vitoria, Spain. The paper closes with a discussion on further application perspectives of the OD demand estimation model and further research directions.

2. The problem formulation

This section describes the most critical issue in OD matrix estimation, whether static or dynamic, the relationship (mapping) of the observed traffic condition data with unobserved OD flows. From a modelling point of view, the most distinguishing difference between the OD demand estimation approaches presented in the literature, is how the relationship between OD flows and any available traffic data (e.g. link traffic counts, speeds, densities, etc.) is defined, calculated and re-calculated throughout the estimation process. This relationship is often accomplished by means of an assignment matrix. In the dynamic problem, the assignment matrix depends on link and path travel times and traveller route choice fractions - all of which are time-varying, and the result of dynamic network loading models and route choice models. These dynamics are reflected in travel times between each origin and destination trips on a network, influenced by traffic link flow. While a vast body of literature has been developed in this area over the past two decades, this section focuses on some of the efforts that highlight the basic problem dimensions.

The general OD estimation problem is to find an estimate of OD demand matrix by effectively utilizing traffic and demand observations. Let $\Omega \subseteq N \times N$ be set of all n OD pairs in the network, and $\mathrm{L}' \subseteq L$ be the set of l links where traffic data observations are available. The time horizon under consideration is discretised into R time intervals of equal duration, indexed by $r = 1, 2, \ldots, R$. The OD matrix, $X = \{x_{nr}\}$, defines the demand for each OD pair $n \in N$ with departure time interval $r \in R$. Prior information on the OD matrix is defined by $\tilde{X} = \{\tilde{x}_{nr}\}$. The vector $\tilde{y} = \{\tilde{y}_{lt}\}$ defines traffic flow observations for time interval $t = 1, 2, \ldots, T$, for each link in L'. Here the dynamic OD demand is represented by a vector, rather than a matrix. It is also assumed that T and R describe the same length of time interval, but their decomposition to time intervals can be different.

The dynamic OD estimation problem can be formulated as a constraint optimization problem (Cascetta 1984) as:

$$\min_{x \geq 0} Z(x) = \alpha f(\mathrm{x}, \tilde{\mathrm{x}}) + (1 - \alpha) f(\mathrm{y}, \tilde{\mathrm{y}}) \quad (1)$$

Regardless of the function f used, the purpose is to obtain an OD demand that yields OD flows and traffic data as closely as possible to their observed values. When solving the OD problem in Equation (1) the relationship between traffic observations and OD demand has to be defined, implicitly or explicitly. Most dynamic OD demand estimation methods, define this relationship implicitly by the traffic assignment model that can be expressed as:

$$\hat{y}_t = \sum_{h=r-\kappa}^{r} A_t^h x_r \quad (2)$$

There are two main drawbacks of relationship defined in Equation (2):

1. *Separability of traffic count observations*: it assumes that the traffic flow observed at the link l during time interval t can always and only be changed by changing one of the OD flows that passes link l in time interval h when x_k is assigned in the network. This assumption of separability is incompatible with some typical phenomena in congested networks, such as congestion spillback between links and time lags due to the delay during congestion. In these cases, it is very likely that increasing an OD flow will cause delays to other flows that do not pass that time-space interval, hereby altering the amount of flow passing the link in the considered time interval. This issue has been addressed in past studies (Yang and Zhou 1998; Tavana and Mahmassani 2001; Lundgren and Peterson 2008). Frederix, Viti and Tampére (2013) suggested using the Taylor approximation to specify the linear approximation of Equation (2) using non-separable response function, given by

$$\hat{y}_t = \sum_{h=r-\kappa}^{r} A_t^h(x_0) x_r + \sum_{h=r-\kappa'}^{r} (x_h - x_0) \left[\sum_{h'=r-\kappa'}^{r} \frac{d(A_t^{h'}(x_{h'}))}{dx_{h'}} x_{h'} \right] \quad (3)$$

2. *Limited only to one data source*: formulation of relationship by assignment matrix in Equation (2) and Equation (3) restricts dynamic OD demand estimation problem to use of traffic count data only, which can potentially over-fit to counts at the expense of traffic dynamics. Relationship between traffic condition data, such as speeds and densities, and OD flows are expected to be non-linear and approximations similar to the assignment matrix cannot be justify (Balakrishna and Koutsopoulos 2008). This issue has been addressed in the past studies (Balakrishna and Koutsopoulos 2008; Cipriani et al. 2011; Cantelmo et al. 2014; Antoniou et al. 2015) who proposed use of traffic simulation models to capture the nonlinear relationship between OD flows and traffic observations instead of the assignment matrix.

Although presented solutions significantly contributed to quality improvement of dynamic OD demand estimates, they still share a common challenge to overcome high computational costs. A complicating factor in utilizing these methods for estimation or prediction purposes, is that OD matrices are very large data structures, that grows rapidly in large networks. Even in case such high-dimensional OD flows can be reduced (see e.g. (Djukic et al. 2012) and this is not entirely unlikely, there are serious methodological difficulties in finding optimal solutions (e.g. getting stuck in local minima, slow convergence, high number of simulation runs, etc.), aside from the computational and memory requirements for such a procedure on the basis of thousands (to millions) of traffic observations. For example, computing the exact Jacobian vector in the second term of Equation (3) with respect to changes in OD flows for each OD pair remains intractable even when efficient, well calibrated, DTA model is used.

In the next section, we propose a different solution approach for exploring the relationship between OD flows and traffic observations by applying modified bi-level optimization framework. We relax assumption to rely on the assignment matrix from DTA by evaluating marginal effect of demand flow changes on traffic conditions in the network for the subset of the OD pairs.

3. Methodology

Solution algorithms proposed in literature, to solve the problem given in Equation (1), incorporate computing the marginal effects of demand changes on traffic observations that lead to high computational costs for medium- or large-scale networks. In this situation, dimensionality reduction of simulation runs required to capture these marginal effects is necessary, leading to improve computational performance.

In order to overcome problem related to dimensionality of OD demand problem, we propose the following heuristic approach. First of all, we propose the use of Equation (3) to capture marginal effects with respect to changes in OD flows, rather than using a more conventional approach with constant assignment proportions given by Equation (2). Secondly, we propose to use Equation (3) on the subset of the OD pairs, whose variation in demand creates the divergence of the cost function given by the objective function defined by Equation (1). Thirdly, we suggest using an initial OD matrix that produces the same congestion pattern as is observed in reality, i.e. that allows one to start with correct traffic regime.

It is convenient to start the presentation of the proposed solution approach with reference to the idea of OD demand estimation problem formulation, as a bi-level optimization framework. Then, we provide a modified bi-level optimization framework, with recursive step to account the marginal effects of the link-flow proportions for a selected set of OD pairs.

3.1 Standard OD model formulation in bi-level optimization framework

The dynamic OD demand estimation problem can be defined as a bi-level optimization framework. The main advantage of using the bi-level formulation is the ability to capture the network congestion effects in the dynamic OD demand estimation problem, as the traffic assignment model can be defined as an optimization problem in itself. The upper level is formulated as an ordinary least square (OLS) problem, which estimates the dynamic OD demand based on the given link-flow proportions. Assuming that errors are independently and identically normally distributed, the objective function aims to minimize the square distance between computed and observed traffic flows, as well as the estimated and prior OD demand matrix, defined in Equation (4) as follows:

$$\min_{x \geq 0} Z(x) = \|A(x)x - \tilde{y}\|^2 + \alpha\|x - \tilde{x}\|^2 \quad (4)$$

subject to

$$y = DTA(x)$$

The lower level is formulated as DTA problem which estimates the elements of the assignment matrix. These elements are a function of the OD flows, and they represent both the propagation of the OD flows, the departure time and the route choice decisions related to an OD flow. The link-flow proportions are, in turn, generated from the dynamic traffic network loading problem at the lower level, which can be solved through a simulation-based DTA procedure. In this paper, simulation based DTA model in Aimsun software TSS-Transport Simulation Systems (2015) has been used. Route choice set is a result of the dynamic user equilibrium. The dynamic network loading has been performed at the mesoscopic level in Aimsun, synthesizing microscopic and macroscopic traffic flow propagation modelling

concepts. Here we assume that the entire set of link traffic counts for the analysis period, $L' \times T$, is used to simultaneously estimate OD demand for all time intervals, $N \times R$.

In general terms, all dynamic OD demand estimation methods defined as a bi-level optimization problem aim to find the most probable OD matrix by iteratively solving problems defined at upper and lower-level. The iterative solution algorithm is given as follows:

Step 1. *Initialization*. Initiate prior OD demand matrix, set $k = 0$.

Step 2. *Assignment*. Assign the demand to the network to obtain assignment matrix, A_k^h and estimated link traffic counts on the links with traffic observations, by Equation (2) or Equation (3).

Step 3. *Convergence test*. Check objective function value convergence. If objective function value has converged, stop and accept the current demand. Otherwise, proceed to step 4.

Step 4. *Update OD demand*. Estimate OD demand with link flows obtained from DTA, as given by Equation (2). Go to step 2, $k = k + 1$.

When non-separability of traffic observations is considered, as we shown in the previous section, Equation (3) has to be applied in Step 2 to capture the marginal effects of the demand variation on the changes in traffic flow observations. Traffic assignment relationship given in Equation (3) can be solved by computing the numerical derivatives using finite or central differences method of the traffic link flows with respect to changes in all OD pairs. This requires perturbing each OD pair in the OD demand two times, one at the time, resulting in $2NR$ traffic simulation runs and objective function evaluations per iteration step. It is obvious that such solution approach will result in computationally expensive tasks, that has to be overcome.

3.2 Modified bi-level optimization framework

The bi-level optimization framework presented in the previous section is modified to meet the following requirements for congested, large-scale networks:

- Relax assumption on fixed link-flow proportions from assignment matrix derived by DTA procedure by computing the marginal effects of the demand deviations on link flows given by Equation (3);
- Reduce number of OD variables in Equation (3) through the inclusion of only the OD pairs whose change in demand values cause significant deviations in the link flows;
- Keep the computational costs lower.

These requirements are implemented through the following modified iterative solution algorithm with recursive step:

Step 1. *Initialization*. Initiate prior OD demand matrix, set $k = 0$, $I' = \{\}$,

Step 2. *Assignment*. Assign the demand through DTA to the network to obtain assignment matrix, A_k^h and simulated link traffic counts on the links with traffic observations

Step 3. *Convergence test*. Check the objective function value with traffic flows and demand for convergence. If objective function value has converged, stop and accept the current demand. Otherwise, proceed to step 4.

Step 4. *OF performance test*. Check performance of the objective function value. If objective function decreases proceed to step 5. Otherwise, proceed to step 6, $k = k - 1$.

Step 5. *Update OD demand*. Estimate OD demand with link flows obtained from DTA, as given by Equation (2). Go to step 2, $k = k + 1$.

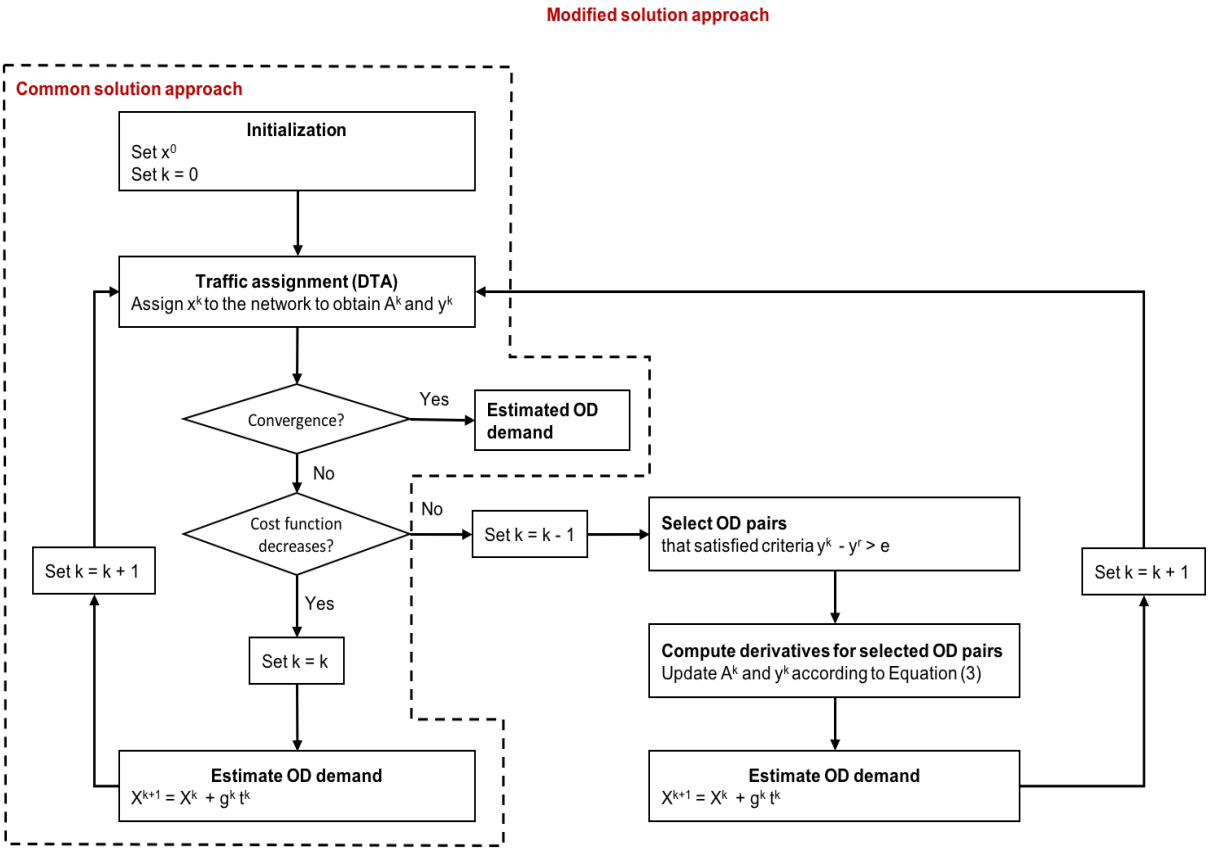
Step 6. *Select OD pairs.* Determine OD pairs whose variation has a considerable impact on link flows variation in the previous iteration and insert them in I'.

Step 7. *Update assignment*. Update the link-flow proportions in the assignment matrix, A^{k-1}, with values obtained from Equation (3) for the selected OD pairs in I'.

Step 8. *Update OD demand*. Estimate OD demand with link flows obtained from Equation (3). Go to step 2, $k = k + 1$.

The common and modified bi-level optimization framework with inputs and outputs for OD demand estimation is illustrated in Figure 1.

Figure 1. Generic algorithm based on the proposed solution



Note that the proposed solution algorithm in Steps 6 and 7 uses a demand with one iteration step latency because the updated demand at iteration step k caused the increase of the objective function value. Therefore, following the modified bi-level framework, Step 8 denotes correction of the state variable for iteration k, using the information from the link-flow proportions and link flows for iteration $k = k - 1$, obtained in Steps 6 and 7. In step 6, we analyse the change in the link flows obtained with demand assigned in iteration step k and $k = k - 1$ and determine the link flows with the highest variation. Using the information from the link flow proportions matrix, we can identify which OD flows are crossing these links with the highest flow variation and set them in I'. Then, in step 7 the elements of the link-flow proportion matrix are corrected for these OD pairs using the Equation (3). The upper-level problem in steps 5 and 8 is solved using the gradient decent method. The OD demand estimation results are evaluated in step 3 against termination criteria and the procedure would continue if termination criteria is not met.

Finally, the steps 6-8 reflect our corrected knowledge on the OD demand state at iteration $k = k - 1$ to improve the performance of the solution algorithm. Reducing the problem dimensionality through selection of the most significant OD pairs replaces the generic approach of computing derivatives for all OD pairs. The importance of this approach lies in the possibility to capture the most important effects of congestion by relaxing assumption of

constant link-flow proportions without loss of accuracy and considerable decreasing the model dimensionality and computational complexity.

3.3 Method for solving the upper-level problem

Let x_k be the demand at step k, and A_k and y_k the assignment matrix and simulated counts given by this demand. As an approximation to the OD estimation objective function given in Equation (4) that we want to minimize, we consider the auxiliary objective function

$$Z_k(x) = \|\tilde{y} - y_k - A_k(x - x_k)\|^2 + \alpha\|x - x_h\|^2 \quad (5)$$

There are different types of exact and heuristic methods proposed in the literature that can be employed to solve the optimization problem defined in Equation (5) with nonnegative variable constraints. At every external iteration, the gradient descent method is selected to minimize the above function, which uses the gradient as the search direction:

$$d = -\nabla Z_k \quad (6)$$

where

$$\nabla Z_k(x) = 2\,\alpha\,(x - \tilde{x}) + 2({A_k}^T A_k x - {A_k}^T \tilde{y} + {A_k}^T y_k - {A_k}^T A_k x_k) \quad (7)$$

To perform this gradient method, we start at $x = x_k$ and we perform N gradient steps, the direction of them being given by the latter equation. At internal step $n \leq N$, let us denote the estimated demand by x_k^n. After determining the search direction, which is given by $\nabla Z(x_k^n)$, the optimal step length, θ^n needs to be obtained in each iteration. To obtain it, we use the following criterion:

$$\theta^n = \min_{\theta^n} Z(x_k^n - \theta^n \nabla Z(x_k^n)) \quad (8)$$

The exact line search procedure proposed by Cauchy (1847) is used to compute the step size. In the case where Z is a quadratic function, the optimal step can be computed analytically. In our case, the optimal step size is computed using the following expression:

$$\theta^n = \frac{\|\nabla Z(x_k^n)\|^2}{\|\nabla Z(x_k^n)\|^2 + \|A_k \nabla Z(x_k^n)\|^2} \quad (9)$$

4. Numerical experiment design

In this section, we will first describe the input data used by method, e.g. historical OD demand generation and DTA traffic assignment procedure. We consider two assessment scenarios in terms of link-flow proportions derivation (i.e. with and without computation of derivatives). Furthermore, we have selected complementary method proposed by (Shafiei et al. 2017) that is based on heuristic approach to solve high computational costs for computing derivatives for the subset of the OD pairs. These scenarios will be discussed in more detail below. Numerical experiments are performed on large-scale network, (Vitoria, Basque Country, Spain) with real data to evaluate the performance of the proposed solution approach.

4.1 DTA with mesoscopic simulation model

In the experiments, we use the mesoscopic event-based demand and supply models in Aimsun, each synthesizing microscopic and macroscopic modelling concept. They couple the detailed behaviour of individual drivers' route choice behaviours with more macroscopic models of traffic dynamics. The travel demand in Aimsun is represented by dynamic OD demand matrices. Vehicle generation is done for each OD pair separately with arrival times that follow an exponential distribution. The iterative interaction between demand and supply

models allows the system to update the set of routes and the travel times after each iteration leading to robust estimation and prediction of traffic conditions in the network.

For this study, a route choice set will be pre-computed in Aimsun and used as fixed for all the simulation runs in the experiments. In this way, dependence of re-routing effects on the changes in the OD demand is ignored. Here we focus to investigate effects of travel time variation and congestion spill-back on traffic observations in the network.

4.2 Network and traffic data

The proposed OD estimation approach is evaluated for large-scale network in Vitoria, consisting of 57 centroids, 3249 OD pairs with a 600km road network, 2800 intersections and 389 detectors, presented as black dots in Figure 2. This network is available in the mesoscopic version of the Aimsun traffic simulation model for the reproduction of traffic propagation over the network. Vitoria network has been selected due to availability of the well calibrated model, its availability to various researchers as a benchmark network, and authors familiarity with the model. The true OD demand is available for this network, which allows analysts to assess the performance of the proposed method. The true assignment matrix and traffic counts on detectors are derived from the assignment of true OD matrix in Aimsun for the evening period from 19:00 to 20:00 reflecting a congested state of the network. The simulation period is divided into 15 minute time intervals with an additional warm-up time interval, R = 5. The link flows resulting from the assignment of the true OD demand are used to obtain the real traffic count data per observation time interval.

The historical OD demand flows are derived by adding a uniform normal component in the range of $\pm$ 40% to the real OD demand to produce uncertainty in the historical demand and congestion in the network.

Figure 2. The Vitoria network, Basque Country, Spain

4.3. Assessment scenario

Here we present a choice of dynamic OD demand estimation methods used within today's dynamic traffic management systems for the performance assessment of the proposed solution approach. Since the main goal of this task is to evaluate the expected improvements due to implementation of nonlinear relationship, in this performance evaluation task we will focus on dynamic OD estimation methods that share same performance measure and solution framework, i.e. least square error measure defined in bi-level solution framework. The

selection of least square objective function indicates that considered methods belong to the common ”family” and ensures to get a better grasp of the algorithms performance and improvements due to application of various solution approaches. Further, the benchmark approach has been selected due to its ability to capture the non-linear relationship and designed heuristic solution approach to deal with high-dimensionality issue of the non-linear problem. As such, it shares the common solution properties of the proposed solution in this paper. For that purpose, in this evaluation task, we consider the following methods:

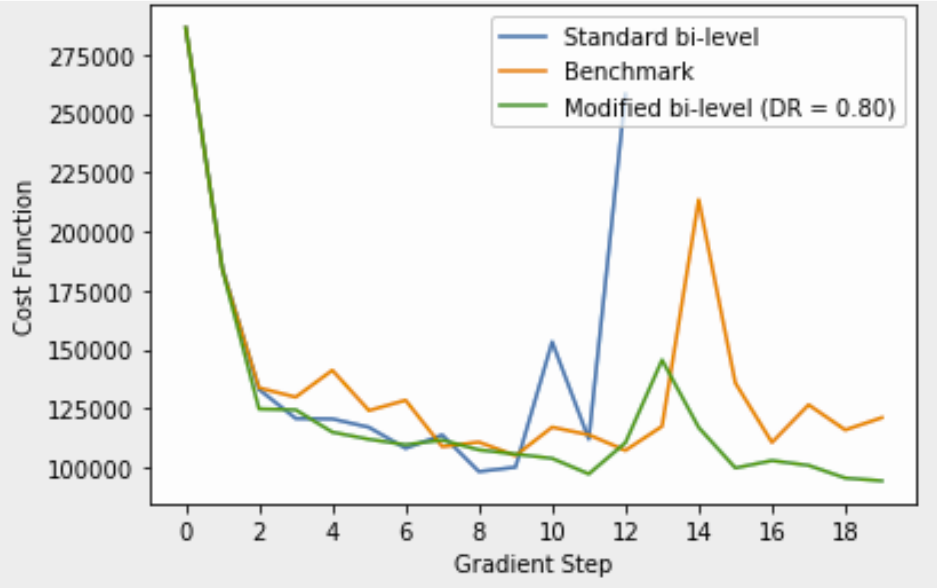
1. Conventional approach: ordinary least square solved by standard bi-level solution approach without explicit non-linear relationship formulation;
2. Benchmark approach: Ordinary least square solved by heuristic bi-level solution approach, where the impact of the OD demand variation on traffic counts have been evaluated for the subset of the OD pairs. For the more detail explanation of this algorithm we refer to the reference paper (Shafiei et al. 2017).

A point of interest now is finding out to which extent the estimation accuracy and computational time are improved. To get a better grasp of the algorithms real world performance, results are presented in following section.

5. Results

The performance of the objective function values for all three methods are presented in Figure 3. For the purpose of this study, convergence was defined as reaching an objective function value that is three times lower than initial value obtained (by any of the algorithms) within 20 iterations. The performance of the proposed modified bi-level solution approach demonstrates satisfying results, since it is able to maintain the decrease of the objective function value through iteration steps.

Figure 3. Comparison of the objective function performance

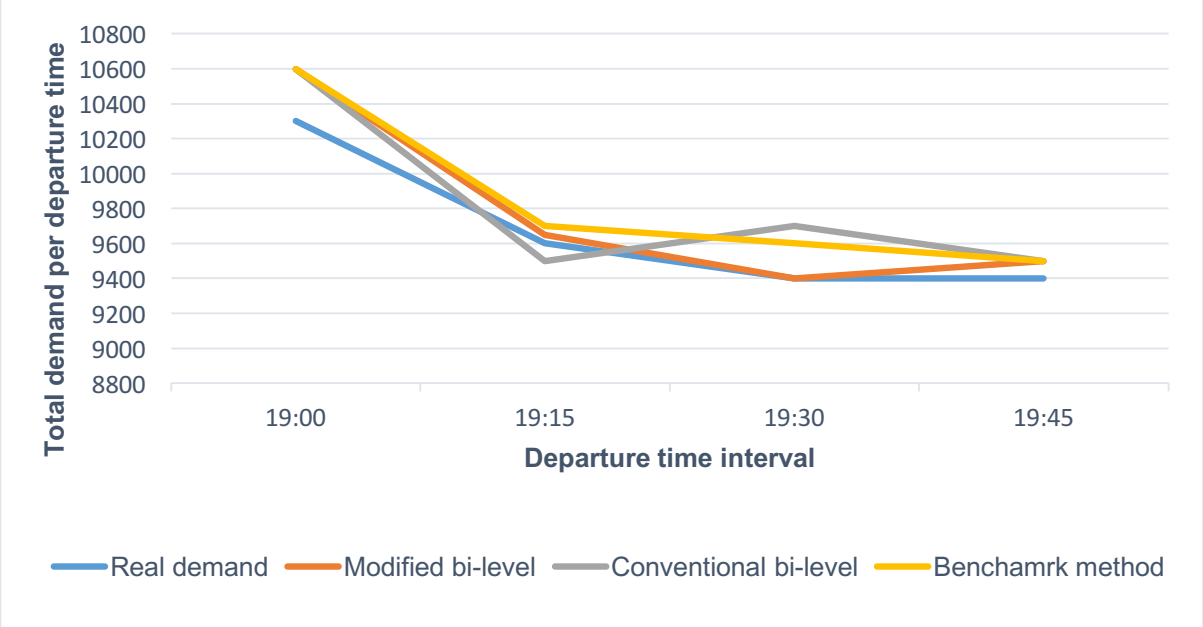


The modified bi-level and benchmark approaches demonstrate convergence trend to a local minimum, very similar to a conventional bi-level optimization framework. Although conventional bi-level approach had a tendency to convergence already in iteration step 8, the convergence trend indicates that the search directions that the algorithm produces are increasingly inefficient as the algorithm progresses. For the purpose of the results visualisation, we have shown results for conventional bi-level approach till iteration step 13, and stored the results for further analysis from the iteration step 8, when method reached the best solution. One may observe that objective values of the proposed approach increased in the iteration steps 12 and 13. Even the recursive step was performed to update the elements

of link flow proportions based on computed derivatives. This effect can be explained by definition of the recursive step in a modified approach, where derivatives have been computed for the subset of the OD pairs, as a result of the trade-off between computational time and performance accuracy. However, this effect can be overcome by extending the list of the OD pairs involved in the computation of the first Taylor approximation given by Equation (3).

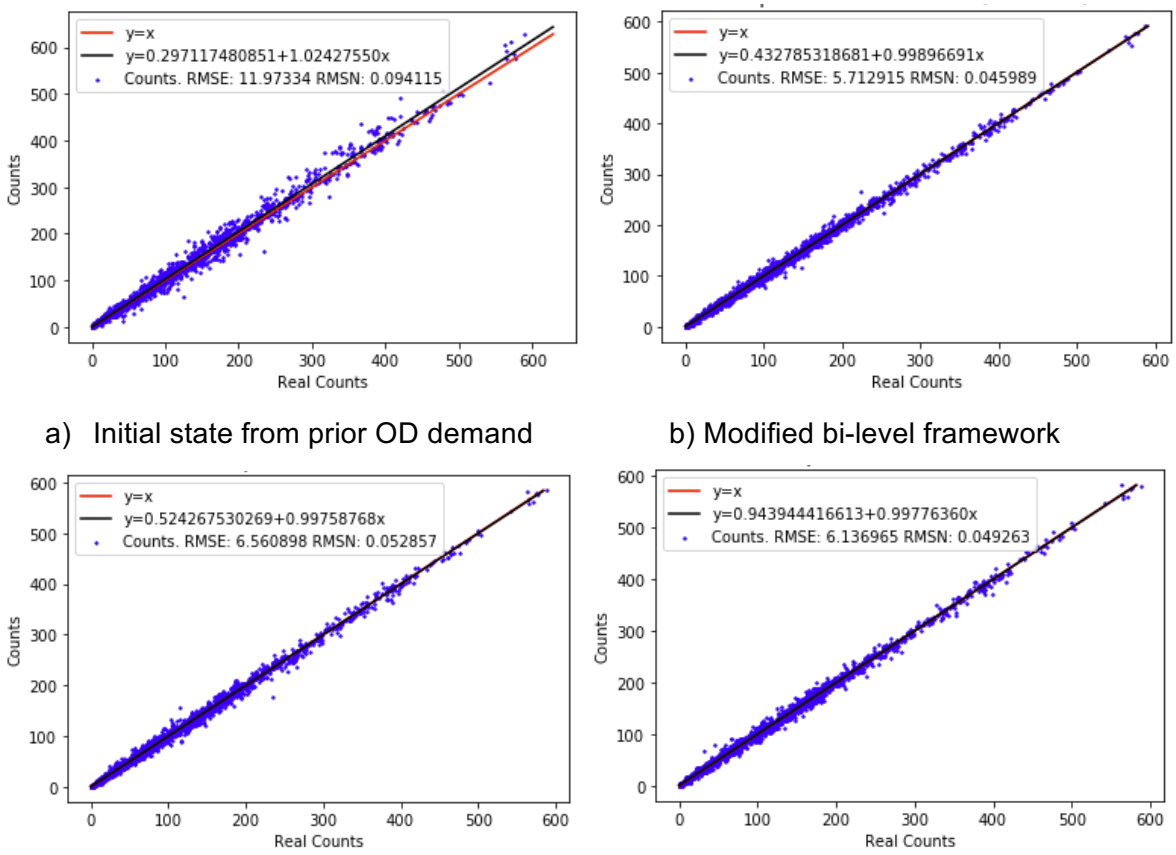
Figure 4 shows the estimated total OD demand per departure time interval for tested OD demand solution approaches. We have included the real OD demand in the figure as a point of reference. All three methods demonstrate tendency to slightly overestimate demand as a consequence of wrong data interpretation from loop detectors when congestion level increase in departure time intervals 2 and 3, i.e. 19:15 and 19:30 respectively. However, performance of the proposed modified bi-level approach demonstrates capacity to recognize overestimation trend and improves demand estimation using the first order approximation given by Equation (3) to update elements of the assignment matrix in congested interval. Further, results indicate unstable performance of the conventional bi-level approach, while benchmark method keeps more stable tendency to overestimate the demand.

Figure 4. Estimated OD demand per departure time interval



Next, it is important to investigate how these estimated OD demands once assigned to network can produce traffic flows close to their real observations. Figure 5 provides a performance overview of solution approaches in terms of relationship between simulated and observed traffic flows. It appears clearly from Figure 5 that all the methods show R-squared value increase and significant reductions in RMSE and RMSN error. Results in Figure 5b) indicate that proposed modified bi-level approach has the best performance with the improvement of 52,29% compared to initial value before OD demand estimation. It is interesting to observe from Figure 5 c) and d) that conventional bi-level and benchmark approach demonstrate slightly lower reduction of the RMSE value, for 48,73% and 45,19% respectively. However, note that results presented here for the conventional bi-level approach are obtained from iteration step 8, since the method was the closest to the convergence in this step as discussed above. In turn, performance results of conventional bi-level approach should be considered with caution.

Figure 5. R^2 for real and simulated traffic flows.



a) Initial state from prior OD demand

b) Modified bi-level framework

c) Benchmark method

d) Conventional bi-level framework

Note that initial idea in our work is to solve the computational complexity of the OD demand estimation problem for real time applications while maintaining reliable estimation results in the congested networks. Therefore, in Table 1 we show the run time and number of Aimsun simulation runs for each of the tested methods (e.g. constant and computed elements of the assignment matrix) on Vitoria network.

Table 1. CPU computation time and the number of the DTA simulations

CPU time	No. of assignment simulations	Aimsun simulation time	Demand estimation time (Python)
Conventional bi-level method	13	26min	31 min
Modified bi-level method	92	3h 4min	1h 16min
Benchmark method	366	12h 12min	14h 32min

Table 1 reports that conventional bi-level method requires the least number of simulation runs. This can be explained by conventional algorithm based on the linear approximation of the assignment matrix requires one Aismun simulation evaluation in each iteration, compared to other two methods that require three simulation evaluations for each OD pair within one iteration step to compute the numerical derivatives of the first Taylor approximation as defined by Equation (3). As a result, the gain in terms of run times obtained by the use of a linear

assignment matrix approximation is large but with trade off on the lower quality of estimation results. However, we can observe significant CPU computation time reduction of the proposed modified bi-level solution approach compared to the benchmark method. This effect can be explained by definition of solution approaches. The proposed modified bi-level approach calculates derivatives for the subset of the OD pairs when deterioration of the objective function is observed in contrasted to the benchmark method that requires in the second iteration step evaluation of the derivatives for all the OD pairs. Furthermore, this computational time gain is expected to increase for larger and more complex networks. These times have been obtained by running Aimsun and Python on DELL Latitude E6430 with processor Intel Core i5-3320M, and 2.6 GHz memory.

6. Conclusions

The common approach usually adopted in dynamic OD demand estimation and prediction consists in solving an optimization problem in which the distance between observed and simulated traffic conditions is minimized by assuming the relationship between OD flows and traffic observations independent of traffic conditions in the network. This approach has a sever shortcoming as it does not take into account the impact of demand flows variation on traffic observations in congested networks. Modelling of non-linear relationship between traffic observations and OD flows, and its dependency on variations in OD flows has been identified by many researchers as a key challenge in the estimation and prediction of an OD matrices.

In this paper, we proposed a modified bi-level optimization framework to solve the high-dimensionality of nonlinear OD estimation problem by computing the derivatives only for the most significant OD pairs with respect to traffic observations that allows the modeller to control the trade-off between simplicity of the model and the level of realism. The proposed algorithm is used to address an estimation with real traffic observations for large scale network: the Vitoria urban network with 3249 OD pairs, 389 detectors, and 2800 nodes. Several specific solution approaches that differ in the assumptions on the link-flow proportions derivation and solution algorithms were used in the performance evaluation study. From the results presented in this contribution we can conclude that proposed approach captured the effect of congestion in the network resulting in the improvement of 52,29% in the estimation of the link flows compared to the initial one. Furthermore, the proposed solution approach significantly decreases the computation time compared to the benchmark method. This result has been achieved by computing the derivatives for the subset of the OD pairs when deterioration of the objective function is observed in contrasted to the benchmark method that requires in the second iteration step evaluation of the derivatives for all the OD pairs. We have also observed that the conventional method can also provide a good solution in a computationally faster manner compared to the other methods. This can be explained by the nature of the Vitoria network selected in the experiments; it is well calibrated network model, especially route choice that has a significant impact on the relationship between demand and supply side. Also, the prior OD matrix has been generated as a random error from the ground true and total demand is very close to the real demand on the network. Further investigation is required, especially if we challenge the model's performance against more bias prior OD demand and even more larger networks over longer demand time intervals of three to four hours, such that methods needs to capture the demand when congestion is building up and resolving.

In this paper we show that deriving non-linear relationship between OD demand and traffic counts for the subset of the OD pairs in proposed solution approach to estimate dynamic OD demand for large-scale networks will lead to computational efficiency with a guaranteed improvement in result's accuracy. An improvement of the algorithm presented in this paper can been seen in two directions: 1) identification and definition of the optimal number of OD pairs to compute the first Taylor approximation such that the computational efficiency, results accuracy and state observability are satisfied; 2) adaptation of the model when additional data (i.e., speeds, density, demand derived from floating car data) can be considered to improve

the quality of the estimated OD demand; 3) explore alternative gradient solution approaches to avoid convergence in local minima.

Acknowledgments

We acknowledge the support of the SETA project funded from the European Union Horizon 2020 research and innovation program under grant agreement No 688082.

References

Ashok, K and Ben-Akiva, ME 2002, Estimation and Prediction of Time-Dependent Origin-Destination Flows with a Stochastic Mapping to Path Flows and Link Flows, *Transportation Science*, 36, (2), pp. 184–198.

Balakrishna, R, Ben-Akiva, M and Koutsopoulos, H 2007, Off-line Calibration of Dynamic Traffic Assignment: Simultaneous Demand and Supply Estimation, *Transportation Research Record: Journal of the Transportation Research Board*, 2003, pp. 50–58.

Ben-Akiva, M, Koutsopoulos, HN and Mukundan, A 1994, Dynamic traffic model system for ATMS-ATIS operations, *I V H S Journal*, 2, (1), pp. 1–19.

Cipriani, E, Florian, M, Mahut, M and Nigro, M 2011, A gradient approximation approach for adjusting temporal origin–destination matrices, *Transportation Research Part C: Emerging Technologies*, 19, (2), pp. 270–282.

Cipriani, E, Gemma, A and Nigro, M 2013, A bi-level gradient approximation method for dynamic traffic demand estimation: sensitivity analysis and adaptive approach, *Proceedings of the IEEE conference on Inteligent Transport Systems, 16th IEEE ITSC*, 1, (2), pp. 2100–2105.

Cantelmo, G, Cipriani, E, Gemma, A and Nigro, M 2014, An Adaptive Bi-Level Gradient Procedure for the Estimation of Dynamic Traffic Demand, *Intelligent Transportation Systems, IEEE Transactions on*, 15, (3), pp. 1348–1361.

Antoniou, C, Azevedo, CL, Lu, L, Pereira, F and 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, pp. 129–146.

Toledo, T and Kolechkina, T 2013, Estimation of Dynamic Origin-Destination Matrices Using Linear Assignment Matrix Approximations, *IEEE Transactions on Intelligent Transportation Systems*, 14, (2), pp. 618–626.

Frederix, R, Viti, F and Tampére, CMJ 2013, Dynamic origin-destination estimation in congested networks: theoretical findings and implications in practice, *Transportmetrica A: Transport Science*, 9, (6), pp. 494–513.

Shafiei, S, Saberi, M, Zockaie, A and Sarvi, M 2017, A Sensitivity-Based Linear Approximation Method to Estimate Time-Dependent Origin-Destination Demand in Congested Networks, *Proceedings of Transportation Research Board - 96th Annual Meeting*, (Washington D.C.), pp. 1–16.

Cascetta, E 1984, Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator, *Transportation Research Part B: Methodological*, 18, (4–5), pp. 289–299.

Yang, H and Zhou, J 1998, Optimal traffic counting locations for origin-destination matrix estimation, *Transportation Research Part B: Methodological*, 32, (2), pp. 109–126.

Tavana, H and Mahmassani, HS 2001, Estimation of Dynamic Origin-Destination Flows from Sensor Data using Bi-level Optimization Method, *Presented at the 80th annual meeting of the Transportation Reasearch Board, Washington DC, USA*.
Lundgren, JT and Peterson, A 2008, A heuristic for the bilevel origin-destination matrix estimation problem, *Transportation Research Part B: Methodological*, 42, (4), pp. 339–354.

Balakrishna, R and Koutsopoulos, H 2008, Incorporating Within-Day Transitions in the Simultaneous Off-line Estimation of Dynamic Origin-Destination Flows without Assignment Matrices, *Transportation Research Record: Journal of the Transportation Research Board*, 2085, pp. 31–38.

Djukic, T, van Lint, H and Hoogendoorn, SP 2012, Application of Principal Component Analysis to Predict Dynamic Origin-Destination Matrices, *Transportation Research Record: Journal of the Transportation Research Board*, 2283, (1), pp. 81–89.

TSS-Transport Simulation Systems 2015, Aimsun Dynamic Simulators Users Manual v8, TSS-Transport Simulation Systems, Barcelona, Spain.

Cauchy, AM 1847, 'Methode generale pour la resolution des systemes d'equations simultanees', *Comptes Rendus Hebd.*, vol. 25, pp. 536–538.