Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 42 (2020) 19–31

Transportation Research Procedia

www.elsevier.com/locate/procedia

46th European Transport Conference 2018, ETC 2018, 10-12 October 2018, Dublin, Ireland

Travel demand matrix estimation methods integrating the full richness of observed traffic flow data from congested networks

Luuk Brederode[a,b*], Kurt Verlinden[c]

[a] DAT.mobility, Deventer, The Netherlands
[b] Department of transport and planning, Delft University of Technology, The Netherlands
[b] Significance, Den Haag, The Netherlands

Abstract

Road travel demand matrix estimation fuses prior or synthetic travel demand matrices with observed flow data. Due to technological advances, ever more observed link flows, speeds and densities are available, whereas rising congestion levels trigger the urgency to use robust and sound estimation procedures on them. This paper addresses difficulties when estimating travel demand using link flows observed on congested networks. Active bottlenecks on these networks influence flow values both upstream (queues will form) and downstream (flow is metered). This implies that, on such a network, observed link flow values may represent either 1) the unconstrained travel demand for that link, 2) a proportion of the capacity of a set of upstream links, 3) the capacity of the normative downstream link; or 4) a combination of these quantities. Which quantity each observed link flow represents depends on the specific traffic conditions in the network. If the assignment model used to assess the relationship between travel demand and link flow does not strictly adhere to link capacity constraints, flow metering effects of bottlenecks (2) are not accounted for and all traffic is considered unaffected (1), thereby forcing incorrect assumptions upon the estimation. Current practice is to derive unconstrained link demand values from flows affected by congestion (2, 3 or 4) and then, instead of the actual observed flows, use these link demand values during matrix estimation. As such, these methods exhibit poor tractability and robustness and do not integrate any information from the assignment model about the composition of routes on the observed links. This paper describes and compares three novel demand matrix estimation methods for large scale strategic congested transport models that use assignment models that strictly adhere to link capacity constraints and explicitly consider the conditions under which link flows are observed. It compares these methods to the current practice and gives practical insights from applications, demonstrating that these methods are more tractable and robust and allow for usage of observed congestion patterns and travel times from (big) data sources. Furthermore, these methods reveal inconsistencies between model link capacities and observed congestion patterns and between count values, allowing the modeler to correct the model

* Corresponding author.
E-mail address: lbrederode@dat.nl

Peer-review under responsibility of the Association for European Transport.
10.1016/j.trpro.2019.12.003

network and other matrix estimation input.

Keywords: demand matrix estimation, congested networks, strict capacity constraints, big data, floating car data, ANPR data, Bluetooth data

1 Introduction

In strategic transport models, road travel demand matrices are usually estimated using estimation methods that fuse prior or synthetic travel demand matrices with flow data observed on individual roads ('links') in the network. On the one hand, ever more data on flows, speeds and/or densities on link level is available, driven by technological advances (e.g. PnD's, smartphones, IoT), trends in transport policy towards smarter usage instead of expansion of the network and the smart mobility concepts arising from them. On the other hand, the urgency of robust and sound estimation procedures is triggered by rising congestion levels on these networks that are at an all-time high. In this paper we address the known difficulties when estimating travel demand using link flows observed on a network with high levels of congestion.

1.1 Interpretation of observed flows under different network conditions

Congested networks incorporate at least several active bottlenecks, which influence flow values both upstream (queues will form) and downstream (flow is metered). This implies that, on such a network, observed link flow values may represent either 1) the unaffected travel demand for that link, 2) a proportion of the capacity of (a set of) upstream link(s), 3) the capacity of the normative (in terms of capacity deficit) downstream link or 4) a combination of these quantities.

These four conditions are illustrated in a unidirectional corridor network with two active bottlenecks in Figure 1, where:

1. The **unaffected links** (continuous black arrows) are unconstrained by active bottlenecks. This means that on links 1 and 2, link outflow equals the demand from centroid 1, whereas link outflow form link 3, equals the demand from centroids 1 and 2.
2. The outflow on **flow metered links** (continuous grey arrows) is determined by active bottlenecks upstream. This means that the outflow on bottleneck link b1 equals the capacity of this link, whereas the outflow on bottleneck link b2 and link 11 equals the capacity of link b2. The outflow on flow metered links 6 and 7 equals the capacity of bottleneck b1 multiplied by the turn proportion from link b1 to link 6, whereas the outflow on link 6a equals the capacity of b1 multiplied by the turn proportion towards link 6a.
3. The outflow on **links in queue** (short dashed black arrows) is determined by the normative downstream link. This means that on links 4 and 5, outflow equals the capacity of b1, whereas outflow of link 10 equals the capacity of b2. Note that for illustrative purposes, in this example bottleneck b1 affects two upstream links, whereas bottleneck b1 only affects one upstream link. In reality, the influence of downstream bottlenecks depends on the severity of the bottleneck in relation to the flow towards it and the buffer capacity on the links upstream from the bottleneck.
4. The outflow on **partially metered links** 8 and 9 (long dashed black arrows) is a combination of unconstrained link demand from centroid 3 and the capacity of active bottleneck (metered link) b1 reduced by the turn proportion towards link 6. Note that combinations of unaffected links (1) and links in queue (3) and combinations of flow metered links (2) and links in queue (3) are not included in this example. In practice, these situations can occur, but only when the considered link has at least one outlink that is not affected by the bottleneck causing the queuing to occur. This requires that exit lanes allowing traffic towards the unconstrained outlinks to freely traverse the queue must exist on the link. These conditions are outside of scope of this paper, as in macroscopic traffic assignment models these conditions cannot occur due to the first-in-first-out (FIFO) assumption in these models, which is required to maintain the route choice of travelers on the network.

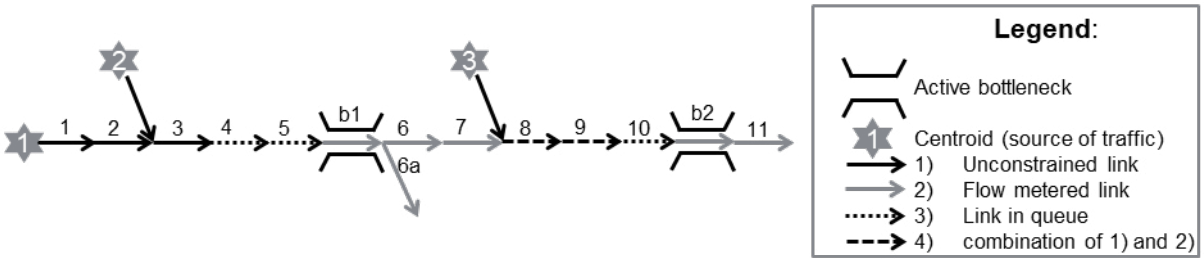


Figure 1: example network illustrating different quantities that link (out)flow may represent

As illustrated in the example from Figure 1, the specific traffic conditions in the network define which quantity each observed link flow represents. Note that only flows measured under conditions (1) or (4) contain information about the absolute level of traffic demand, whereas flows measured under conditions (2) or (3) only contain information about network capacities and bottleneck locations (hence a lower bound on the level of traffic demand).

1.2 Problem formulation and current practice

Demand matrix estimation methods use a traffic assignment model to assess the relationship between travel demand and link flow in intercept information. In current (strategic transport modelling) practice, intercept information is provided by traffic assignment models that cannot distinguish between the different conditions, because they do not strictly adhere to link capacity constraints. Therefore, flow metering (2, 4) nor queuing effects (3) of bottlenecks are taken into account and all traffic is implicitly considered to be unaffected (1), thereby forcing incorrect assumptions upon the estimation. Therefore, matrix estimation methods using these models should only be applied on observed flows values that are unaffected (1), rendering them mostly useless on networks with high congestion levels. Note that by nature these assignment models should not be applied on study areas with congestion altogether.

1.3 Contributions

This paper describes existing demand matrix estimation methods for large scale strategic congested transport models that use assignment models that strictly adhere to link capacity constraints, allowing them to explicitly consider the conditions under which link flows are observed. It compares these methods to the current practice and gives practical insights from applications of methods that are already implemented and applied, thereby demonstrating that these methods allow for usage of (big) data sources such as floating car data and congestion patterns (used in methods 2 and 3) and (route) travel time observations from e.g. Bluetooth or ANPR data (intended to be used in method 3).

2 Methodologies

All methodologies are described in the bi-level optimization framework summarized in equation (1), where in the upper level, the origin-destination (OD) demand matrix is altered to minimize differences between observed and modelled link flows and between the prior and modelled OD demand matrix, while in the lower level a traffic assignment model is used solving a user equilibrium problem translating the new OD demand into modelled link flows.

$$\mathbf{D}^* = \underset{\mathbf{D}}{\text{Argmin}}(F) = \underset{\mathbf{D}}{\text{argmin}}[(\mathbf{D} - \mathbf{D_0})^2 + (\mathbf{y}(\mathbf{D}), \tilde{\mathbf{y}})^2] \quad (1)$$

where F denotes the upper level objective function to be minimized, $\mathbf{D}^*$, $\mathbf{D}$ and $\mathbf{D}^0$ denote vectors containing posterior, current and prior (or observed) OD demand respectively for all OD pairs, $\mathbf{y}(\mathbf{D})$ and $\tilde{\mathbf{y}}$ denote vectors of estimated and observed link flows. Furthermore, we define $L = \{L_1, L_2, L_3, L_4\}$ as the set of observed links split up into the four different traffic condition types, to be used in the remainder of this paper.

2.1 Assignment model classes

In the lower level, the function assign represents the assignment model used. The method from current practice (section 2.2) requires a static capacity restrained traffic assignment (SCRTA) model, whereas the other methods (sections 2.3 through 2.5) require a static capacity constrained assignment (SCCTA) model. The essential difference between these model classes is that the SCCTA strictly respects link capacity constraints (link flow can never be larger than link capacity), whereas in the SCRTA model, only the route choice is influenced by capacity constraints (and link flow can be larger than link capacity). We refer to (Bliemer et al., 2017) for concise definitions of these assignment model classes.

2.2 Solution method used in current practice

Current practice to use observed flows affected by congestion (conditions 2, 3 or 4) is to estimate unconstrained link demand values from the observed flow values, for example using the 'Tonenmethodiek' (Transpute, 2003) used in the Dutch LMS/NRM models, or similar techniques that shift observed flows to upstream unconstrained links. Then, instead of the actual observed flows, the post-processed link demand values are used during OD demand matrix estimation. These models do not make use of any information from the assignment model about the network conditions on the observed links. Therefore, even flow metered observations (which by definition only contain information on network capacity) are erroneously used in the demand estimation instead of network supply calibration. For these reasons, these methods exhibit poor tractability and robustness.

The objective function of this method is defined as:

$$F = w_1 \sum_{od \in OD} (D_{od}^0 - D_{od})^2 + w_2 \sum_{l \in L} \left(y_l^{SCRTA}(\mathbf{D}) - f(\tilde{y}_l) \right)^2, \tag{2}$$

where f denotes a function (like the 'Tonenmethodiek') that estimates corresponding unconstrained link demand values from observed link flows, $y_l^{SCRTA}(\mathbf{D})$ represents the link flows as calculated by a SCRTA assignment model and w_1 and w_2 represent parameters that express the relative importance of the prior demand component in relation to the link flow component in the objective function.

Note that although an SCRTA model must be used to provide the link flows in the upper level, the final assignment of the estimated OD demand matrix can be done using a SCCTA model to increase accuracy. This is effectively being done by the assignment model QBLOK in the LMS/NRM model system, which uses capacity constraints model for route choice and the final assignment results but omits capacity constraints to determine the link flows used in the upper level.

2.3 Solution method 1: Using SCCTA instead of SCRTA model

This method uses the SCCTA model to isolate unmetered from total demand and apply the upper level only on the unmetered demand. To this end, metered demand is subtracted from both the observed and modelled flows, yielding the following objective function:

$$F = w_1 \sum_{od \in OD} (D_{od}^0 - D_{od})^2 + w_2 \sum_{l \in \{L_1, L_4\}} (\alpha_l(\mathbf{D}) y_l^{SCCTA}(\mathbf{D}) - \alpha_l(\mathbf{D}) \tilde{y}_l)^2 \tag{3}$$

where α_l denotes the proportion of flow that arrives at link l unaffected by any upstream bottleneck(s). Proportion factors α_l are derived from od specific proportion factors α_l^{od} outputted by the SCCTA model using:

$$\alpha_l = \frac{\sum_{od \in OD} \delta_l^{od} \alpha_l^{od} D^{od}}{\sum_{od \in OD} \delta_l^{od} D^{od}} \tag{4}$$

where δ_l^{od} is the link-od incidence indicator which equals one if link l is used by od pair od, and zero otherwise.

Note that this method (correctly) only estimates demand using observed flow on links in $\{L_1, L_4\}$, but does not use the information on bottleneck locations that can be derived from observed flows on links in $\{L_2, L_3\}$. This method was initially applied in the 2018 version of the transport models of Noord Brabant, but due to the omission

of information on L_3 links, queue lengths where structurally underestimated and not all bottleneck locations where modelled. This led to adoption of method 2 (described in the next section) in these transport models.

2.4 Solution method 2: Adding information on bottleneck locations

This method is an extension of the method described in 2.3 and adds usage of information from speeds observed on links in $\{L_2, L_3\}$ to determine bottleneck locations. In the applications presented, observed speeds from floating car data where used. To extract bottleneck locations from these speeds, first all links for which the observed speed v_{fcd} is lower than its critical speed v_{crit} are identified as being in queue (i.e.: part of set L_3). The required critical speeds can be derived from the speed limit that applies on the considered link. Once set L_3 has been defined, bottleneck locations can be identified as the node between the last (most downstream) link from a (spatial) sequence of L_3 links and the first (most upstream) link in a (spatial) sequence of other links.

The queue on the (spatial) sequence of L_3 links upstream from the bottleneck location can be translated into an excess demand (D_l^{exc}), which, added to the capacity (C_l) of the first link downstream from the bottleneck location, is treated as a (indirect) observation of demand just upstream from the bottleneck link. Note that, by definition, this first link must belong to L_2.

The method requires excess demand D_l^{exc} to be calculated for all L_3 links in the network. To do so, the observed speeds and the fundamental diagram of each link can be translated into the density that, according to the fundamental diagram, would apply on that link. The densities and lengths of al links in the considered sequence of L_3 links together with the capacity of the bottleneck link can then be translated into the excess demand D_l^{exc}. Alternatively, the set of links in queue (L_3) may be derived from annual summaries of daily traffic reports (e.g. the File Top 50 in the Netherlands (VID, 2017)). These annual summaries provide observed bottleneck locations along with observed queue lengths and durations. Using a bottleneck model these queue lengths can be translated into excess demand, either assuming some value for the density in queue, or using the density values derived from the observed speeds and fundamental diagrams as described above.

Note that both sources for location and excess demand estimation may be combined, allowing the modeler to choose the most accurate source available to be used. For example, annual summaries of traffic reports may provide more accurate bottleneck locations, but they are typically not available for lower order roads, whereas accuracy of the densities derived from observed speeds may provide better estimates for excess demand than the bottleneck model would. This might lead the modeler to choose to use observed bottleneck locations on the higher order roads and derived bottleneck locations on the lower order roads.

Figure 2 illustrates the procedure for bottleneck location detection and excess demand estimation on a corridor, merge and diverge network. The formulae for the corridor and merge cases indicate that the observed link speeds provide enough information to estimate the (indirect) observed demand for the bottleneck link. However, the formula for the diverge network indicates that more information is required to determine the normative outgoing (bottleneck) link. This information cannot be derived from observations on link level and would require observations on node and turn level and modelling of traffic flow on lane and turning movement level. Such observations are not (widely) available yet, but more importantly, such a level of traffic flow modelling is beyond the scope of the macroscopic traffic assignment models used in strategic transport models.

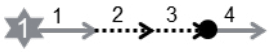
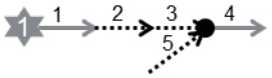
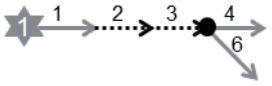
Network		**(Indirect) observed demand**
Corridor		$\tilde{y}_3 = C_4 + D_2^{exc} + D_3^{exc}$
Merge		$\tilde{y}_3 + \tilde{y}_5 = C_4 + D_2^{exc} + D_3^{exc} + D_5^{exc}$
Diverge		$\tilde{y}_3 = C_4 + D_2^{exc} + D_3^{exc}$ or $\tilde{y}_3 = C_6 + D_2^{exc} + D_3^{exc}$

Figure 2: derivation of bottleneck locations and (indirect) observed demand from observed link speeds

Method 2 implies the following objective function:

$$F = w_1 \sum_{od \in OD} (D_{od}^0 - D_{od})^2 + w_2 \sum_{l \in \{L_1, L_4\}} (\alpha_l(\mathbf{D}) y_l^{SCCTA}(\mathbf{D}) - \alpha_l(\mathbf{D}) \tilde{y}_l)^2 + w_3 \sum_{b \in B} \left(C_{\underline{b}} + D_{\underline{b}}^{exc} - \sum_{\overline{b}} y_{\overline{b}}^{SCCTA} \right)^2 \quad (5)$$

where B denotes the set of bottleneck nodes, $C_{\underline{b}}$ and $D_{\underline{b}}^{exc}$ denotes the capacity and excess demand on the normative outlink (the link that has caused activation of the bottleneck location) of bottleneck node b respectively, $y_{\overline{b}}^{SCCTA}$ denotes the modelled flow on inlink $\overline{b}$ of bottleneck node b and w_3 denotes the relative importance of the objective function component that covers the demand for bottleneck links.

This method was applied on the NRM-West: the Dutch regional strategic transport model of the Randstad Agglomeration (including the 4 largest cities in the Netherlands) as described in (Brederode et al., 2017) and is recently implemented in the 2018 version of the strategic transport models of the province of Noord Brabant.

2.5 Solution method 3: Adding sensitivities of proportion factors and travel times

Equations (3) and (5) indicate that the unmetered proportion factors α_l depend on the current OD demand matrix $\mathbf{D}$. This means that any changes made to $\mathbf{D}$ in the upper level have an immediate effect on the value of the unmetered proportion factors, whereas these are considered constant in objective functions (3) and (5). For this reason, this method approximates the sensitivity of the proportion factors to changes in demand ($\partial \alpha_l / \partial \mathbf{D}$) using marginal simulation of the node model component within the assignment model and adds these sensitivities to the objective function assuming a first order Taylor approximation.

Equation (5) indicates that the bottleneck component in the objective function is competing with the prior demand and link flow components. This means that adding the bottleneck component reduces the chance that bottlenecks switch from active to inactive state during matrix estimation. Bottlenecks that switch from active to inactive or vice versa disturb the matrix estimation process is undesirable, because 1) it causes changes to the definition of sets L_1, L_2, L_3 and L_4, thereby non-convergence; and 2) because the (added) sensitivities of the proportion factors are point approximations which are only valid when the considered link remains in the state in which the sensitivity was estimated. For these reasons, this method removes the bottleneck component from the objective function and instead, adds it as a constraint to the optimization problem.

Lastly, this method adds, when available, travel times to the objective function, as these can also be expressed as a function of α_l. Observed travel times can be derived from e.g. floating car data on link level or from ANPR or Bluetooth measurements on route level. These changes and additions yield the following optimization problem:

$$\mathbf{D}^* = \underset{\mathbf{D}}{\operatorname{argmin}} \left[w_1 \sum_{od \in OD} (D_{od}^0 - D_{od})^2 + w_2 \sum_{l \in \{L_1, L_4\}} \left(\left[\alpha_l(\mathbf{D}) + \frac{\partial \alpha_l}{\partial \mathbf{D}} (\mathbf{D} - \mathbf{D^0}) \right] y_l^{SCCTA}(\mathbf{D}) - \alpha_l(\mathbf{D}) \tilde{y}_l \right)^2 + w_4 \sum_{p \in \tilde{P}} \left(\tau_p(\mathbf{D}) - \tilde{\tau}_p \right)^2 \right] \quad (6)$$

$$\text{s.t.:} \quad \mathbf{y}(\mathbf{D}) = assign(\mathbf{D}),$$
$$\mathbf{D} > 0,$$
$$\sum_{\overline{b}} y_{\overline{b}}^{SCCTA} = C_{\underline{b}} + D_{\underline{b}}^{exc} \quad \forall\, b \in B,$$

where $\tilde{P}$, $\tilde{\tau}_p$ and τ_p denote the set of paths with observed travel times, the observed travel time on path p and the modelled travel time on path p respectively. Weighing parameter w_4 expresses the relative importance of the travel time component of the objective function.

This method is a continuation of the method described in (Brederode et al., 2014) and is implemented in

prototype form. The method has proven to outperform methods 1 and 2, both in accuracy as well as speed of convergence on small test networks. The prototype is still under development as its runtimes make it currently not practically applicable on large networks.

3 Software

Section 2 describes four methodologies in terms of problem formulations and solution methods, which give insight in the theoretical added value of the different methods. However, the extent to which this theoretical value is translated into practical value is determined by its software implementation. Therefore, this chapter briefly describes the different software implementations used by the authors gaining insights in practical applications, before these insights are described in chapter 4. Table 1 summarizes the applications and software examined.

Table 1: applications examined in this paper

Method	Transport model used	Software lower level	Software upper level
Current practice	LMS/NRM models	QBLOK	AVVMAT
Method 1	models of Noord Brabant	STAQ	OtMatrixEstimation
Method 2	models of Noord Brabant	STAQ	OtMatrixEstimation
Method 2	NRM West (Randstad model)	STAQ	AVVMAT
Method 3	Various transport models	STAQ	MATLAB

Furthermore, in this chapter, requirements for alternative software implementations (not used by the authors) are given to allow readers to adopt methods from section 2 in their own preferred software.

3.1 Assignment model used in current practice

Authors have gained experience of current practice using the assignment model used in the LMS/NRM methodology of the dutch national and regional strategic transport models: QBLOK (Bakker et al., 1994). QBLOK is a deterministic equilibrium model that extends traditional SCRTA models on the following three points:

1) It not only models actual flow that uses the network within the study period, but also the flow that would have wanted to travel in the study period but did not reach its destination in time due to congestion.
2) It takes the network effects of congestion (flow metering and spillback) into account using a heuristic, but these effects are only included in link travel time calculation (and thus route choice), not in the modelled traffic flows.
3) It uses a fixed number of iterations and prefixed weights to approximate the user equilibrium, as convergence to equilibrium is infeasible within acceptable computation times.

3.2 Assignment model used in methods 1 through 3

Authors have gained experience of methods 1 through 3 using the SCCTA model STAQ described extensively in (Brederode et al., 2018). STAQ is implemented as a propagation model within the StreamLine framework in OmniTRANS transport planning software.

The model supports any concave, two regime fundamental diagram, but insights in this paper where gained using the quadratic linear diagram (QL) from Bliemer et al., 2014. To describe interaction of flows on nodes STAQ uses the explicit node model from Tampère et al., 2011, which allows to explicitly calculate and output OD specific proportion factors α_l^{od} used in methods 1 through 3. This node model is also used in method 3 for the marginal simulation that determines the sensitivity of the proportion factors to changes in demand ($\partial\alpha_l/\partial\mathbf{D}$). In most studies, the node model was extended with the junction modelling component of OmniTRANS transport planning software to account for the effect of limited supply due to conflict points on the junction itself (i.e. crossing flows), and to calculate travel-time delays due to geometry of the node and conflicts on turning-movement level.

The assignment model can be used with different route choice models, but insights in this paper where gained using the multinomial logit (MNL) model with scale parameters set to one over 14% of the minimal route cost of the

considered OD pair. This means that the route choice model is only sensitive to the ratio of different route costs, not their absolute values. In all three methods, the stochastic user equilibrium (SUE) is used as underlying route choice paradigm. The adapted relative duality gap derived in (Bliemer et al., 2013) that accounts for perception errors and thus reaches zero upon convergence when using the MNL route choice model is used as convergence criterion with a threshold value of 5E-04. The method of self-regulating averages described in (Liu et al., 2009) is used to average route demands over iterations providing fast convergence.

The assignment model makes use of pre-generated route sets. Insights in this paper where gained using the routeset generator from the StreamLine framework, which uses the Dijkstra algorithm to find the shortest path between each OD pair and then uses a repeated random sampling process on free flow link travel times using a gamma distribution known as the accelerated Monte Carlo method (Fiorenzo-Catalano 2007) to generate additional alternative routes. Route filters are applied after the repeated random sampling process to reduce route overlap, remove irrelevant routes and restrict the size of the set of potential routes.

For methods 1 and 2, any SCCTA model that can output OD-specific proportion factors α_l^{od} can be used as an alternative for STAQ. For method 3, the assignment model must be suitable to be used to efficiently approximate the sensitivity of the link-based proportion factors to changes in demand ($\partial\alpha_l/\partial\mathbf{D}$)

3.3 Upper level solvers used in current practice and methods 1 and 2

In current practice and in the NRM West application of method 2, the AVVMAT software is used to solve the upper level. AVVMAT is based on the Combined Calibration matrix calibration program develop by Hague Consulting Group in the 1990's. AVVMAT assumes a multiplicative model in which each matrix cell is a function of its initial value and a set of parameters (count, trip ends, trip length class, etc.). Furthermore AVVMAT assumes that the parameters are statistical of nature and therefor have a level of reliability. AVVMAT assumes a Poison distribution and applies the BFGS algorithm. Derivation and implementation of the AVVMAT OD matrix estimator is described in (Lindveld, 2006) in more detail.

In the application of methods 1 and 2 on the models of Noord-Brabant OtMatrixEstimation is used to solve the upper level. OtMatrixEstimation is part of OmniTRANS transport planning software and uses a heuristic to iteratively scale relevant matrix cells in the prior matrix to better match with observed flows on link or screen line level. It threats trip ends and trip length distribution from the prior as constraints. A more extensive description of the heuristic can be found in (Smits, 2010).

For methods 1 and 2, any solver that can handle the convex quadratic objective function and the non-negativity constraint may be used. However, because OD demand matrix estimation problems in strategic transport models are usually very large and very sparse, the solver should be able to exploit the sparsity of the problem to be able to solve the problem within constraints of available computer memory and required computation time. For the same reason, the solver should not rely on finite differences to approximate the gradient. Instead, the solver should use an analytically calculated gradient.

3.4 Upper level solver used in method 3

The upper level and interaction with the lower level of method 3 is implemented in prototype form in Matlab and uses the fmincon interior point algorithm within the optimization toolbox of Matlab. The interior point algorithm uses a conjugate gradient descent method. To prevent memory issues, it is set to use the limited memory version of the BFGS algorithm (Nocedal, 1980) to approximate the Hessian. Furthermore, functions to calculate the gradient of the objective function and the gradient of the bottleneck constraints are included in the implementation and passed to fmincon, to prevent it from doing finite difference analysis on every OD pair (which would take too much time).

Like method 2, method 3 can be solved using any solver that is suitable for large sparse quadratic optimization problems. However, it must also be able to include the linear (bottleneck) constraints. Furthermore, analytical calculation of the gradient for both the objective function and constraints is possible (Rijksen, 2018), but due to the inclusion of the sensitivity of the bottleneck proportion factors the calculations are more complex than for methods 1 and 2.

4 Practical Insights from applications

This section discusses insights from the practical applications listed in Table 1.

4.1 Insights in conditions on observed links

Preliminary analysis on the input data of the Noord-Brabant and NRM West models reveal using floating car data to identify the links in queue (L_3) and STAQ to determine the distribution over unaffected (L_1), metered (L_2) and partially metered (L_4) links. Results of this analysis reveal that the majority of observed link flows are unaffected or partially metered (i.e. they belong to $\{L_1, L_4\}$) and could thus be used for demand estimation using method 1.

To illustrate this, we describe results from the preliminary analysis for the AM peak period of the base year of the NRM-West, which describes the most congested region of the Netherlands. For this model, there were no flow metered observed link flows ($L_2 = \emptyset$) and only 6% of the count locations where observed in queue. Covering the other 94% of the count locations, the black line in Figure 3 shows the portion of flow unaffected by upstream bottlenecks per count location according to the assignment results of the prior OD demand matrices. In the graph, count locations are ordered by their portion of unaffected flow increasingly. This reveals that about 66% of the count locations where partly metered (the percentile where the black line hits 100%) and about 34% was unaffected (the remainder of the locations). These findings suggest that although most observed link flows are influenced by congestion, there are only few observed links that are not suitable to be able to apply method 1. Since the NRM-West describes the most congested region in the Netherlands, other Dutch models are expected to exhibit even lower proportions of link flows unsuitable for use with method 1.

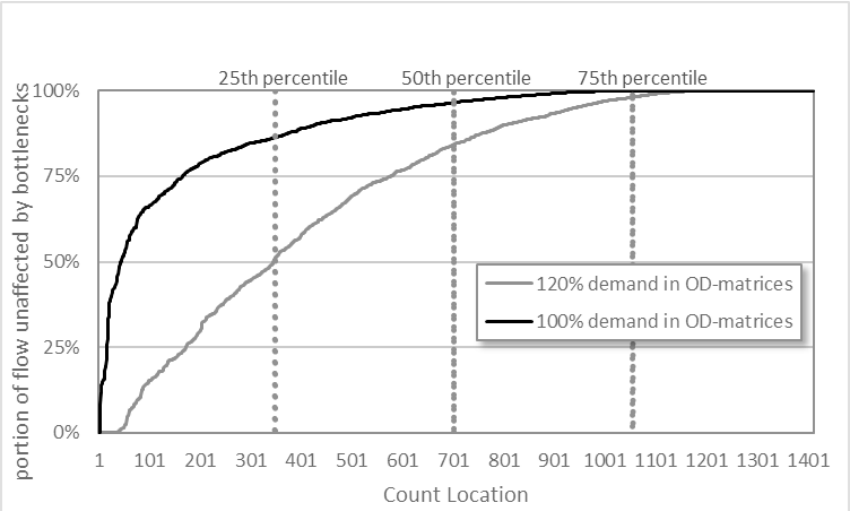


Figure 3: portion of flow unaffected by bottlenecks per count location in NRM-West model, AM peak

To determine robustness of these findings, sensitivity analysis was carried out in which the prior OD demand matrix was increased by 20%. The result is displayed as the gray line in Figure 3. In this case, around 3% of the count locations that where not in queue became flow metered, whereas the share of partially metered count locations increased to about 76%, leaving 21% unaffected. Although the 20% increase of demand yields slightly more links to become unsuitable for application of method 1, it is still only a small minority. Since methods 2 and 3 follow the same underlying principle but add (indirect) estimation using observations on links in queue (L_3) these insights about applicability holds to an even greater extent for those methods. For these methods no more than 3% of observed link flows is metered and could therefore not be used in demand matrix estimation methods 2 and 3.

4.2 Insights from method used in current practice

The method used in current practice (as described in 2.2) circumvents computational issues that arise from inclusion of strict capacity constraints in OD demand matrix estimation methods for strategic transport models by projecting the (estimated) effect of capacity constraints on the input of the methodology (the observed flows), rather than adapting the methodology itself to include the constraints. This approach allows for the usage of proven technology: SCRTA models and (upper level) solution methods widely available since the 1990's; see e.g. (Abrahamsson, 1998) for an overview. Because of the (desirable) mathematical properties of the SCRTA model and its corresponding (upper level) problem, solutions are found relatively easily and fast.

However, using (estimated) link demands instead of observed link flows as primary input gives rise to the following myriad of problems all related to the fact that link demand is a quantity that cannot be measured. Firstly, this means that the accuracy of methods that estimate link demands (e.g. Tonenmethodiek) cannot be determined directly. Instead, only the accuracy of the solution method as a whole can be evaluated by comparing the result of a capacity constrained assignment of the estimated OD demand matrix with the observed flows. Differences between these modelled and observed link flows can be caused by either errors in the method used to estimate link demands, the assignment model or the solution method. Formulated differently: although the methodology minimizes differences between observed and modelled link demand, it does not necessarily minimize differences between observed and modelled link flows. This means that calibration of the parameters of the matrix estimation method and the assignment model, as well as finding and fixing input errors needs to take place in a single process. In practice, this leads to extensive estimation procedures that aim to provide acceptable outcomes using (structured) trial and error. This causes high and uncertain lead times for projects including OD demand matrix estimation with only reasonable outcomes.

Secondly, the process described in the previous paragraph is highly sensitive to changes in input. This means that a process that has produced acceptable outcomes for a set of observed link flows representing a certain study area or base year might not give acceptable outcomes for set of observed link flows representing another study area or base year. This reasoning also holds for different sets of parameters for the assignment model and/or upper level solution method. In practice, this requires estimation procedures to be changed when the input data or parameter set of the considered project gives rise to it. This causes expensive matrix estimation projects with poor tractability and comparability of model outputs.

4.3 Insights from application using method 1

By replacing the SCRTA model with a SCCTA model and considering only the unmetered demand in the upper level, the problems related to the usage of link demands described in section 4.2 are effectively removed. Method 1 allows to directly compare modelled flows with observed link flows and to isolate effects of changes in parameters of the upper level solution method and SCCTA model. Furthermore, there is no need to change the estimation procedure when input or parameter sets change.

Although the share of observed link flows in queue is only small (section 4.1), these links are most important for a transport model to describe accurately. However, as mentioned in section 2.3, method 1 does not use information on links in queue, hence it neglects observed queues. Instead, the hypothesis behind method 1 is that demand estimation on the other (majority) of the count locations will cause the correct demand on the queued links as well. This hypothesis proved wrong, as it turns out that fitting flows on unconstrained or partially constrained links only does not (substantially) improve the fit of link demand for links in queue.

The reason for this is explained using the example in Figure 4. Assume that in this network observed flows are available for links 2 and 3. Method 1 would not use any information from link 3 (as this link is in queue) and thus would only try to minimize differences between modelled and assigned flow on link 2. Assume that modelled flow on link 2 is underestimated. Method 1 would then evenly increase demand on all OD pairs using link 2, neglecting the effect that demand on OD pairs towards link 6 would have on the queue on link 3, whereas a different (more uneven) distribution over OD pairs could effectively improve the fit on link 3 with the same improvement of fit on link 2.

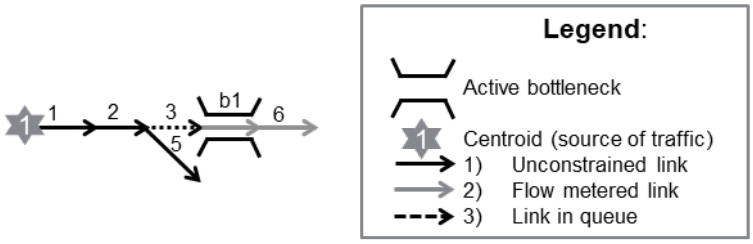


Figure 4: example where estimating demand matrices using method 1 and a single count on link 2 does not imply a positive effect on the fit on link 3

The example shows that the OD demand matrix estimation problem has too many degrees of freedom (too many OD pairs to choose from) to expect a method to improve the overall fit on links that the method does not explicitly consider. For method 1, this means that the level of congestion in the final assignment results will mainly be determined by the level of congestion in the assignment result of the prior demand matrix[1], as was seen in its application in the transport models of Noord-Brabant.

Another potential issue of method 1 is that it contains no mechanism that prevents bottlenecks to switch from active to inactive or vice versa during the matrix estimation process. As described in section 2.4, this causes poor convergence. In the Noord-Brabant application, this issue did not clearly manifest itself, probably because the prior demand matrices generally underestimated the congestion levels causing a limited set of active bottlenecks and hence a small chance of state switches during estimation. However, the more theoretical tests described in (Brederode et al., 2014; Frederix, 2012) clearly demonstrate this issue.

4.4 Insights from applications using method 2

In addition to the findings described in (Brederode et al., 2017) the application of method 2 on the strategic transport model of the Randstad Agglomeration proved that the addition of (indirect) observation of demand just upstream from the bottleneck link allows for accurate representation of observed queues while maintaining the fit on unconstrained and (partly) flow metered links, thereby solving the problems described in section 4.3.

The application also demonstrated that by changing the ratio between weights w_2 and w_3, inconsistencies between model link capacities and observed congestion patterns and inconsistencies between count values can be isolated, allowing the modeler to correct the model network and other matrix estimation input. Often errors with respect to the exact bottleneck location, its normative outlink or the combination of observed flows and observed link speeds from different data sources proved to be the cause of these inconsistencies. The methodology proved an asset in removing these errors and inconsistencies.

However, the application also showed that weighing parameter w_3 needs to be set carefully. It should be high enough to ensure and maintain activation of the correct bottlenecks throughout the estimation procedure, but low enough to allow accurate representation of unaffected and partly metered link flows near bottleneck locations.

4.5 Insights from applications using method 3

As mentioned in section 2.5, method 3 outperforms methods 1 and 2, both in terms of accuracy as well as convergence properties. On top of that, removes the issue of choosing w_3 by replacing this component of the objective function with an equivalent constraint. Furthermore, it supports observed travel times as an additional input data type.

However, the prototype is still under development as its runtimes make it currently not applicable on large networks. This is mainly caused by large sparse matrix multiplications that are required to translate the sensitivities of the proportion factors from the marginal node model runs to their effect on the objective function. Until this implementation issue is fixed, the method is best applied excluding the sensitivities of the proportion factors but including the constraints that ensure and maintain correct bottleneck states.

[1] Assuming that no wide-spread unidirectional changes to the demand matrix are being made by the estimation method

5 Conclusion and recommendations

Active bottlenecks in congested networks influence observed link flow values both up- as well as downstream. In strategic transport models, this means that an observed link flow value is either unaffected, metered, in queue or partially metered due to active bottlenecks. Flow observed on unaffected or partially metered link contains information about travel demand that can be directly used for OD demand matrix estimation, whereas observed flow on links in queue is only useful when supplemented by observed link speeds or queue lengths. Flows observed on metered links only contain information on network supply and can therefore not be used for travel demand estimation.

Data and sensitivity analysis on the transport model describing the most congested region of the Netherlands indicates that it is highly unlikely that more than 3% of observed link flows of any Dutch strategic transport model is flow metered, meaning that more than 97% contains information on OD demand. This data can be used, provided that observed link speeds (from e.g. floating car data) or observed queue locations (from e.g. daily traffic reports) are available and that a method that supports partial metered and links in queue is used.

The most common OD demand matrix estimation method used for strategic transport models can only handle observed flows that are unaffected by active bottlenecks (which is the case for 34% or less of the observations), and therefore needs to translate observed link flows into estimated link demands to account for bottleneck effects. This approach allows for the usage of conventional SCRTA models and relative quick solution of the matrix estimation problem. However, the use of input that is estimated rather than measured, makes the method non-transparent and input sensitive resulting in poor tractability, comparability and transferability of estimation processes. This has led to high and uncertain project lead times with outcomes of only reasonable accuracy.

Therefore, this paper assessed three methods that take bottleneck effects into account by replacing the SCRTA model with a SCCTA model. Method 1 can handle observed flows on partially metered links in addition to unaffected links and allows for direct use of and comparison with observed link flows. Methods 2 and 3 additionally provide support for observed queue lengths on links in queue, thereby integrating the full richness of traffic flow data on congested networks.

For network diverges, methods 2 and 3 require information on the normative outgoing link which demands for observations on node and turn level and modelling of traffic flow on lane and turning movement level. Such observations are not (widely) available yet, but more importantly, such a level of traffic flow modelling is beyond the scope of the macroscopic traffic assignment models used in strategic transport models as it would violate the first-in-first-out assumption. Although (Wright et al., 2017) describe a node model that would allow for such violations, development in this direction will (further) degrade on mathematical properties that are desirable in the strategic context: existence and uniqueness of the SUE solution of the assignment model.

Compared to method 2, method 3 provides greater accuracy and faster convergence, removes the need to set a sensitive w_3 parameter and it supports observed travel times as an additional input data type. However, its implementation is still in prototype form limiting its applicability on large networks.

The upper level problem of all three methods can be solved using widely available software packages for large sparse quadratic programming problems with linear constraints. Calculation of the gradient efficiently and correctly is a point of attention when implementing these methods.

Currently, the authors are working on extension of methods 2 and 3 to support estimation of OD demand matrices that cover multiple period(s), which should eventually lead to a method that supports 24 hour estimation. This requires the SCCTA model to be extended to be able to transfer residual traffic (traffic that has not reached its destination within a previous time period) to the next time period, and an upper level extension that can simultaneously estimate matrices for all considered time periods. Both extensions are viable from a methodological point of view, but especially implementation of the latter is expected to create new challenges.

References

Bakker, D., Mijjer, P.H., Hofman, F., 1994. QBLOK: an assignment technique for modelling the dependency between bottlenecks and the prediction of grid lock, in: Proceedings of Colloquium Vervoersplanologisch Speurwerk. Presented at the Colloquium vervoersplanologisch speurwerk 1994. Implementatie van beleid.de moeizame weg van voornemen naar actie., Delft, pp. 313–332.

Bliemer, M.C.J., Raadsen, M.P.H., Brederode, L.J.N., Bell, M.G.H., Wismans, L.J.J., Smith, M.J., 2017. Genetics of traffic assignment models for strategic transport planning. Transport Reviews 37, 56–78. https://doi.org/10.1080/01441647.2016.1207211

Bliemer, M.C.J., Raadsen, M.P.H., De Romph, E., Smits, E.-S., 2013. Requirements for traffic assignment models for strategic transport planning: a critical assessment, in: Paper Presented at: Proceedings of the 36th Australasian Transport Research Forum 2013, ATRF, Brisbane, Australia, 2-4 October, 2013. Australasian Transport Research Forum.

Bliemer, M.C.J., Raadsen, M.P.H., Smits, E.-S., Zhou, B., Bell, M.G.H., 2014. Quasi-dynamic traffic assignment with residual point queues incorporating a proper node model.

Brederode, L., Pel, A., Wismans, L., de Romph, E., Hoogendoorn, S., 2018. Static Traffic Assignment with Queuing: model properties and applications. Transportmetrica A: Transport Science 1–36. https://doi.org/10.1080/23249935.2018.1453561

Brederode, L.J.N., Hofman, F., van Grol, R., 2017. Testing of a demand matrix estimation method Incorporating observed speeds and congestion patterns on the Dutch strategic model system using an assignment model with hard capacity constraints. Presented at the European Transport Conference, AET 2017 and contributors.

Brederode, L.J.N., Pel, A.J., Hoogendoorn, S.P., 2014. Matrix estimation for static traffic assignment models with queuing. hEART 2014 - 3rd symposium of the European association for research of transportation, Leeds UK.

Lindveld, K., 2006. O-D matrix estimation using the Combined Calibration method as applied to the LMS/NRM.

Liu, H.X., He, X., He, B., 2009. Method of Successive Weighted Averages (MSWA) and Self-Regulated Averaging Schemes for Solving Stochastic User Equilibrium Problem. Networks and Spatial Economics 9, 485–503. https://doi.org/10.1007/s11067-007-9023-x

Nocedal, J., 1980. Updating quasi-Newton matrices with limited storage. Mathematics of computation 35, 773–782.

Rijksen, B., 2018. Matrix Estimation With STAQ (Masters Thesis). University of Twente, Deventer.

Smits, E.-S., 2010. Origin-Destination Matrix Estimation in OmniTRANS (Masters Thesis). Utrecht University, Deventer.

Tampère, C.M.J., Corthout, R., Cattrysse, R., Immers, L.H., 2011. A generic class of first order node models for dynamic macroscopic simulation of traffic flows. Transportation Research Part B: Methodological 45, 289–309. https://doi.org/10.1016/j.trb.2010.06.004

Transpute, 2003. “Flowsimulator; beschrijving van het model”, uitgebracht aan de Adviesdienst Verkeer & Vervoer, Gouda.

VID, 2017. File top 50 [WWW Document]. VID | file top 50 over 2017. URL https://rijkswaterstaatverkeersinformatie.nl/top50.2017.html (accessed 9.7.18).