Transportmetrica A: Transport Science

ISSN: 2324-9935 (Print) 2324-9943 (Online) Journal homepage: http://www.tandfonline.com/loi/ttra21

STAQ – Static Traffic Assignment with Queuing: Model properties and applications

Luuk Brederode, Adam Pel, Luc Wismans, Erik de Romph & Serge Hoogendoorn

To cite this article: Luuk Brederode, Adam Pel, Luc Wismans, Erik de Romph & Serge Hoogendoorn (2018): STAQ – Static Traffic Assignment with Queuing: Model properties and applications, Transportmetrica A: Transport Science, DOI: 10.1080/23249935.2018.1453561

To link to this article: https://doi.org/10.1080/23249935.2018.1453561

Publisher: Taylor & Francis & Hong Kong Society for Transportation Studies Limited

Journal: *Transportmetrica A: Transport Science*

DOI: 10.1080/23249935.2018.1453561

STAQ – Static Traffic Assignment with Queuing: Model properties and applications

Luuk Brederode[*, 1,2], Adam Pel[1], Luc Wismans[2,3], Erik de Romph[4], Serge Hoogendoorn[1]

[1]*Department of Transport & Planning, Delft University of Technology*

[2]*DAT.mobility – a Goudappel Company*

[3]*Centre for Transport Studies, Faculty of Engineering Technology, University of Twente*

[4]*Sustainable Urban Mobility & Safety, TNO, The Hague.*

[*]Corresponding author: lbrederode@dat.nl

Acknowledgements

We would like to thank the Dutch municipality of Den Haag and province of Noord-Brabant, and the Flemish traffic centre, as well as Goudappel Coffeng (the Netherlands) for providing the strategic transport models used in this paper.

STAQ – Static Traffic Assignment with Queuing: Model properties and applications

This paper describes the road traffic assignment model STAQ that was developed for situations where both static (STA) and dynamic (DTA) traffic assignment models are insufficient: strategic applications on large-scale congested networks. The paper demonstrates how the model overcomes shortcomings in STA and DTA modeling approaches in the strategic context by describing its concept, methodology and solution algorithm as well as by presenting model applications

on (small) theoretical and (large) real-life networks. The STAQ model captures flow metering and spillback effects of bottlenecks like in DTA models, while its input and computational requirements are only slightly higher than those of STA models. It does so in a very tractable fashion, and acquires high-precision user equilibria (relative gap < 1E-04) on large scale networks. In light of its accuracy, robustness and accountability, the STAQ model is discussed as viable alternative to STA and DTA modeling approaches.

Keywords: traffic assignment; strategic planning; large-scale; congested networks; model; quasi-dynamic

1 Introduction

Since the late 1950's strategic transport models are used to assess the long-term impact of transport policies and the design and management of transport systems. Since then, road traffic congestion has become a common and structural part of many transport systems around the world. However, strategic transport models differ strongly with respect to how such structural congestion and the effects thereof are accounted for within the model used in the traffic assignment (TA) step. The TA model uses the travel demand and network supply as input and usually solves a user equilibrium (UE) problem determining the routes that travelers choose as well as the resulting traffic state on the network (i.e. traffic conditions, including congestion). The TA model is often the most computational expensive component of the model system because an iterative approach is required to solve the UE problem. Given the above, we argue that there is a need for computationally efficient TA models in strategic transport models for large-scale[1] transport systems with structural congestion.

[1] As a rough reference, we consider a network to be large-scale when it contains more than 1 million used OD pairs. This is typically the case with networks containing 3000 or more centroids. Note that the size in this context need not be directly related to the size of the

From the combined perspectives of policy makers and TA model users, the authors argue that apart from computational efficiency and the ability to accurately capture the effects of structural congestion, TA models should also be based on input data that can be forecasted with sufficient certainty for (distant) future years, and should produce accurate, robust and accountable model results for all vehicle classes and for both urban roads and motorways upon assessing policy, design and management measures for transport systems. These desired properties for TA models are in line with (Bliemer et al. 2013; Flügel et al. 2014; Flötteröd 2015)[2] and are defined below.

The ability to capture congestion effects pertains to how bottlenecks lead to flow metering and spillback as well as how it affects route choice. Robustness and accountability are desired properties, because when comparing model outputs of different scenarios (e.g. sets of policy measures), we aim to single-out differences only caused by or related to the different scenario inputs. Hence, differences caused by random variables (e.g. due to stochastic processes in the model) or because the model output does not (sufficiently) represent a stable system state[3] should be negligible or non-existent. Accountability also means that it should be possible to pinpoint and size

---

study area; a fine grained network of a relatively small area with lots of socio-economic activity can also be qualified as large.

[2] The referred papers also mention extendibility (the extent to which the model is prepared for future modes) and consistency with microscopic models as desired properties. However, we argue that these are no distinctive features for strategic TA models, since extendibility is primarily determined by the adopted demand model, while theoretical consistency with microscopic TA models is always possible at some level of aggregation.

[3] Even when the exact stable state never occurs on any given day in reality (because of e.g. differences in perception and lack of information among travelers), for planning purposes it is important that the model outcomes are not influenced by these effects: they must remain comparable.

the contribution of each of the different model components in terms of scenario outputs. This requires model components that can be isolated and that are mathematically tractable (i.e.: all calculations can be verified given the theory behind it). Finally, computational efficiency, low input requirements and applicability allow for fast calibration and application of the model on any network. These desired properties for TA models within large-scale strategic transport models are summarized in Table 1 for later reference.

<<insert table 1 here>>

A quick-scan of strategic transport model systems of large urban areas in Western Europe shows that in general two types of traffic assignment models are being used. Most strategic transport model systems use traditional static traffic assignment (STA) models (e.g. Paris, Berlin, Amsterdam, Lisbon, Vienna, Copenhagen, Rotterdam, The Hague). These models are computationally efficient, have low input requirements and are robust, tractable and accountable. However, they are not sufficiently accurate (and thus applicable) in congested conditions, because they do not capture flow metering and spillback effects due to congestion (Flötteröd 2015). To the best of the authors' knowledge, there are three (quasi dynamic) traffic assignment models in use in strategic transport model systems that, to some extent, capture flow metering and spillback effects: QBLOK (Bakker, Mijjer, and Hofman 1994) used solely in the Dutch national models system, Saturn (Van Vliet, Hall, and Willumsen 1980) used in e.g. London Highway Assignment Models, and the blocking back assignment in PTV VISUM (Bundschuh, Vortisch, and Van Vuuren 2006) used in e.g. the UK west midlands PRISM model and Flemish strategic traffic models. Although they provide more accuracy than traditional STA models, all three models suffer from a solid theoretical basis, as they are merely presented as algorithms, while the underlying

mathematical problem formulation and assumptions are not specified. This leads to poor accountability and makes calibration of parameters using observed data cumbersome and model-specific. Furthermore, queues and delays predicted by these models are not consistent with (simplified) kinematic wave theory (Lighthill and Whitham 1955; Richards 1956 or Newell 1993), causing poor mathematical tractability.

Over the last decades, there has been much emphasis on development of dynamic traffic assignment (DTA) models and their application in the operational (and sometimes tactical) context. However, as suggested by (TRB 2011; Szeto and Lo 2006; Peeta and Ziliaskopoulos 2001) DTA models lack the convergence properties that are needed for applications within the strategic context. This means that the robustness and accountability of these models is insufficient to be used in strategic transport model systems. Indeed, researchers and practitioners state that a duality gap value (DG, the metric most used to measure the level of disequilibrium) of 1E-04 or lower is needed in strategic context (Boyce, Ralevic-Dekic, and Bar-Gera 2004; Han et al. 2015; Brederode, Heijnickx, and Koopal 2016; Caliper 2010), whereas, to the best of the authors' knowledge, no DTA algorithms exist that can converge to this level on realistic congested networks of reasonable size. Furthermore, the existence of a time dimension within DTA models is a major contributor to their high computational cost and memory usage and therefore limited scalability. The time dimension also causes DTA models to require much more input data in comparison with STA models, because demand-matrices (or demand models that can deliver these), traffic counts and route choice parameters (may) become time dependent. This input data is often not available, especially for longer-term scenarios (i.e. 5-20 years into the future). A quick-scan of DTA models in the strategic context (especially in the US) confirms that these model

applications are all relative small-scale (<1300 centroids) and most do not converge well[4].

Based on these considerations, we first of all argue that traditional STA models are insufficiently accurate to be applied on strategic transport model systems with structural congestion, whereas the accountability and robustness of existing quasi-dynamic assignment models is questionable, and their calibration cumbersome due to the lack of a solid theoretical basis. Second of all, we argue that DTA models are sufficiently accurate to describe congestion effects, but their low computational efficiency, high input requirements and poor robustness and accountability prohibit application in large-scale strategic models. To overcome these shortcomings, STAQ (Static Traffic Assignment with Queuing): an assignment model for road traffic within strategic transport models was developed as an alternative to the traditional STA model, providing more accuracy on congested networks without reducing robustness, applicability and accountability and without increasing input requirements, whilst keeping computational requirements to acceptable levels. This makes the model suitable for applications where both static and dynamic assignment models may fail, i.e. strategic applications on large-scale congested networks.

STAQ consists of two submodels, both consisting of several components. For each component variations are possible which, combined, result in a large set of possible implementations of STAQ. We shall first describe the concept, methodology

[4] e.g. Miami (284 centroids, DG 1E-02, Florida international university et al 2013), Sacramento (1279 centroids, DG 1E-01, TRB 2014), Sydney (1131 centroids, DG 5E-02, Duell et al 2015), Ljubljana (323 centroids, DG 1E-02, PNZ 2009), Sundsvall (330 centroids, DG 1E-04, Contenti 2013)

and implementation using STAQ in its most accurate[5] (or 'reference') form. Thereafter, the role and performance of variations applied in this paper will be described (from section 2.5 onwards). All model variations represent simplifications, thus leading to lower accuracy, but at the same time benefitting from equal or higher tractability, accountability, robustness or efficiency, or equal or lower input requirements compared to the reference form.

The mathematical problem formulation of STAQ, its theoretical advantages over STA and DTA models as well as earlier versions of its solution algorithm have been described before by (Brederode, Bliemer, and Wismans 2010; Bliemer et al. 2012; Bliemer et al. 2013). Since the most recent publication, there have been a few minor methodological improvements, and various STAQ variants have successfully been tested and put to practice on several large-scale real-life strategic models[6]. Now that mathematical development and conceptual testing of the model is completed, this paper focuses on the key aspects of practical model applications. The main contributions of the paper are 1) to provide a complete and up-to-date description of the model concept, methodology and implementation, 2) to (explicitly) show how, and to what extent the model addresses the shortcomings of STA and DTA models in practice in the strategic context, and 3) to demonstrate the model performance in terms of the desired properties listed in Table 1, both for the reference model form and several model variations,

[5] by the definition from Table 1

[6] Note that (Bliemer et al. 2014a) did test their assignment model on two large scale networks, but that paper only describes a single variation of STAQ that does not account for queue spillback nor junction modelling, and only the MSA averaging scheme is used.

thereby helping model users to choose the variation best suited for their application. The latter is done using (small) theoretical and (large) real-life model applications.

The remainder of this paper is organized as follows. Section 2 describes the models concept and methodology, and Section 3 describes the algorithmic implementation of its reference form. Throughout both sections, where appropriate, we discuss how STAQ qualitatively overcomes the shortcomings of STA and DTA models and adheres to the desired properties. Then Section 4 demonstrates the model performance also quantitatively based on recent real-life model applications conducted in the past 5 years, and presents how different model variations affect the desired model properties. We end with discussion and conclusions in Section 5.

2 Concept and methodology of STAQ

This section describes the concept underlying STAQ as well as its methodology and variations. STAQ is implemented in c++ and available and applied for policy makers as a part of OmniTRANS transport planning software since early 2015. Sections 2.1 till 2.4 provide insight into how STAQ combines assumptions from static and dynamic assignment models to satisfy the desired properties for strategic transport model systems (Table 1) and form a prerequisite for the sections afterwards. Section 2.5 describes the STAQ variations used in this paper.

2.1 General concept and properties

STAQ achieves the desired properties in Table 1 by combining some (implicit) assumptions from STA models with some assumptions from DTA models. In order to include flow metering and spillback effects of congestion, its network loading submodel (section 2.3) respects strict capacity (maximum flow) and storage (maximum density) constraints respectively. Note that strict capacity constraints can be added straightforwardly as mathematical constraints to the STA model formulation (e.g.

Larsson and Patriksson 1999), but when added for all links, this yields unrealistic (equilibrium) solutions without congestion, because all links are forced into free flow regime. Furthermore, solving the model becomes much more tedious (Nie, Zhang, and Lee 2004). DTA models on the contrary simulate the full on-set and off-set of congestion due to the flow and density constraints, but calculate much more (dynamic information) than required in the strategic context at the cost of computational efficiency, convergence and scalability properties. STAQ resolves this trade-off by including strict capacity and storage constraints (as in DTA models), but excluding the time dimension by assuming stationary travel demand throughout the study period (as in STA models) and instantaneous propagation of unconstrained flow (as STA models assume for all flow). It does this in a way maintaining most of the robustness, accountability and low level of computational and input requirements from STA models. It uses a concave two-regime fundamental diagram for the relation between speed, flow and density on link level (section 2.3.1), and uses an explicit node model to describe merging, diverging and crossing flow interactions on node level (section 2.3.2). Additionally to the node model, to allow for application in the urban context, STAQ has a junction modeling component, taking into account capacity and delay effects on the level of turning movement caused by e.g. traffic rules, geometry and/or signal schemes on junctions (section 2.3.3) allowing for model application on both urban roads and motorways. Furthermore, STAQ allows for multi-user-class assignment, where each vehicle class has its own route choice parameters, free flow speed and set of network restrictions making the model applicable for all vehicle classes.

The specific assumptions in STAQ are beneficial for its purpose to overcome the shortcomings of STA and DTA models in the strategic context, but also have consequences for its usage and interpretation of its outcomes. First (contrary to STA

and similar to DTA), its strict capacity constraints and explicit node model can lead to residual traffic: traffic that cannot reach its destination within the studied period. Second (similar to STA and contrary to DTA), its omission of a time dimension means that all model results (e.g. flows, travel times, densities) are averages over all travelers departing in the study period. Third (similar to STA and contrary to DTA), it forces the modeler to make an assumption on the network state before and after the study period, as there are no warm up or cool down periods to take care of this. On the one hand, just like static models implicitly assume, STAQ assumes an empty network before and zero demand after the study period. On the other hand, all travel time (and contributions to density and flow) of traffic that departed within the study period is accounted for in the average outputs, also when part of a trip takes place after the end of the study period (the latter cannot occur in STA models).

2.2 Modelling framework

The assignment model is split into two submodels: network loading, and routing. The network loading submodel uses route-specific travel demand to compute the resulting (route) travel times, whereas the routing submodel uses route travel times to compute the resulting travel demand per route. As shown in (Bliemer et al. 2012; Brederode, Bliemer, and Wismans 2010), the network loading submodel of STAQ can be seen as a static version of the generalized link transmission model of (Gentile 2010).
Therefore, STAQ is categorized using the framework for macroscopic DTA models displayed in figure 1 (adapted from (Cascetta 2009)). Note that STAQ uses a route submodel that is common to macroscopic DTA models, but has a very different network loading submodel. Note that the unit of demand is the number of car equivalents (in case of single-user-class assignment) or the number of vehicles per user-class (in case of multi-user-class assignment).

The remainder of this section describes the model components within both the network loading submodel (further elaborated in section 2.3) and the route submodel (as used in the case studies in section 2.4). Note that mathematical definitions of the different model components are omitted in this paper, as these have been described before in other publications (Bliemer et al. 2014a; Raadsen, Bliemer, and Bell 2016). Instead, we provide references to those publications, and here conceptually elaborate how the various components are combined within the model and its variations. The model variations that are used and/or tested in section 4 are described in section 2.5.

<<insert figure 1 here>>

2.3 Network loading submodel

The network loading submodel of STAQ consists of two phases that both use the same node and junction model components, but use a different link model component as to the adopted fundamental diagram.

First, *the squeezing phase* models the effect of the flow metering of bottlenecks using the path-based network loading model with strict capacity constraints as described in (Bliemer et al. 2014a). This model assumes a fundamental diagram with in the density-flow plane a concave free-flow branch and a linear horizontal congested branch in the link model (Figure 2, middle) implying vertical queues on nodes for which the node model calculates active capacity constraints. Note that the squeezing phase implicitly assumes instantaneous flow propagation for all flow that is not held up in queues just like STA models assume for all flow (in free flow and congested state).

Second, *the queuing phase* models the effect of the spillback and secondary effects of bottlenecks using the event-based generalized dynamic link transmission model described in (Raadsen, Bliemer, and Bell 2016), assuming stationary demand and initial in- and outflow rates and fixed turn-fractions derived from the turn flows

calculated in the squeezing phase. This model assumes a concave free-flow branch and a linear downward-sloping congested branch in the link model (Figure 2, right) implying storage constraints, while the node model calculates the effects of changes of in- and outflow rates on adjacent links. Note that although no 'normal' time dimension exists, the queuing phase uses a time dimension internally (referred to as 'queuing time') to capture the amount of spatial interaction between all the different spillback and flow metering effects. A specific queuing time however, cannot be related to, or interpreted as, a specific moment in time because the queuing phase starts with the instantaneously propagated flow rates from the squeezing phase, and demand is assumed to be stationary. Only the (demand averaged) flow rates and travel times are consistent with the assumptions in STAQ and as such form the primary output.

The most important reason for splitting the algorithm into two phases is to maintain scalability when calculating spillback and secondary effects of bottlenecks. Additional reasons are that the squeezing phase compensates for the lack of a pre-study-period warm-up and that flow metering and spillback effects can be analyzed separately.

2.3.1 Link model

Figure 2 illustrates the density-flow relation of the fundamental diagrams of STAQ (middle and right) with the BPR-type travel-time functions that are typically used in STA models (left). In the figure, the free-flow branch of each diagram is blue and the congested branch is red. Note that the travel time functions in STA models have no capacity constraint. Hence their fundamental diagram does not contain a congested branch. Considering the fundamental diagram of STAQ– squeezing (middle): it has a free-flow branch very similar to that of the STA model, but it has a congested branch that satisfies the capacity constraint on maximum flow and as such accounts for flow metering. However, because there is no constraint on maximum density, vertical queues

are implied (i.e. point queues with infinite density). The fundamental diagram of STAQ – queuing (right) has the same concave free-flow branch as STAQ – squeezing, and a congested branch that complies with both the capacity constraints on maximum flow (accounting for flow metering) and maximum density (accounting for spillback). The mathematical formulation of this Quadratic-Linear (QL) fundamental diagram can be found in (Bliemer et al. 2014b).

<< insert figure 2 here>>

2.3.2 Node model

The node model seeks for a consistent solution in terms of flows transferred over the intersection, assuming individual flow maximization and accounting for all demand and supply constraints of the adjacent links. This means that the node model can transfer the effect of capacity restrictions on downstream links to upstream links and can transfer the effect of changes in demand on upstream links to downstream links. STAQ uses the node model proposed in (Tampère et al. 2011) and (Flötteröd and Rohde 2011) which complies to a set of generic requirements for first order macroscopic node models described in the first paper. Later, (Smits et al. 2015) generalize all feasible supply distribution schemes complying to the requirements of (Tampère et al. 2011) into a family of macroscopic node models, of which the model used in STAQ is a member.

2.3.3 Junction model

The junction model is an extension of the node model. It has two purposes in STAQ. Firstly, it accounts for the effect of limited supply due to conflict points on the junction itself (i.e. crossing flows), since the node model itself only accounts for flow restrictions due to merge and diverge interactions between flows leaving in-links and entering out-links to the node. The junction model thus imposes further constraints onto the node model. Secondly, the junction model calculates travel-time delays due to passing the

junction, caused by conflicts on turning-movement level depending on junction type (e.g. roundabout, prioritized or signalized). In the current implementation, the junction model uses the method described in (Bovy 1991) for roundabouts and the Highway Capacity Manual (TRB 2000) for other junction types[7]. The junction model first calculates effective turn capacities given the local demand, and then derives turn delays using these capacities. These turn delays consist of deceleration and acceleration delays when approaching and leaving the node, and delays due to direct interference of other traffic or signaling on the node itself. Note that delays as a result of queuing are excluded from the junction model because its turn capacities are used in the node model that potentially triggers the link model to account for queuing.

2.3.4 Travel-time calculator

The travel-time calculator is used to derive travel times from the output as calculated by the link, node and junction models. The travel-time calculator has two functions. Firstly, it uses cumulative inflow and cumulative outflow curves created by the link model of each link to derive the link travel time (e.g. (Long, Gao, and Szeto 2011)). Note that in this way, the effects of queues and spillback as a result of demand and (internal) supply constraints imposed by the node and junction models are automatically accounted for. Secondly, it translates these link-based travel times into route-based travel times, and includes delays from the junction model. It calculates the travel time of a route from an origin to a destination; flow averaged over all car equivalents *departing* within the study period. It includes the travel time experienced after the study period by car (equivalents)

[7] The junction modelling component is currently being updated to also include the US HCM2010, the German HBS2015 and other state of the art junction models as part of the research described in (Bezembinder, Wismans, and van Berkum 2015).

that did not reach their destination within the study period. This is achieved by setting outflow from all centroids to zero after all demand is put on the network, and letting the queuing phase continue until all traffic has reached its destination.

2.4 Route submodel

The advantages of STAQ are derived more from its unique network loading submodel than its route submodel, and hence the latter is interchangeable. Nevertheless, for sake of completeness and clarity we describe the route submodel here briefly.

2.4.1 Route set generator

The route set generator creates routes based upon a digitized transport network. It uses the Dijkstra algorithm to find the shortest path between each origin-destination (OD) pair. By use of 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), alternative routes are generated. 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.

2.4.2 Route choice model and convergence criterion

The route choice model uses the generalized route costs (based on the network loading submodel) to compute route fractions for all route alternatives between an OD pair. Here we assume random utility maximization with perception errors, and hence use the multinomial logit (MNL) model to calculate route choice probabilities, such that route demand f_p is defined by:

$$f_p = \exp(-\mu_{od} c_p) / \sum_{p' \in P_{od}} \exp(-\mu_{od} c_{p'}) D_{od}\,, \tag{1}$$

where c_p is the route cost on route p, D_{od} is the travel demand for OD pair od and μ_{od}

is the scale parameter describing the degree of travelers' perception errors on route travel times (where perfect knowledge is assumed when μ_{od} approaches infinity). Here (and in most real world applications) μ_{od} is determined using a global scale parameter μ normalized over ODpairs by $\mu_{od} = \mu / \min_{p \in P_{od}} c_p^0$, where c_p^0 is the free flow cost on route p. This normalization ensures that the relative effect of perception errors is the same on all OD pairs (regardless of their average route travel time).

Together with the feedback loop in Figure 1 and an averaging scheme, this leads to flow assignment complying to the stochastic user equilibrium (SUE) . To check for convergence, we use the adapted relative duality gap as derived in (Bliemer et al. 2013) that accounts for perception errors and thus reaches zero upon convergence when using the MNL route choice model:

$$G = \frac{\sum_{(o,d)} \sum_{p \in P_{od}} f_p (c_p + \mu_{od}^{-1} \ln f_p - \psi_{od})}{\sum_{(o,d)} D_{od} \psi_{od}}, \tag{2}$$

where $\psi_{od} = \min_{p \in P_{od}} [c_p + \mu_{od}^{-1} \ln f_p]$ represents the minimum stochastic path cost.

2.4.3 Route demand calculation and averaging scheme

The route demand calculation component has two functions. Firstly, it computes the travel demand at route level, based on the OD-demand and route fractions. Secondly, it enforces and speeds up convergence by averaging route demands over iterations. STAQ uses the method of self-regulating averages (SRA) to average route demands over iterations. SRA complies to the convergence conditions derived by (Blum 1954) stating that the influence of priori iterations must decrease in every subsequent iteration. SRA is described in detail in (Liu, He, and He 2009) and tends to provide fast convergence with high precision. The concept of SRA is to let the influence of prior iterations decrease with either a larger or smaller step size depending on the difference in levels of

disequilibrium (in terms of 'excess' vehicle hours) between the last and second-to-last iteration.

2.5 STAQ variations

As mentioned in section 1, model components can relatively easily be exchanged or adapted thereby creating STAQ variations. A variation is a STAQ model application in which one or more of the components described in subsections 2.3 and 2.4 are replaced or altered. Variations are applied to change the balance between accuracy and applicability on the one hand and input requirements, tractability, accountability, computational efficiency and convergence properties on the other hand. Below, the five variations that will be used in section 4 are described. Note that each variation can be applied in combination with other variations; e.g. in section 4.3, three variations are used to construct the twelve different combinations listed in Table 3. Further note that more variations are feasible (and have been implemented), but are omitted here for reasons of relevance and brevity.

The first variation mainly influences the balance between accuracy and convergence properties by omitting the queuing phase until equilibrium has been reached, and then apply it only in the last iteration to translate the equilibrated vertical (point) queues into horizontal (spatial) queues. When applying this variation, route choice is based on vertical queues, and effects of horizontal queues are only included in the final network traffic states (i.e. link flows, speeds and densities). This variation is tested in subsections 4.3 and 4.4. It is expected to improve convergence and thereby computational efficiency at the expense of accuracy and applicability, especially around heavy bottlenecks where in reality spillback would influence route choice.

The next two variations also mainly influence the balance between accuracy and convergence properties and are related to the junction model. Firstly, flow restrictions due to junction modeling can be omitted, in which case only the turn delays are taken into account in the travel time calculator. Secondly, junction modeling can be omitted entirely, in which case no additional flow constraints are imposed on the node model nor are turn delays considered in the travel time calculator. Both variations are tested in subsections 4.3 and 4.4.

The fourth variation is to increase model tractability at the expense of convergence properties by applying the MSA, instead of SRA, averaging scheme. Because MSA uses predefined fixed step sizes that are independent of results of previous iterations it is much easier to verify its outcomes. The effect on convergence properties (and thereby computational efficiency) is discussed in subsections 4.3 and 4.4.

The fifth variation is also to increase model tractability and relates to the form of the fundamental diagram. Instead of the QL diagram, the triangular fundamental diagram proposed by (Newell 1993) can be used. This diagram implies no delays in the free flow branch, which means that it is less accurate in these circumstances. The diagram is especially useful to demonstrate tractability, because a flow/density tuple can easily be calculated using simple geometric algebra as will be shown in subsection 4.1.

3 Model implementation

This section describes the implementation of STAQ in terms of input, algorithm and output. Recall from section 1 that all variations are simplifications of the reference form. This section thus describes the normative input requirements, most advanced algorithm and most accurate output of the model. In line with section 2, mathematical or

pseudo-code representation of the model is omitted here, as these have been provided before in publications to which we shall refer.

3.1 Model input

STAQ needs less input than DTA models and only slightly more input than STA models. Therefore, we first describe model input required for STA models, and then describe the additional input required for STAQ.

In STA models, the infrastructure (supply) is described by a (graph) network of the study area consisting of centroids, directed links and nodes. Centroids represent aggregated trip origins and destinations. Links represent road segments and have attributes pertaining to the free flow speed and the theoretical link capacity. Nodes represent merges, diverges and intersections. Only those nodes where junction modeling is applied have attributes, which pertain to the junction type, approach and exit lane configuration and dimensions and optionally the traffic light schema. Travel demand is assumed stationary during the study period and described for each origin-destination pair in a single OD matrix.

Most STA models have (implicit or explicit) link-flow propagation functions that only describe a free-flow branch of the fundamental diagram. To construct the fundamental diagrams for each link (Figure 2), STAQ uses the free-flow speed and capacity like in a STA model to determine the slope and height of the free-flow branch. Additionally, STAQ requires the jam density per lane to determine the point of intersection of the congestion branch with the density axis, and requires the critical speed to determine the slope of the free-flow branch at capacity. Note that the critical speed can be derived from free-flow speeds from an existing STA network and jam density can be derived or assumed based on the average car length. Further note that STAQ does not need a link typing (as most STA models employ), since all link

characteristics are derived from the (link specific) fundamental diagram. STAQ does not need any additional input on the travel demand.

Although STAQ needs little extra input compared to STA models, its strict capacity constraints put emphasis on the required level of precision and accuracy of the input data. Firstly, the strict capacity constraints make it necessary to define the stationary demand matrix more explicitly: it should contain all the traffic that chooses to *depart* in the study period, no matter if it reaches its destination within that study period. This means that when using traffic counts to calibrate the OD matrix, flow metering and spillback effects of congestion should be somehow taken into account (something that is usually *not* accounted for in matrix estimation procedures for static traffic assignment models). Another consequence of this more explicit definition of travel demand is that the modeler will have to think about the translation from the 'real' time-varying travel demand to an 'averaged' or 'peak' travel demand for the study time period, depending on the desired outcomes ('average' or 'peak' flows and travel times). This means the study period length and the static travel demand level should be defined consistently. Note that this is also the case with STA models, but the lack of strict capacity constraints prevents manifestation of erroneous choices[8].

Secondly, (similar to any macroscopic DTA model) when using a variation with the queuing phase and junction modeling, the strict capacity constraints require junctions to be modelled integrally using a single node, and not as an 'expanded node' (i.e. a constellation of short links and nodes that jointly represent a junction). In STA

[8] Because of the lack of strict capacity constraints, no queuing occurs in STA models, which means that the relation between demand and (modelled) delay due to congestion is much less sensitive compared to models with strict capacity constraints, preventing manifestation of erroneous choices.

models, the latter is sometimes done to maintain (digital) network consistency with environmental models. Although not correct, the error introduced in the STA context is relatively small, because only the (additive) turn delays from junction modeling are used to influence route choice within the model. Therefore, the induced error could be traded-off for network consistency. However, because STAQ also uses the turn capacities from junction modeling as strict capacity constraints in the network loading submodel, this trade-off can no longer be made[9].

The effects of the capacity constraints described above can be considered a pain, but they do increase the accuracy of the model substantially by adding flow metering and spillback effects. Furthermore, they force the definitions of travel demand, study period length and junctions to be defined explicitly and more precise, thereby increasing the model accountability.

From the above, we conclude that with respect to the desired property of low input requirements, STAQ requires more input with a higher accuracy than STA models. However these requirements are very modest compared to those of DTA models, and most of the additional input can be derived and refined from STA model input. Hence, STAQ requires much less (precise) input compared to DTA models.

---

[9] Because capacity is not additive, each path using the junction will only be affected by the first turn on the path that forms an active constraint. If this is a turn on a node originating from a 'junction-link' a queue will form on the junction-link, whereas in reality this would be prohibited (on signalized junctions), impossible (on junctions without mid verges) and/or would only occur when a queue formed downstream of the junction spills-back onto the junction. In the first two situations, a queue that in reality would form on the upstream links of the junction is modeled on the junction itself, potentially blocking other turns on the junction. Because junction-links are relatively short, spillback on these links occurs rapidly causing almost instant gridlock on the junction, whereas this would not happen in reality.

3.2 Algorithm description

Below the algorithms underlying the STAQ network loading submodel (left part of Figure 1) are described using flow charts. A full mathematical description of the squeezing and queuing algorithms can be found in (Bliemer et al. 2014a; Raadsen, Bliemer, and Bell 2016) respectively.

The squeezing phase (Figure 3) primarily detects the locations and severity of active bottlenecks in the network, given the demand for all routes from the route submodel. It calculates a consistent set of reduction factors on turning movement ('turn') level that express the fraction of flow that can traverse the turn, given the capacities of the turn itself (as defined by the junction model), the capacity of its downstream link (as defined in the link attributes of the network) and all the reduction factors upstream from the turn (on routes that use the considered turn). The algorithm initializes reduction factors at a value of 1 (so no reduction) on all turns, and continues iterating[10] until on all turns the difference between the flow of the previous and current iteration is small enough. At this stage the final link (in)flows and turn flows are known, and (not shown in flowchart) vertical queues (on turn and node level) and link and route travel times can be derived using the final reduction factors and the route demand. Note that (Bliemer et al. 2014a) have proven that the squeezing phase converges to a unique fixed point under very mild assumptions.

<<insert figure 3 here>>

The queuing phase (Figure 4) adds spillback and secondary interaction effects between queues on the network. It tracks shockwaves through space using link

[10] Note that these are iterations within the network loading submodel (inner loop), not to be confused with iterations between the network loading and route submodel (outer loop).

discretization as in the link transmission model (LTM, (Yperman 2007)), but does so in continuous time (using events) starting at the beginning of the study period. The queuing phase initializes by storing splitting rates derived from the turn flows from the squeezing phase and by translating the reduction factors from the squeezing phase into trigger events containing the flow rate upstream and downstream from the shockwave it represents. Then, the algorithm loop starts by running the link model for each trigger event. The link model updates the cumulative in- or outflow curve of the considered link and uses these to apply simplified kinematic wave theory (Newell 1993) to calculate the release event time: the expected arrival time of the resulting shockwave at the other end of the link. After all trigger events are handled, the release events are placed on the event list that is then sorted ascending by time. Then, the first event is selected and its event time is validated. Validation is needed, because whilst the selected event was on the event list, other events on the same link may have updated its cumulative in- and/or outflow curve. If it is valid, time is set to the event time and the node model of the corresponding link end is run, given the updated in- or outflow rate from the event and the splitting rates stored during initialization. This generates new trigger events at links adjacent to the node, which closes the loop. If it is invalid, the event time is either updated (when other events have sped up or slowed down the shockwave) or the event is deleted (when other events have reversed the direction of the shockwave).

<<insert figure 4 here>>

The assumption of zero demand after the study period (section 2.1) is implemented by artificial trigger events at time T carrying zero flow on all upstream ends of links connecting origins to the network (not shown in flowchart). The algorithm

stops when there are no more scheduled events on the event list, which means that the network is empty.

Note that the number of events in the queuing phase can become quite large in large networks, mainly due to forward moving shockwaves that spread out according to the turn fractions causing the change of flow rate between upstream and downstream end of the shockwave to approach zero quite quickly. To reduce the computational burden at the cost of model precision, the queuing phase can be configured to skip processing trigger events for which the difference between the updated flow rates from the node model are smaller than some threshold value epsilon. (Raadsen, Bliemer, and Bell 2016) discuss the effect of different epsilon values and conclude that a value of 5.0 veh/h provides a good trade-off between computation speed and precision. Throughout this paper we use a far more conservative value of 1.0 veh/h, for which negligible effects are reported in the same paper. Note that because the queuing phase is an event based algorithm, it only does calculations when and where needed. This makes the algorithm much faster compared to regular LTM implementations that evaluate all links in the network for each time step.

3.3 Model output

The primary output of STAQ consists of average flows, speeds and densities on link- and turn-level. All primary output is derived from the cumulative in- and outflow curves that are created in the queuing phase in a way that is consistent with simplified kinematic wave theory (Newell 1993) and the assumptions of STAQ as described in section 2.1. This is illustrated using the example of cumulative flow curves for a link displayed in Figure 5 in which the dents in the cumulative flow curves correspond to the events in the queuing phase leading to an increase or decrease in the in- or outflow rate. This figure exhibits four phenomena directly related to the assumptions from section

2.1.

Firstly, the assumption of instantaneous propagation of unconstrained flow means that the initial inflow rate (the angle of the cumulative inflow curve at queuing time 0) is equal to the route flow per link from the squeezing phase (as defined in Figure 3). It also means that at queuing time 0, this flow rate applies to the entire link, from start to end. This means that the cumulative inflow curve does not start at zero, but at a value equal to the inflow rate times the free flow travel time on the link, to reflect that traffic has reached the link end before the queuing phase starts.

Secondly, due to the strict capacity constraints, the initial outflow rate (the angle of the initial cumulative outflow curve) may be lower than the initial inflow rate due to a vertical queue at the downstream side of the link, which means that it is equal to the route flow per link from the squeezing phase multiplied by the reduction factor of this link.

Thirdly, due to the strict storage constraints, density (the difference between cumulative in- and outflow at any point in queuing time) can never be larger than jam density, and the actual densities and changes in flow rates through queuing time are consistent with simplified kinematic wave theory (Newell 1993).

Fourthly, the assumption of stationary travel demand during a single time period implies that the assignment is finished when on all links, the cumulative outflow curve has reached the unconstrained travel demand for the respective link (i.e. the total demand using this link according to the estimated demand matrix during the study period duration and route choice model). Considering Figure 5, the cumulative inflow curve shows that only after t_1 all travel demand has entered the link. Because $t_1 > T$, demand for this link is being held up by active bottlenecks upstream or due to spillback of the link itself (indeed the cumulative inflow curve shows periods where the inflow

rate is decreased). Similarly, the last vehicle leaves the link at t_2, which includes the delay of all active bottlenecks upstream, delay due to spillback caused by the considered link but also any congestion on the link itself that does not lead to spillback.

<<insert figure 5 here>>

Note that the squeezing and queuing phases both yield flows and speeds, where the output of the squeezing phase is predominantly used internally in STAQ, while the output of the queuing phase forms the primary model output. Further note that output of both phases is consistent with the route choice model, and that the squeezing phase does not yield densities because there exists no (internal) time dimension in this phase.

Other STAQ output consists firstly of vertical queues on turn-level and node-level, as calculated by the squeezing phase. These queues are defined as the number of car-equivalents that depart within the study period and have not yet exited the queue at the end of the study period. Secondly, the junction model yields effective turn capacities and turn delays on turn-level. And thirdly, the route choice model yields all common output on the route-level consisting of route fractions and costs.

4 Demonstration of model properties using case study examples

In this section the properties of STAQ are demonstrated using several model applications, and discussed with respect to the desired properties from Table 1. In sections 4.1 till 4.4 we subsequently discuss: tractability, accuracy in congested conditions and accountability, robustness, and computational efficiency. The sixth desired property regarding input requirements is already discussed in section 3.1. The seventh desired property regarding applicability is already briefly mentioned in section 2.1, but also plays a role in sections 4.2.2 and 4.3.

4.1 Tractability

Recall from section 1 that we have defined tractability as the extent to which the

calculations in each of the components can be verified using the methodology underlying the component or submodel. In this subsection, we demonstrate the tractability of STAQ using the illustrative network displayed in Figure 6, by showing that all calculations can be done and understood using only the law of flow conservation and the shape of the fundamental diagram as underlying methods. For the reader to more easily verify the calculations, in this section the triangular fundamental diagram of Newell is used as a variation on the quadratic-linear (QL) fundamental diagram used by STAQ. Because only the shape of the fundamental diagram (one of the two inputs for demonstrating tractability) of the model variant is different to the reference form, conclusions drawn in this section will also hold for the reference form itself and thus for all variations (since these are simplifications of the reference form).

In the illustrative network, all links are unidirectional and have a length of 2 kilometers and a free flow speed of 100 km/h. Capacities per link are displayed in the middle part of the figure, jam density is set to 180 veh/lane. There is only one OD-pair that has its origin top left and destination top right carrying a stationary travel demand of 8000 veh/h. Four routes exist in this network, shown in the right part of the figure.

<<insert figure 6 here>>

First we show the mathematical tractability of the multinomial route choice model. Assuming $\mu=1/0.14$ and given the free-flow travel-times derived from link lengths, $\mu_{od} \approx 89.28$. Then applying equation (1) yields most vehicles (5867) choosing the shortest route 1, fewer vehicles choose routes 2 and 3 (984 vehicles each) and the longest route 3 is used the least (165 vehicles).

Given these route demands, the squeezing phase (Figure 3) detects that there are potential bottlenecks at the turning movements towards link 9 (demand: 6851 (5867+984), capacity: 3000), link 12 (demand: 6851, capacity: 2500) and link 3

(demand: 8000, capacity: 2000). For the sake of briefness, we only consider the first potential bottleneck here: the diverge upstream from link 9. Without going into details of the node model, one can apply the law of conservation of vehicles here to see that 3851 vehicles will be left in the vertical queue not able to enter link 9 yielding a reduction factor of 0.44 for all vehicles leaving link 4. Because of this queue at link 4 another 646 vehicles on route 3 and 4 towards link 10 are also caught in the same vertical queue, due to the conservation of turning fractions (one of the properties of the node model described in 2.3.2). Further iterations of the squeezing phase yield flows and vertical queues displayed in the left part of Figure 7, where one can verify that for each node, the summation of flow on its incoming links is equal to the summation of flow on its outgoing links plus the vertical queue on the node, proving tractability of the squeezing phase.

Given the flows and vertical queues, the queuing phase (Figure 4) starts out with three initial backward shockwaves. Shock 1 starts from the downstream end of link 12, shock 2 starts from the downstream end of link 9 and shock 3 starts from the downstream end of link 4. The conservation law implies that that shockwave speed is equal to the difference in flows divided by the difference in density in front and behind the shockwave. Using this and the link lengths, one can verify that shock 1 is the first to arrive to its upstream link end (after 446 seconds), whereas shockwave 3 arrives at its upstream link end after 576 seconds, and shockwave 2 arrives at the upstream end of link 9 after 792 seconds. From this moment onwards, links 4, 9 and 12 are spilling back, whereas the other links are in free flow state and derivative shocks are cycling through the two loops in the network. Shock 2 cycles through links 10 (forward), 11 (forward) and 9 (backward), whereas shock 1 cycles through links 12 (backward), 13 (forward) and 14 (forward). After one hour, inflow on all routes is set to 0, triggering a forward

shockwave in link 4 that empties the network. Due to the heavy congestion (more than half of the demand is already being held up at the first bottleneck), it takes another 3 hours before the last vehicle has left link 3.

<<insert figure 7 here>>

To demonstrate how the node and link models work together we analyze shock 1 through time by looking at the cumulative in- and outflow curves of link 12 (Figure 8).

(1) At time 0 the shockwave starts at the downstream end (the slope of the cumulative outflow curve is lower than the slope of the cumulative inflow curve at this time).

(2) 446 seconds later (which is exactly the link length divided by the backward wave speed) the shockwave arrives at the upstream end (the slope of the cumulative inflow curve *decreases*), triggering an update of the node model at the upstream end. Because link 12 is now in spillback state, it can process *less* flow and thus has a *lower* effective capacity.

(3) Because link 12 is the normative link, this means that the reduction factor on link 9 is *decreased*, which also causes *less* flow towards link 13 (due to the conservation of turning fractions) and *less* inflow into link 14 at time 518.

(4) This leads to *less* demand from link 14 to link 3 at time 590, which causes the node model between these links to assign *more* flow from link 12 to link 3 and thus *increasing* outflow (the slope of the cumulative outflow curve slightly *increases*).

(5) The *increased* outflow triggers a backward shockwave, and the events described in step 2 till 5 are repeated, but now starting with the opposite effect causing all words in italics to be replaced by their respective opposites.

Note that each cycle of shockwave 1 corresponds to a downstream event followed by an

upstream event on link 12. These events always occur 446 seconds apart (the time that a backward wave traverses the link), as can be derived from Figure 8. Due to the linear free flow branch of the fundamental diagram, the travel time for shockwaves to move forward through links 13 and 15 is also fixed at 144 seconds (as can be derived from Figure 8 by comparing durations between subsequent event times on the up- and downstream end of link 12). When using the QL fundamental diagram, or when other routes would influence this cycle, these time intervals would vary. Note that from t = 2218 onwards, no more events occur on link 12. This means that the differences between updated flow rates from the node models due to the shockwave that is cycling through links 12, 13 and 14 have become smaller than the epsilon value of 1.0 veh/h. Indeed, the differences in flow rate (the slope of the cumulative in- and outflow curves) in Figure 8 before and after the last events where the epsilon is still greater than 1.0 (upstream at t = 2218 and downstream at t = 1771) is already very small. The cumulative curves in Figure 8 also show that the last vehicle enters link 12 at t=14125 and leaves the link at t = 14435.

<<insert figure 8 here>>

To demonstrate how the average cumulative outflow curve (the red dashed line in the example of Figure 5) is used to calculate the link outflows as displayed in the right part of Figure 7, we acknowledge that only routes 1 and 3 make use of link 12, yielding an unconstrained demand of 6851 vehicles for link 12. From the cumulative outflow curve, we can see that the 6851th vehicle leaves the link at time 14435, which means that the average outflow per hour is equal to 6851/14435 * 3600 = 1709 veh/h, which corresponds to the outflow rate displayed on link 12 in the right part of Figure 7.

In this section we have demonstrated that given a network, all calculations within the route submodel and the network loading submodel and the interaction between these

components can be verified using only the specification of the route choice model, the law of flow conservation and the shape of the fundamental diagram. Such a level of tractability is matched by STA models (using a shortest path algorithm, some link delay function and an averaging scheme), and in theory also by non-heuristic DTA models (e.g.: CTM, LTM). However, in practice, DTA models cannot easily be traced in this way, mainly because they use time discretization requiring all time steps to be traced individually and sequentially requiring very large amount of calculations, even on small networks. Furthermore, time discretization implies discretization errors that make outcomes of these models dependent on the level of precision of their implementation. From this we conclude that STAQ satisfies the desired property of tractability both in theory and practice (whereas only some DTA models do in theory).

4.2 Accuracy in congested conditions and accountability

In section 0, we defined model accuracy in congested conditions as the accuracy of flow metering and spillback effects as well as route choice effects due to congested conditions. In the same section, we defined accountability as the extent to which different submodels can be isolated. To assess both properties, we first isolate the flow metering and spillback effects by comparing congestion patterns (location and severity of queues) on a corridor network without route choice with observed congestion patterns and patterns from STA and DTA models (section 4.2.1). Thereafter, we add route choice effects by looking at congestion patterns and route choice effects in a case study on an urban network with route choice (section 4.2.2). This way, we isolate how the different model components capture the different mechanisms that occur in the transportation network, thus demonstrating the accountability of STAQ. Finally, in section 4.2.3 we show the impact of the model accuracy on the societal value of the

measures taken in the same urban network as used in 4.2.2.

4.2.1 Accuracy of network loading submodel on A12 Gouda – Den Haag

In the following analysis, loop detector data of the A12 morning peak on a representative workday in 2006 are used to compare observed congestion patterns and travel times with model outputs from STAQ. For reference we also compare these with model outputs from an STA model and a second-order macroscopic DTA model (MaDAM, (Raadsen et al. 2010)). We stress here that the DTA model is of second order, meaning that anticipation (deceleration) and relaxation (acceleration) effects are accounted for in this model. Figure 9 shows the A12 corridor network in which there are no route choice alternatives. Also, given that all network nodes are simple on-ramps and off-ramps, no junctions exist in this network. Hence, application on this network focuses on the link and node model within the network loading submodel. The OD-matrix has been calibrated on the observed demand just downstream from knooppunt Gouwe (on the motorway) and on all on-ramps indicated in Figure 9.

The congestion patterns are displayed in Figure 9, showing three active bottlenecks: 1) spillback from the traffic lights around Centrum Zuid, 2) the weaving section between Prins ClausPlein and off-ramp Voorburg and 3) the merge from on-ramp Zevenhuizen. Furthermore, the entire stretch of road between Zevenhuizen and Prins Clausplein is congested due to spillback from bottleneck 2, meaning that any potential bottlenecks along this stretch of road cannot clearly be identified from the data.

The first bottleneck (centrum Zuid) is not reproduced by any of the assignment models because spillback from outside the network is not modeled.

The second bottleneck (Voorburg) is identified by both STAQ and the DTA model. However, both models identify the merge from Prins ClausPlein as the only

problem, whereas in reality the weaving section between Prins ClausPlein and Voorburg also causes problems that are not being picked up by STAQ nor the DTA model. The STA model wrongly identifies multiple links downstream from the true bottleneck as a bottleneck, because there is no flow metering in this model.

The third bottleneck (Zevenhuizen) is identified by the DTA model, causing a flow metering effect that results in a free-flow section between Zoetermeer and Zoetermeer Centrum that is not present in the observed data. STAQ does not detect the bottleneck at Zevenhuizen, although the capacity between Zevenhuizen and Bleiswijk and the demand from Knooppunt Gouwe and on-ramp Zevenhuizen is exactly the same. This must mean that the second order effects due to traffic merging from on-ramp Zevenhuizen lowers the effective capacity causing this bottleneck in reality and the DTA model. The omission of this bottleneck by STAQ causes activation of a downstream bottleneck at Zoetermeer Centrum. The STA model gives some delay at the link downstream from the bottleneck, although capacity has not been reached; meaning that the definition of the BPR function causes this link to be identified as a bottleneck.

Based on this comparison, we conclude that STAQ, contrary to the STA model, successfully detects and models primary bottlenecks, but may overlook bottlenecks that are activated due to second-order and lane-distribution effects. These conclusions hold on any network, since they are a direct result of properties of the network loading submodel.

Although second-order and lane-distribution effects cannot be directly modelled using a first order network loading submodel such as STAQ[11], they could be added to

[11] Note that some second order DTA models (e.g. METANET) contain a correction term for merging sections

the assignment model by decreasing the link capacities on weaving sections and merges following guides like the US Highway capacity manual (TRB 2000) or the Dutch CIA (Rijkswaterstaat 2015). This could be done before the assignment, using merging and weaving proportions from the OD matrix assuming free-flow route choice, or incorporated within the assignment model using the actual proportions from the previous iteration. Note that this problem will mainly occur on motorways, because bottlenecks on urban roads typically occur at intersections.

<<insert figure 9 here>>

4.2.2 Accuracy and accountability of assignment model on case Den Bosch

In this section, the accuracy of STAQ compared to STA models is further analyzed using a bottleneck location close to the city of Den Bosch in the Netherlands. During the AM peak period the bottleneck manifests itself on the A59 motorway from Den Bosch towards Oss around the off-ramp Rosmalen (indicated by the black circle in the left part of Figure 10). In the reference situation, the STAQ results (right side of Figure 10) show a vertical queue between the off and on-ramp and a second, much smaller, vertical queue at the end of the off-ramp, together causing a queue spilling back all the way onto motorway intersection Empel (the upper left of the network cut out area displayed in the figures), whereas the static results only exhibit minor speed drops directly on the bottleneck links.

<<insert figure 10 here>>

For sake of analyses, we consider a network variant in which the capacity of the intersection at the end of the southern off-ramp is increased and an extra lane between the southern off- and on-ramp is added, leading to the assignment results displayed in Figure 11.

<<insert figure 11 here>>

These assignment results lead to the following findings (demonstrating accuracy) and mechanisms (demonstrating accountability) for which the STA and STAQ model results are similar:

(1) The two bottlenecks around the off-ramp are effectively removed as a result of the capacity increase. In STAQ this finding is a result of the removal of an active supply constraint in the node model of the node connecting the motorway and the southern off-ramp and the removal of supply constraints of the junction model of the node at the end of the southern off-ramp.

(2) As a result of 1, the on-ramp itself and all arterial roads towards it are used more (i.e. higher flows). In STAQ this finding is a result of decreased travel times on turning movements over, and links around, the nodes mentioned in bullet 1, which cause the route submodel to increase route-fractions of routes using the on-ramp and adjacent arterial roads.

(3) The southbound traffic crossing the A59 returns from alternative routes to the arterial that uses the intersection with the considered off-ramp (indicated by the increased southbound flow on the arterial from the original bottleneck location). The mechanism causing this is thus the same as in finding 2.

Findings that the STA model results omit, but the STAQ model results do correctly show, thereby demonstrating its better accuracy under congested conditions, are:

(4) On the A59, the queue spilling back from the considered bottleneck towards the northwest is much shorter because the squeezing phase predicts the bottleneck to be much smaller and further downstream, which causes the shockwaves calculated in the queuing phase to travel at a lower speed and over a longer distance towards the northwest. Furthermore, due to increased flow from this

direction (calculated by the route choice model), a new bottleneck is activated at the merge of the motorway intersection Hintham (just west of the original bottleneck location).

(5) On the A59, downstream from the removed bottleneck, the existing bottlenecks intensify, and a new small bottleneck activates at the next off-ramp. This is caused by the increase of the reduction factor at the original bottleneck location as calculated by the squeezing phase in combination with the increase of flow due to the route choice model reacting to lower travel times for eastbound traffic on the motorway.

Comparing the STA and STAQ results we conclude that only effects on the links and nodes where measures were taken and some of the route choice effects of the network variant are captured by the STA model, whereas STAQ also captures the effects up- and downstream from the removed bottleneck. This leads to the conclusion that the addition of flow metering and spillback effects strongly improves the accuracy and realism under congested conditions. This conclusion holds on any network, because it is a direct result of properties of the network loading submodel. Furthermore, we have shown that the STAQ results can be related to (combinations) of model components, demonstrating its accountability. With respect to accountability we conclude that STAQ includes effects of route choice, flow metering and spillback; whereas STA models only include route choice effects. And furthermore, accountability of STAQ is still on a level that makes the results explainable on a level comparable to that of STA models. Also, this section has shown that STAQ is applicable on networks containing both urban roads and motorways.

4.2.3 Accuracy and its impact on the predicted societal value of the measures of case Den Bosch

To demonstrate that the differences between the assignment methods may also (substantially) change the outcomes of a (social) cost benefit analysis, we compare the effect of the network variant in terms of vehicle loss hours per road type for both STA and STAQ assignment results (Table 2). Note that these results are only for illustrative purposes, since no calibration has been performed on either model.

<<insert table 2 here>>

Analysis of this table leads to the following findings:

(1) Although route choice does vary among the two networks and assignment methods (see analysis above), the usage per road type in veh*km is (approximately) the same.

(2) In the STA model most delay occurs on the non-motorways, whereas in STAQ most delay occurs on the motorways. Given the usage and location of bottlenecks (both are concentrated on the motorways in this network) STAQ results are more consistent with the model input, than results from the STA model are.

(3) Both assignment models yield a reduction in vehicle loss hours as a result of the measures taken in the network variant. However, when using STAQ, the reduction is more than twice as large compared to the STA model output (a reduction of 134 vehicle loss hours in the STAQ assignment versus a reduction of 64 vehicle loss hours in the STA model).

For illustrative purposes, the annual societal value of the travel time savings during the morning peak hour induced by the network variant is calculated. Following (Kouwenhoven et al. 2014) we assume an average value of time of €9,- per hour and an

average reliability ratio of 0.6. Furthermore, we assume that per year 260 of these average morning peak hours occur. This means that the societal value of the network variant would approximately be €240.000,- according to the STA model output and €500.000,- according to the STAQ output, an increase of 108%. These findings show that choosing an assignment method that accounts for flow metering and spillback effects has substantial effects on the outcomes of a cost benefit analysis for study areas with structural congestion.

4.3 Robustness

As defined in section 0, we consider a model to be robust when there are no random variables in the model and when it converges to a defined and meaningful stable state. From section 3.2, we know that the model does not contain random variables or stochastic processes. Therefore, in this section we will only look at the convergence of STAQ towards Wardrops' conditions of user equilibrium (Wardrop 1952) (which we consider a meaningful stable state indeed[12]) using the adapted relative duality gap as described in 2.4.2.

Key components within STAQ are chosen or defined to maximize convergence properties. In the route submodel, the stochastic user equilibrium is chosen as the route choice paradigm which means that in each iteration traffic is distributed over all routes

[12] Note that uniqueness of the solution is only guaranteed when the TA model uses an (implicit) cost function that is strictly increasing (theorem 1.8 in (Nagurney 1993)). Just like DTA models, the strict capacity constraints within STAQ cause a violation of this requirement. However, empirical tests show that STAQ approximates the same equilibria in terms of link flows, no matter the start solution.

(instead of choosing one route in the deterministic user equilibrium), leading to better convergence properties on the route level (Bliemer et al. 2013). In the network loading submodel, the node model complies with the two invariance principles described in (Lebacque and Khoshyaran 2013) ensuring that its outcomes are stable under constant link boundary conditions (a numerical example of how this ensures stability is given in (Tampère et al. 2011)). Furthermore, the link model contains no discretization over space or time. This means that its solutions are exact, avoiding any discretization errors as shown in a numerical example in (Raadsen, Bliemer, and Bell 2016).

In the remainder of this section, the convergence of STAQ is assessed using several congested networks taken from strategic transport model systems that normally use an STA model. Largely neglecting the required level of precision and accuracy of the input data for STAQ (described in section 3.1), the travel demand matrices used where taken directly from the original transport models systems, whereas the networks where only refined slightly on locations where effective capacities where incorrect (these errors did never manifest itself in the STA model due to the lack of strict capacity constraints). For each model, a hundred iterations where run for all twelve combinations of the STAQ variations that are known to have substantial influence on the convergence (Table 3). These twelve combinations are built up from two variations regarding the averaging scheme (MSA or SRA), three variations regarding junction modelling (no junction modelling ('NoJM'); take only calculated turn delays into account ('Delays'); take both calculated turn delays and turn flow restrictions into account ('JM')), and two variations regarding spillback effects (see section 2.5 for variation definitions).

<<insert table 3 here>>

An overview of the strategic transport model systems tested is given in Table 4. The models are all strategic, but range from relatively coarse motorway oriented models

(Leuven, NRM-West and NVM), to more fine-grained regional models (BBMB, Vlaams Brabant) and urban models (Breda, Haaglanden). Besides Leuven, all models classify as large-scale by the definition from section 0. Note that the digitized networks of Vlaams Brabant, NVM and NRM-West do not contain modelled junctions (no junction definitions set), and therefore, only combinations 1, 4, 7 and 10 where run for these models.

<<insert table 4 here>>

For each model, the appendix contains a graph that shows the relation between the calculation time[13] and the adapted relative duality gap for each of the STAQ variations tested. Besides showing the trade-off between computational time and convergence, the total computational time needed to do 100 iterations can also be derived from the graphs in the appendix by looking on the vertical axis at the point where the curve stops.

Recall from section 1 that the adaptive relative duality gap should be lower than 1E-04 for the assignment mode to produce outcomes that are suitable to be used in the strategic context. From the graphs in the appendix we conclude that that almost all runs without spillback converge sufficiently within 100 iterations when using the SRA averaging scheme. Models Vlaams Brabant and NVM are the only exceptions, however their duality gap curves do suggest that they would reach 1E-04 when some more iterations would have been conducted. Both models show a lot of bottlenecks and a high percentage of routes affected by them (77% and 91% of all routes respectively; see also Table 5). Further investigation shows that the networks of model Vlaams Brabant and NVM are relatively coarse in relation to its density in urban areas, which can be seen

[13] All runs conducted on a Core I7-950 3.07 Ghz machine with 24 GBytes of memory running Windows 7

when looking at the number of centroids and especially the number of links in relation to the number of inhabitants in the study area. This causes (artificial) problems on locations where centroids representing large and densely populated areas are connected to the network with only a limited number of connectors. This happens especially in the city of Brussels in model Vlaams Brabant, and in the larger cities in NVM. This causes the high number of blocking nodes and large proportions of routes being affected. In turn, this causes high sensitivity of route cost to changes in route demand and thus poor convergence properties. Refining the network around these areas would very likely lead to much better convergence properties.

When using MSA, only the BBMB model converges sufficiently within the first 100 iterations (but only just), all models consistently show a well-known property of MSA: its convergence slows down considerably with higher iteration numbers, which happens long before convergence has been reached. Note that, although far from sufficiently converged, in the initial 10 to 15 iterations, MSA generally outperforms SRA. However, after these initial iterations, broadly when MSA approaches duality gap values between 1E-03 and 1E-02, the convergence properties of runs using SRA are clearly much better; leading to better convergence using far less calculation time.

We now consider the effect of junction modelling on models that have junctions defined in the network and for model variations that have proven to converge without junction modelling (i.e.: variations without spillback and using SRA). The graphs in the appendix show that enabling junction modelling, but neglecting its flow restrictions (thus only adding delays from junction modelling to the route cost) deteriorates high precision convergence properties, but does not prevent any model for reaching the required convergence rate, nor does it increase required calculation times significantly. Applying full junction modelling however does break convergence for Leuven, Breda

and Haaglanden. Although the duality gap curve of Leuven suggests that it would reach 1E-04 when some more iterations would have been conducted. Further investigation showed that on the Breda network, a single junction that flip flops from under- to oversaturation causes the oscillations in duality gap values around 1E-04 that can be seen in its graph. Similar observations were made on the Haaglanden network, although in this model, not a single, but several (clustered) junctions showed oscillating under- and oversaturation. These observations suggests that methods similar to diagonalization (Dafermos 1980) might resolve this problem; e.g. smoothing or less frequent updating of the flow restrictions from junction modelling.

The appendix also shows that on all models except Leuven, model variations with spillback do not converge sufficiently within 100 iterations: the duality gap keeps oscillating and never drops below 1E-04. Spillback effects are thus the most important cause for non-convergence, which makes sense when realizing that spillback is likely to cause the cost of routes that use link(s) affected by this spillback to become diagonally non-dominant (i.e.: the demand for such a route itself is no longer the main contributor to its cost; instead demand on other routes is), whereas the route choice model and averaging scheme do not anticipate for this. Note that the one run with spillback that does converge to below 1E-04 is a variation with SRA and without junction modelling on model Leuven. Further investigation shows that the Leuven model has relatively low demand (thus violating the requirement of an accurate definition of stationary demand as stated in section 3.1) due to demand matrix calibration conducted in a static context using observations in congested conditions thus causing spillback effects to only limitedly occur.

The findings described above suggest that the model variation #8 (SRA-Delays-NoSpillb) in Table 3 has the best accuracy whilst still converging sufficiently on all

tested models. In some cases/models, full junction modelling (variation #9) can be used without losing sufficient convergence. Also, this section has shown that STAQ is applicable on networks ranging from fine grained urban to coarse motorway networks.

4.4 Computational efficiency

In section 1 we defined computational efficiency as the extent to which run times and memory requirements are acceptable for calibration and application of large scale models. Although no formal criteria exist, a general guideline is that it should be possible to run an assignment for all modes and for all modelled periods in a strategic transport model overnight. Assuming that a single car assignment takes up around 25% of the total computational effort, this means that any assignment should not take longer than three to four hours. With respect to memory consumption we assume that it should be possible to run the assignment on a regular high-end desktop computer with 16 Gigabytes of RAM. In the remainder of this section we look at calculation times and memory usage for the STAQ model variation #8 (SRA-Delays-NoSpillb) on the models in Table 4 as it was selected as the most balanced model variation combination in section 4.3.

Since in the considered model variation combination, the queuing phase is only performed in the last iteration, calculation time per iteration is roughly equal to the calculation time for the squeezing phase. Given the mathematical problem solved by the squeezing phase (Bliemer et al. 2014a), calculation time to run the network loading submodel is mainly proportional to the following variables (column names of variables included in Table 5 in parenthesis): the number of routes *(#routes*), the number of active bottleneck locations *(#blocking nodes*) and their usage *(% of routes blocked*), the severity of active bottleneck locations (e.g.: local demand to capacity ratio per active

bottleneck location) and the strength of the relationships between those active bottleneck locations (e.g.: the number of shared routes per active bottleneck location). Note that the severity and strength of relationships per bottleneck location are omitted from Table 5 since they are hard to capture in a single indicator.

<<insert table 5 here>>

Looking at the calculation time per iteration, we see indeed that it is roughly proportional to the variables mentioned above yielding calculation times varying from 0.03 ms to 0.12 ms per route per iteration for the models tested, which translates to about 30 seconds to 2 minutes per iteration for every million routes. From Table 5 however, no relationship between the number of iterations required and other run properties can be identified, whereas total calculation time is roughly[14] proportional to the number of iterations between route and network loading submodel required for convergence *(#Iterations***)** since the route submodel forms a loop around the network loading submodel (Figure 1).

To explain why no relationship is found between the required iterations and other run properties in Table 5, we look again at the adapted duality gap graphs in the appendix. In these graphs, some models and model variation combinations show strongly oscillating curves (e.g. combinations #8 and #9 of BBMB (after 30 minutes of calculation time) and combination #9 on both the models of Breda (after 2 hours of calculation time) and Haaglanden (after 1 hour of calculation time), which slows down and/or prevents further convergence. Analysis of the adapted duality gap values per OD for these models (leaving out the summation over OD pairs in equation (*2*)) confirms

[14] This holds only roughly, since later iterations contain fewer active bottlenecks, yielding less calculation time required for the network loading submodel.

that the least converging OD pair contributes the most to poor gap values. Using this knowledge, the cause of the oscillations could be traced to a limited set of OD pairs and even to a limited set of bottleneck locations. These bottleneck locations proved to be switching between an active and inactive state over (sets of) iterations. Often, by removing only one of such bottlenecks in the network, the duality gap graph could drop substantially (factors of 10 to 1000's at equal calculation times). This extends the finding in 4.3 that not only single (clusters) of flip flopping junctions can cause oscillating duality gap values, but that it can also occur on bottleneck nodes not being modelled as a junction. Although identified, this phenomenon may substantially delay or even prevent reaching the required level of convergence and it also prevents formulation of a relationship between the run properties and expected total calculation time in Table 5.

To analyse the computational efficiency of the different model components, the share of calculation time per model component for model variation #8 for six of the tested models is displayed in Figure 12. This figure shows that the network loading submodel (link, node, junction modals and travel time calculator) take up most (54%-64%) of the calculation time. This share is much lower than the share of the network loading submodels within DTA models, demonstrating the high computational efficiency of the network loading submodel of STAQ. This also indicates, that efforts to further improve computational efficiency might need to be put into the route choice model. This component now claims a relative large proportion of calculation time (between 32% and 41%), which will only increase when using more advance route choice models than the relatively simple MNL route choice model used here.

<<insert figure 12 here>>

Comparing the total calculation times of the different models with the upper bound of three to four hours, we see that all models except for Vlaams Brabant and NVM exhibit acceptable calculation times. Although not further investigated, probably, the coarseness of these networks in relation to their density described in section 4.3 is likely to be the cause for its poor convergence.

With respect to memory usage, Table 5 indicates that that it is also proportional to the number of routes. On average, the peak memory usage per route is around 3 Kilobytes, which roughly translates to around 3 Gigabytes needed for every million routes, which means that the largest model tested here (NVM with more than 4 million routes) requires 9.4 Gigabytes of RAM, thereby easily meeting the requirement of maximum 16 Gigabytes of RAM.

5 Conclusions and Discussion

In this paper, we have provided a complete description of the concept and implementation of the assignment model STAQ and several variations, along with insight into how the model addresses the shortcomings of STA and DTA models in the strategic context for large congested networks. In line with literature we have defined seven desired properties for strategic transport models for large congested networks, and have shown the performance of STAQ and its variants for each of these seven properties in comparison with STA and DTA models.

5.1 Main conclusions

The different mechanisms that occur in a transportation network when applying STAQ can all be isolated and verified using only the law of flow conservation and the shape of the fundamental diagram as underlying methods, proving that tractability and

accountability of STAQ is comparable to that of STA models and amply exceed that of DTA models.

With respect to the accuracy under congested conditions, we conclude that, contrary to STA models, STAQ successfully detects and models flow metering and spillback effects of primary bottlenecks, with the limitation that STAQ may overlook bottlenecks that are activated due to second-order and lane-distribution effects. STAQ allows for assignment of different vehicle classes and the junction modelling component allows application on both urban roads as well as motorways.

Furthermore, we conclude that when evaluating network scenarios, STA models only capture effects on links and nodes where network changes occur and include some of the route choice effects, whereas STAQ also captures the effects up- and downstream from network changes. It was shown that the addition of these effects causes large differences in terms of vehicle loss hours and thus societal benefits of these types of policy measures. This clearly demonstrates that the addition of flow metering and spillback effects strongly improves the accuracy and realism under congested conditions and that choosing an assignment method that accounts for these effects will have substantial effects on the outcomes of a cost benefit analysis for study areas with structural congestion.

Based on analysis of twelve different model variations on seven large scale strategic transport models of largely congested regions we conclude that STAQ with spillback in the last iteration, full junction modelling and the self-regulating averaging scheme proved to be the optimal variation, providing sufficient realism and convergence (duality gap values below 1E-04) within well acceptable calculation times for five of the seven models tested (ranging from 23 minutes up to 3 hours to achieve equilibrium on a regular desktop pc). A limitation of this model variant is that spillback

effects are not included in the route choice behavior. Adding these effects is possible, but at the expense of convergence. The network of the models Vlaams Brabant and NVM prove to be too coarse in relation to its density, creating artificial congestion locations causing high sensitivity of route cost to changes in route demand and thus poor convergence properties. Refining the network in densely populated areas would very likely lead to better convergence properties for both models.

Input requirements of STAQ are much lower than those of DTA and only slightly higher than those of STA models. Although STAQ needs little extra input compared to STA models, its strict capacity constraints put emphasis on the required level of precision and accuracy of the input data. Most importantly, the definition of the study period and the level of stationary demand in the matrices should be consistent, flow metering and spillback effects in observed data should be taken into account while calibrating the OD matrices, and the hard capacity constraints in STAQ require more accurate capacity values on links and junctions to be coded as a single node. Based on the above, we conclude that STAQ is a viable alternative to the traditional STA model, providing more accuracy on congested networks without reducing robustness and accountability and without increasing input requirements, whilst keeping computational requirements to acceptable levels (as opposed to DTA models). This makes the model suitable for applications where both STA and DTA models are insufficient: strategic applications on large-scale congested networks.

5.2 Recommendations and further research

Based on this research, several improvements in the way STAQ and its variations are being applied are proposed. Most importantly, the development of a STAQ based matrix estimation method that takes flow metering and spillback effects on observed data into account. A first attempt for such a method is described and applied in

(Brederode, Pel, and Hoogendoorn 2014; Brederode, Hofman, and van Grol 2017) respectively. When in place, model systems can properly be calibrated using STAQ which enables more thorough validation of the assignment model comparing its outcomes with observed flows, congestion patterns and travel times for a large urban region. Furthermore, when thoroughly validated, the societal value of the model should be determined by comparing a full cost benefit analysis of one or more existing projects using an STA model and STAQ.

As described in section 4.3, there is still room for improvement on the speed and level of convergence of the model, especially for model variations with full spillback enabled. Several research directions are worth mentioning here. Firstly, the parameters that control the step sizes used within the self-regulating averaging scheme (section 2.4.3) should be calibrated (now the default values from (Liu, He, and He 2009) are used). Secondly, in section 4.4 we have already briefly mentioned that the causes for poor convergence can be traced down towards (sets of) bottleneck locations which is in line with findings in (Levin et al. 2015) for DTA models. This provides a starting point for various possible algorithmic enhancements that try to decrease the changes in demand per iteration for these locations by e.g. constraining changes in demand on OD pairs using sensitive bottlenecks through the route choice model and/or averaging scheme (note that some of these enhancements where already tested as described in Brederode et al., 2016). From this same starting point, it might be possible to develop a method to calculate a rough estimate of the expected convergence properties of a model given its network and level of OD demand.

As pointed out in section 4.4, the calculation time per model component indicate that the network loading submodel of STAQ is relatively fast, such that efforts to further

improve computational efficiency of STAQ are better put into other model components, primarily the route choice model.

With respect to the route choice model, the paired combinatorial logit model (PCL, Pravinvongvuth and Chen 2005) is implemented as a STAQ variation. PCL adds support for route overlap and therefore allows inclusion of more relevant routes and thus is expected to improve convergence. To be able to test this hypothesis an adaptation of the duality gap for PCL (as has been done for MNL in equation 2) needs to be derived.

Finally, a recommendation with respect to the concept of STAQ. In its current form, STAQ effectively adds strict capacity constraints to STA models. However it still assumes stationary demand during a single time period. This means that the 'true' demand should always be averaged or aggregated in some way over the time period. To reduce averaging errors, an extension to STAQ that allows for multiple time periods would be needed. This would close the gap with DTA models further, however at the same time most likely will introduce new problems, such as more input requirements, poor convergence properties and longer calculation times. If these can be accepted or overcome, it would require for residual traffic to be transferred from one period to the next period. Such a mechanism would also solve another problem: residual traffic due to trip durations longer than the duration of the single time period, which can occur when dealing with large networks and/or short time periods.

Acknowledgements

We would like to thank the Dutch municipality of Den Haag and province of Noord-Brabant, and the Flemish traffic centre and Goudappel Coffeng (the Netherlands) for providing the strategic transport models.

References

Bakker, D., P. H. Mijjer, and F. Hofman. 1994. “QBLOK: An Assignment Technique for Modelling the Dependency between Bottlenecks and the Prediction of Grid Lock.” In *Proceedings of Colloquium Vervoersplanologisch Speurwerk, Delft*, 313–332. Rotterdam.

Bezembinder, E.M., L.J.J. Wismans, and E.C. van Berkum. 2015. “Using Decision Trees to Determine Junction Design Rules.” In *Proceedings of the 94th Annual Meeting of the Transportation Board, 11-15 January 2015, Washington DC, USA. (on CD ROM)*., 1–16. Transportation Research Board (TRB).

Bliemer, M.C.J., L.J.N. Brederode, L.J.J. Wismans, and E-S. Smits. 2012. “Quasi-dynamic network loading: adding queuing and spillback to static traffic assignment.” In . Transportation Research Board (TRB). http://purl.utwente.nl/publications/101278.

Bliemer, M.C.J., M.P.H. Raadsen, E. De Romph, and E-S. Smits. 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. http://repository.tudelft.nl/assets/uuid:7ab0d947-bf24-4d6e-bd1d-f85ae682398f/304469.pdf.

Bliemer, M.C.J., M.P.H. Raadsen, E-S. Smits, B. Zhou, and M.G.H. Bell. 2014a. “Quasi-Dynamic Traffic Assignment with Residual Point Queues Incorporating a First Order Node Model.” *Transportation Research Part B: Methodological* 68: 363–384.

Bliemer, M.C.J., M.P.H. Raadsen, E-S. Smits, B. Zhou, and M.G.H. Bell. 2014b. “Quasi-Dynamic Traffic Assignment with Residual Point Queues Incorporating a Proper Node Model.” ITLS working paper. http://www.sciencedirect.com/science/article/pii/S0191261514001246.

Blum, J.R. 1954. “Approximation Methods Which Converge with Probability One.” *The Annals of Mathematical Statistics* 25 (2): 382–386. doi:10.1214/aoms/1177728794.

Bovy, P.H. 1991. “Zusammenfassung Des Schweizerischen Kreiselhandbuchs.” *Straße Und Verkehr* 3: 129–139.

Boyce, D., B. Ralevic-Dekic, and H. Bar-Gera. 2004. “Convergence of Traffic Assignments: How Much Is Enough?” *Journal of Transportation Engineering* 130 (1): 49–55. doi:10.1061/(ASCE)0733-947X(2004)130:1(49).

Brederode, L.J.N., M.C.J. Bliemer, and L.J.J. Wismans. 2010. “STAQ: Static Traffic Assignment with Queing.” In *Proceedings of the European Transport Conference*. Glasgow, UK. http://resolver.tudelft.nl/uuid:2e241e08-1851-4259-8ef8-67f3c4676221.

Brederode, L.J.N., M. Heijnickx, and R. Koopal. 2016. “Quasi Dynamic Assignment on the Large Scale Congested Network of Noord-Brabant.” In . AET 2016 and contributors. https://abstracts.aetransport.org/paper/download/id/4919.

Brederode, L.J.N., F. Hofman, and R. van Grol. 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.” In . AET 2017 and contributors.

Brederode, L.J.N., A. J. Pel, and S.P. Hoogendoorn. 2014. “Matrix Estimation for Static Traffic Assignment Models with Queuing.” *hEART 2014 - 3rd Symposium of the European Association for Research of Transportation, Leeds UK*.

http://repository.tudelft.nl/islandora/object/uuid%3Ad24cfdc2-806e-40ce-97ef-f72012353829?collection=research.

Brederode, L.J.N., A.J. Pel, L.J.J. Wismans, and E. de Romph. 2016. “Improving Convergence of Quasi Dynamic Assignment Models.” In *Proceedings of the 6th International Symposium on Dynamic Traffic Assignment*. Sydney, Australia.

Bundschuh, M., P. Vortisch, and T. Van Vuuren. 2006. “Modelling Queues in Static Traffic Assignment.” In *Proceedings of the European Transport Conference*, 2006. https://stuff.mit.edu/afs/sipb.mit.edu/user/kolya/afs/root.afs/athena/course/11/11.951/oldstuff/albacete/Other_Documents/Europe%20Transport%20Conference/traffic_assignment/modelling_queues_i1596.pdf.

Caliper. 2010. “What Transcad Users Should Know about Traffic Assignment.”

Cascetta, E. 2009. *Transportation Systems Analysis*. Vol. 29. Springer Optimization and Its Applications. Boston, MA: Springer US. http://link.springer.com/10.1007/978-0-387-75857-2.

Dafermos, S. 1980. “Traffic Equilibrium and Variational Inequalities.” *Transportation Science* 14 (1): 42–54.

Fiorenzo-Catalano, M.S. 2007. “Choice Set Generation in Multi-Modal Transportation Networks.” TRAIL. http://resolver.tudelft.nl/uuid:ef3b9c22-b979-4f46-9b02-110c82d67535.

Flötteröd, G. 2015. “Traffic Assignment for Strategic Urban Transport Model Systems.” https://www.researchgate.net/profile/Stefan_Fluegel/publication/280553979_Traffic_assignment_for_strategic_urban_transport_model_systems/links/55b8baaa08aed621de0715f7.pdf.

Flötteröd, G., and J. Rohde. 2011. “Operational Macroscopic Modeling of Complex Urban Road Intersections.” *Transportation Research Part B: Methodological* 45 (6): 903–922. doi:10.1016/j.trb.2011.04.001.

Flügel, Stefan, Gunnar Flötteröd, Chi Kwan Kwong, and Christian Steinsland. 2014. “Evaluation of Methods for Calculating Traffic Assignment and Travel Times in Congested Urban Areas with Strategic Transport Models.” *TØI Report* 1358: 2014.

Gentile, G. 2010. “The General Link Transmission Model for Dynamic Network Loading and a Comparison with the DUE Algorithm.” *New Developments in Transport Planning: Advances in Dynamic Traffic Assignment*, 153–178.

Han, Ke, Terry L. Friesz, W. Y. Szeto, and Hongcheng Liu. 2015. “Elastic Demand Dynamic Network User Equilibrium: Formulation, Existence and Computation.” *Transportation Research Part B: Methodological* 81, Part 1 (November): 183–209. doi:10.1016/j.trb.2015.07.008.

Kouwenhoven, M., G.C. de Jong, P. Koster, V.A.C van den Berg, E.T. Verhoef, J. Bates, and P.M. J. Warffemius. 2014. “New Values of Time and Reliability in Passenger Transport in The Netherlands.” *Research in Transportation Economics*, Appraisal in Transport, 47 (November): 37–49. doi:10.1016/j.retrec.2014.09.017.

Larsson, Torbjörn, and Michael Patriksson. 1999. “Side Constrained Traffic Equilibrium Models— Analysis, Computation and Applications.” *Transportation Research Part B: Methodological* 33 (4): 233–264. doi:10.1016/S0191-2615(98)00024-1.

Lebacque, J.P., and M.M. Khoshyaran. 2013. “A Variational Formulation for Higher Order Macroscopic Traffic Flow Models of the GSOM Family.” *Procedia -*

Social and Behavioral Sciences 80 (June): 370–394. doi:10.1016/j.sbspro.2013.05.021.

Levin, Michael W., Matt Pool, Travis Owens, Natalia Ruiz Juri, and S. Travis Waller. 2015. “Improving the Convergence of Simulation-Based Dynamic Traffic Assignment Methodologies.” *Networks and Spatial Economics* 15 (3): 655–676. doi:10.1007/s11067-014-9242-x.

Lighthill, M. J., and G. B. Whitham. 1955. “On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads.” *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 229 (1178): 317–345. doi:10.1098/rspa.1955.0089.

Liu, H.X., X. He, and B. He. 2009. “Method of Successive Weighted Averages (MSWA) and Self-Regulated Averaging Schemes for Solving Stochastic User Equilibrium Problem.” *Networks and Spatial Economics* 9 (4): 485–503. doi:10.1007/s11067-007-9023-x.

Long, J., Z. Gao, and W.Y. Szeto. 2011. “Discretised Link Travel Time Models Based on Cumulative Flows: Formulations and Properties.” *Transportation Research Part B: Methodological* 45 (1): 232–254. doi:10.1016/j.trb.2010.05.002.

Nagurney, A. 1993. *Network Economics: A Variational Inequality Approach.* Kluwer Academice Publishers, Boston, USA.

Newell, G. F. 1993. “A Simplified Theory of Kinematic Waves in Highway Traffic, Part I: General Theory.” *Transportation Research Part B: Methodological* 27 (4): 281–287. doi:10.1016/0191-2615(93)90038-C.

Nie, Yu, H.M. Zhang, and Der-Horng Lee. 2004. “Models and Algorithms for the Traffic Assignment Problem with Link Capacity Constraints.” *Transportation Research Part B: Methodological* 38 (4): 285–312. doi:10.1016/S0191-2615(03)00010-9.

Peeta, S., and A.K. Ziliaskopoulos. 2001. “Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future.” *Networks and Spatial Economics* 1 (3–4): 233–265.

Pravinvongvuth, Surachet, and Anthony Chen. 2005. “Adaptation of the Paired Combinatorial Logit Model to the Route Choice Problem.” *Transportmetrica* 1 (3): 223–240. doi:10.1080/18128600508685649.

Raadsen, M.P.H., M.C.J. Bliemer, and M.G.H. Bell. 2016. “An Efficient and Exact Event-Based Algorithm for Solving Simplified First Order Dynamic Network Loading Problems in Continuous Time.” *Transportation Research Part B: Methodological*, Within-day Dynamics in Transportation Networks, 92, Part B (October): 191–210. doi:10.1016/j.trb.2015.08.004.

Raadsen, M.P.H., H.E. Mein, M.P. Schilpzand, and F. Brandt. 2010. “Implementation of a Single Dynamic Traffic Assignment Model on Mixed Urban and Highway Transport Networks Including Junction Modeling.” In *DTA Symposium, Takayama, Japan*.

Richards, P.I. 1956. “Shock Waves on the Highway.” *Operations Research* 4 (1): 42–51. doi:10.1287/opre.4.1.42.

Rijkswaterstaat, WVL. 2015. *Handboek Capaciteitswaarden Infrastructuur Autosnelwegen.pdf*. https://staticresources.rijkswaterstaat.nl/binaries/Handboek%20Capaciteitswaarden%20Infrastructuur%20Autosnelwegen_tcm21-76101.pdf.

Smits, E-S., M.C.J. Bliemer, A.J. Pel, and B. van Arem. 2015. “A Family of Macroscopic Node Models.” *Transportation Research Part B: Methodological* 74 (April): 20–39. doi:10.1016/j.trb.2015.01.002.

Szeto, W. Y., and Hong K. Lo. 2006. “Dynamic Traffic Assignment: Properties and Extensions.” *Transportmetrica* 2 (1): 31–52. doi:10.1080/18128600608685654.
Tampère, C.M.J., R. Corthout, R. Cattrysse, and L.H. Immers. 2011. “A Generic Class of First Order Node Models for Dynamic Macroscopic Simulation of Traffic Flows.” *Transportation Research Part B: Methodological* 45 (1): 289–309. doi:10.1016/j.trb.2010.06.004.
TRB. 2000. *(Transportation Research Board) Highway Capacity Manual*. https://en.wikipedia.org/w/index.php?title=Highway_Capacity_Manual&oldid=746856480.
TRB. 2011. “(Transportation Research Board) Dynamic Traffic Assignment - A Primer.” *Transportation Network Modeling Committee*, 1–39.
Van Vliet, D., M.D. Hall, and L.G. Willumsen. 1980. “A SimulationAssignment Model for the Evaluation of Traffic Management Schemes.” *Traffic Engineering & Control*, no. 21 (April): 168–176.
Wardrop, J G. 1952. “Road Paper. Some Theoretical Aspects of Road Traffic Research.” *Proceedings of the Institution of Civil Engineers* 1 (3): 325–362. doi:10.1680/ipeds.1952.11259.
Yperman, I. 2007. “The Link Transmission Model for Dynamic Network Loading.” https://lirias.kuleuven.be/handle/1979/946.

Appendix: duality gap vs calculation time for all tested models and runs

In this appendix, the empirical relation between calculation time (on a Core I7-950 3.07 Ghz machine with 24 GBytes of memory) and convergence is displayed for all 7 models (Table 4) and all 12 model variations per model (Table 3). Each graph shows the runs on one model, and each curve in a graph represents a specific model run, its color and shape indicate the combination of model components tested as displayed in the legend. The reds represent runs using the MSA averaging scheme, the greens represent runs using the SRA averaging scheme. Dashed curves represent runs where spillback is enabled, and continuous curves represent runs without spillback. The three different shades of both reds and greens represent the three different options for junction modelling.

<<insert figure 13 here>>

Tables

Property	Definition
Tractability	The extent to which calculations in each model component can be verified using the theory behind the component or submodel
Accuracy under congested conditions	The extent to which flow metering, spillback and route choice effects caused by congestion are included in the model
Accountability	The extent to which different model components can be isolated and verified
Robustness (1)	The extent to which the model is free from random variables that affect its outcomes
Robustness (2)	The extent to which the model converges to a defined and meaningful stable state
Computational efficiency	The extent to which run times and memory requirements are acceptable for calibration and application of large scale models
Input requirements	The extent to which input requirements are available with acceptable uncertainty for distant future scenarios
Applicability	The extent to which the model is applicable for all vehicle classes and for both urban roads and motorways

Table 1: desired properties and criteria for traffic assignment models within large-scale strategic transport models

	Usage [veh*km]	**Experienced delay [vehicle loss hours]**					
	Both cases	Reference		Network variant		Difference	
Roadtype	Both assignments	Static	STAQ	Static	STAQ	Static	STAQ
Motorways	~57%	334	1130	314	1052	-20	-79
Non-motorways	~43%	1761	339	1717	283	-44	-55
Total	**100%**	**2095**	**1469**	**2031**	**1335**	**-64**	**-134**

Table 2: vehicle loss hours for reference and network variant for both static and STAQ assignment

1 MSA-NoJM-NoSpillb	4 MSA-NoJM-Spillb	7 SRA-NoJM-NoSpillb	10 SRA-NoJM-Spillb
2 MSA-Delays-NoSpillb	5 MSA-Delays-Spillb	8 SRA-Delays-NoSpillb	11 SRA-Delays-Spillb
3 MSA-JM-NoSpillb	6 MSA-JM-Spillb	9 SRA-JM-NoSpillb	12 SRA-JM-Spillb

Table 3: numbering of the twelve combinations of model variations tested

Model	Major cities in study area	model type	Links	nodes	junctions	centroids
Leuven	Leuven (BE)	motorway	2698	1833	587	430
NRM-West	Amsterdam, Rotterdam, The Hague, Utrecht (NL)	motorway	86783	56739	0	3392
BBMB	Eindhoven, Tilburg, Breda, Den Bosch, Helmond (NL)	regional	142336	106780	15979	3321
Breda	Breda (NL)	urban	147253	107984	16241	6043
Haaglanden	The Hague (NL)	urban	140277	94159	3539	5845
Vlaams Brabant	Brussels, Leuven (BE)	regional	34239	23241	0	2999
NVM	all of the Netherlands (NL)	motorway	159920	65272	0	6102

Table 4: properties of models tested, models sorted by size measured in number of routes

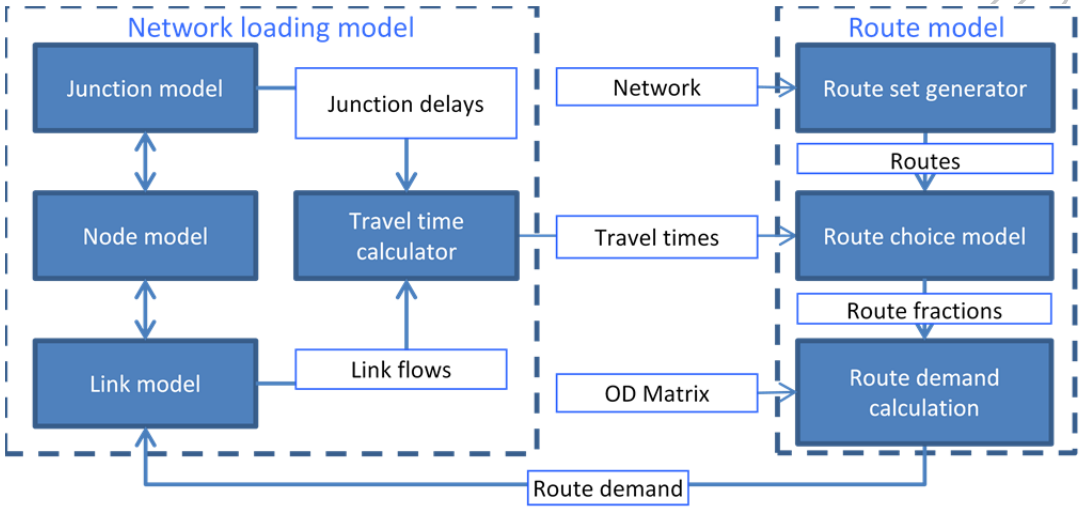
	Run properties					Calculation time			Mem usage	
Model	**Peak period**	**#routes**	**#Itera-tions**	**#blocking nodes**	**% of routes blocked**	**total [hh:mm:ss]**	**per iter [mm:ss]**	**per route per iter [ms]**	**total [Mb]**	**per route [Kb]**
Leuven	PM	74697	49	74	21%	0:01:50	0:02	0.03	404	5.41
NRM-West	AM	1241762	31	863	56%	0:37:19	1:12	0.06	2935	2.36
BBMB	AM	1272227	14	470	27%	0:22:53	1:38	0.08	2245	1.76
Breda	PM	2069672	46	940	53%	3:02:58	3:59	0.12	6470	3.13
Haaglanden	PM	2854246	18	255	32%	1:36:56	5:23	0.11	7631	2.67
Vlaams Brabant	PM	3109173	>100	1354	77%	7:35:37	4:33	0.09	7181	2.31
NVM	AM	4057235	>100	8390	91%	13:43:16	8:14	0.12	9418	2.32

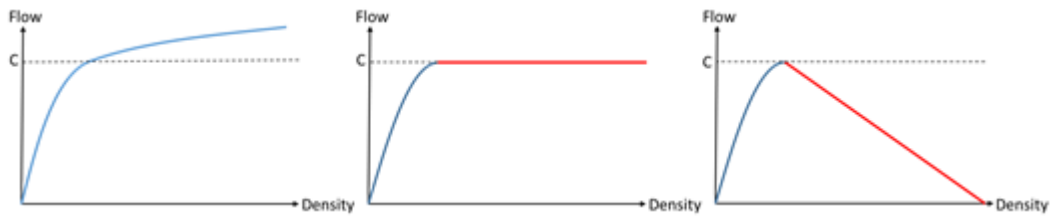
Table 5: calculation times and peak memory usage of model variation combination #8 for all tested models

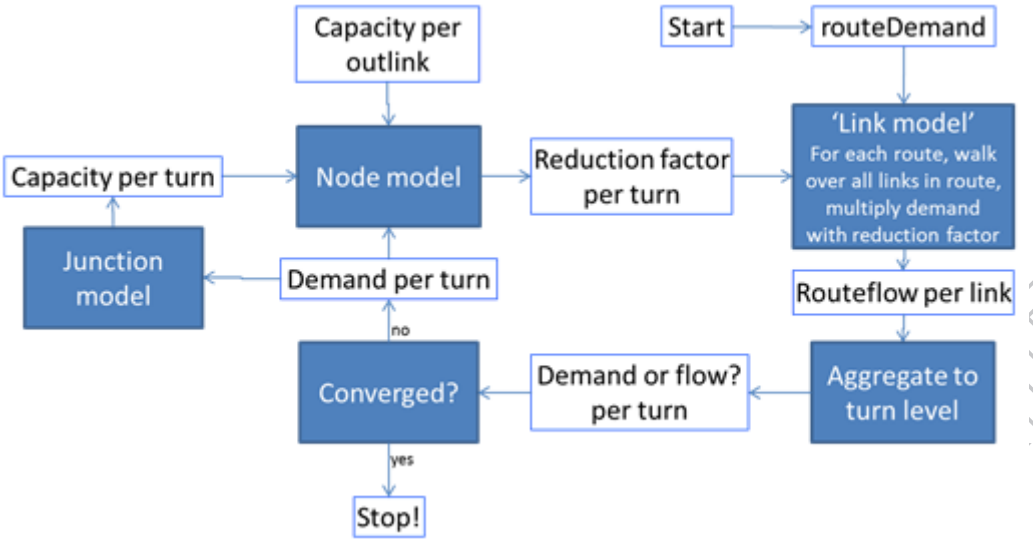
Figure captions

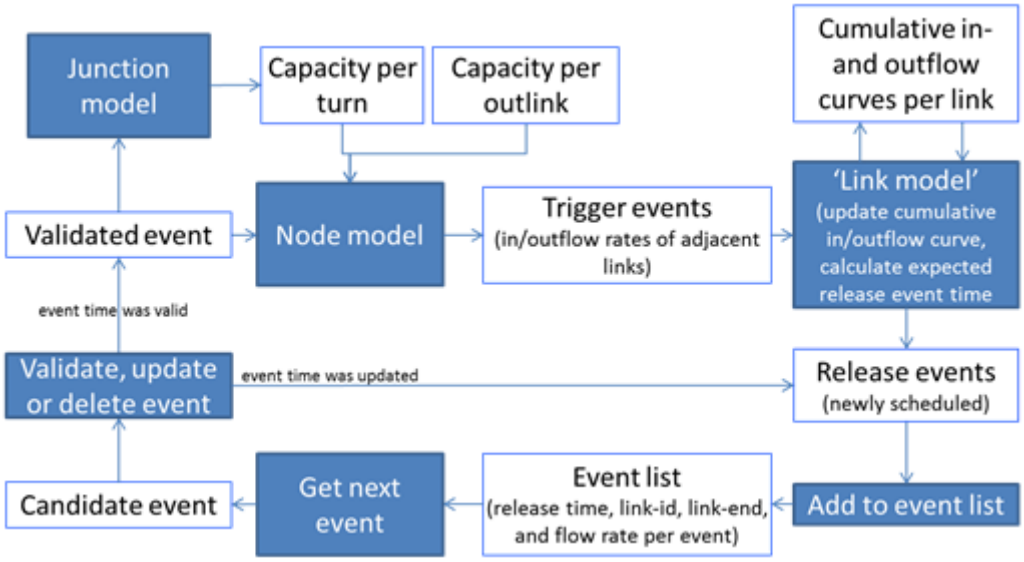
- Figure 1: STAQ modeling framework (adapted from Cascetta, 2009)
- Figure 2: Fundamental diagrams: used in static (left), used in STAQ – squeezing (middle), used in STAQ – queuing (right)
- Figure 3: flowchart of squeezing phase
- Figure 4: flowchart of queuing phase
- Figure 5: example of cumulative flow curves of a link as calculated in the queuing phase
- Figure 6: network with link numbers (left), link capacities (middle) and free flow travel times per route (right) of toy network
- Figure 7: results of iteration 1; inflows (bandwidths / black font) and vertical queues (pie charts / blue font) from squeezing phase (left); outflows (bandwidths) and relative speeds (colours, see legend) from queuing phase (right)
- Figure 8: cumulative in- and outflow curves for link 12. Dots represent events in the queuing phase
- Figure 9: comparison of observed and modelled congestion patterns on the A12 motorway between Gouda and Den Haag
- Figure 10: assignment results for reference scenario; static (left, black circle indicates bottleneck location) vs STAQ (right). Bandwidth colours: modelled speed as ratio of free flow speed; Bandwidth widths: modelled flow; Blue circles in STAQ results (right): vertical queues (radius indicates queue size)
- Figure 11: assignment results of the network variant; static (left) vs STAQ (right)
- Figure 12: calculation time shares per model component

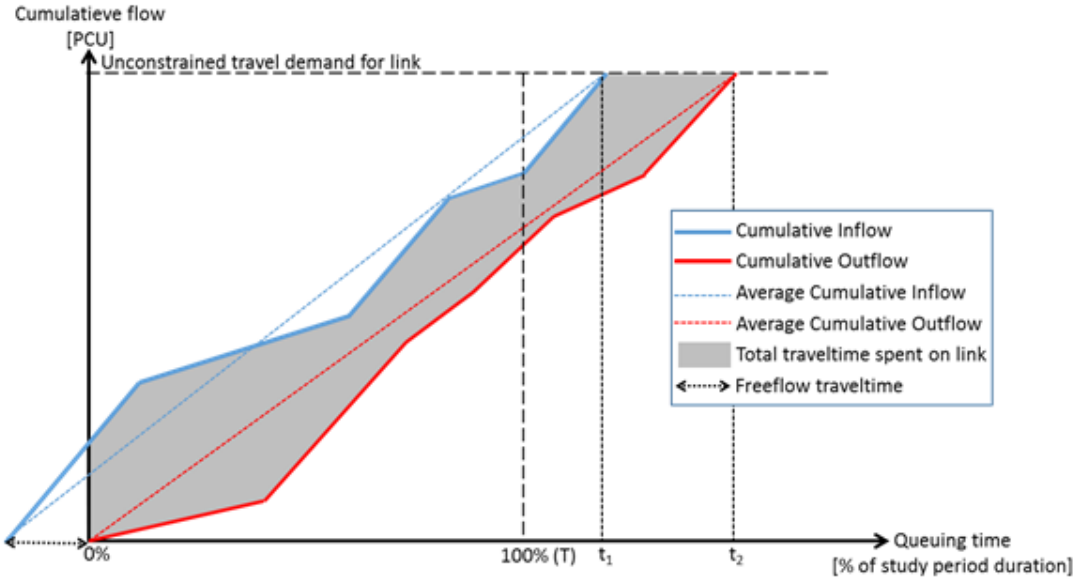
- Figure 13: empirical relation between calculation time and convergence for all 9 models and all 12 model variations

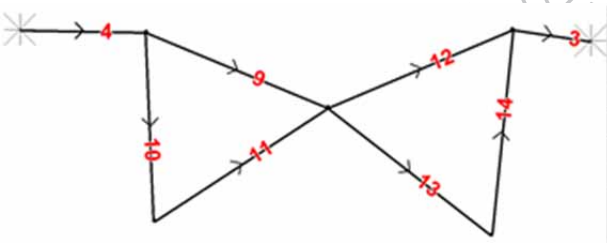






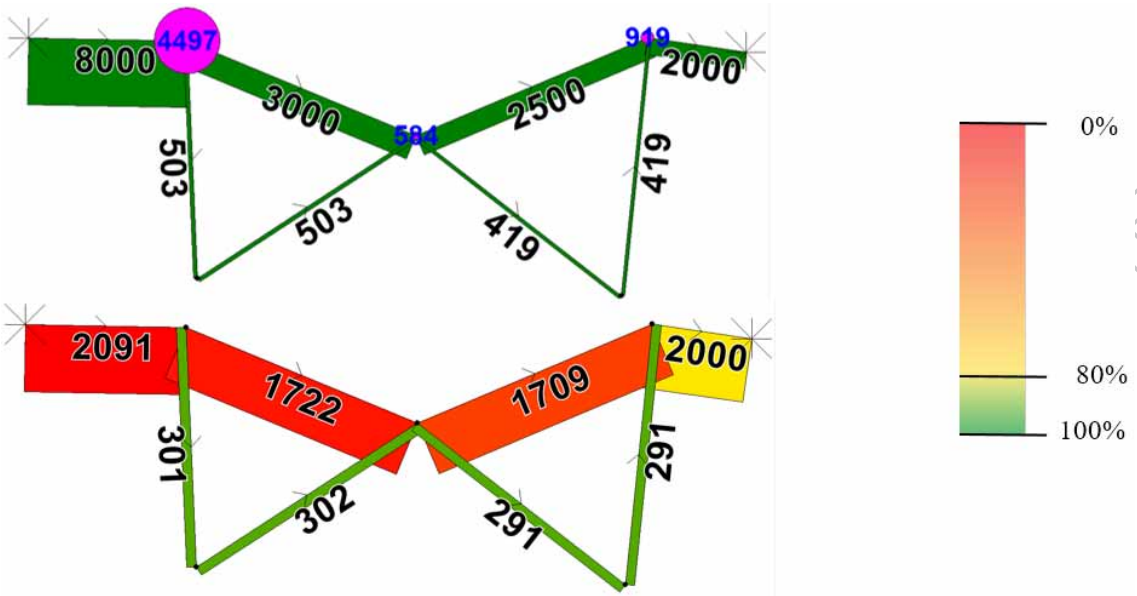


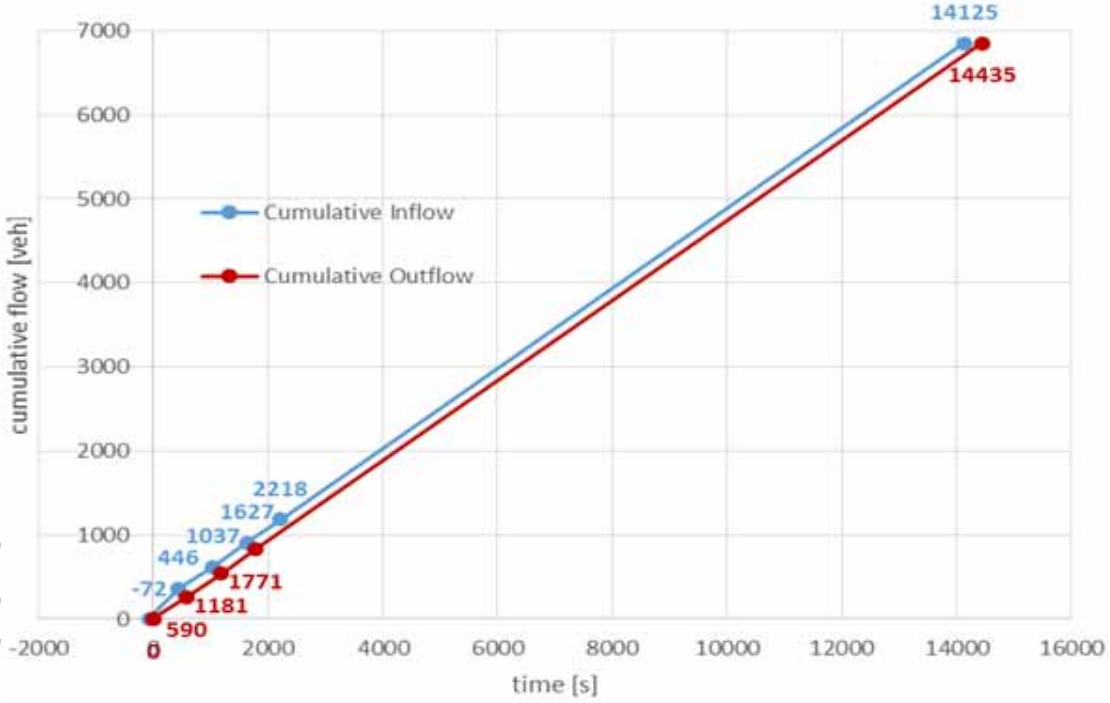


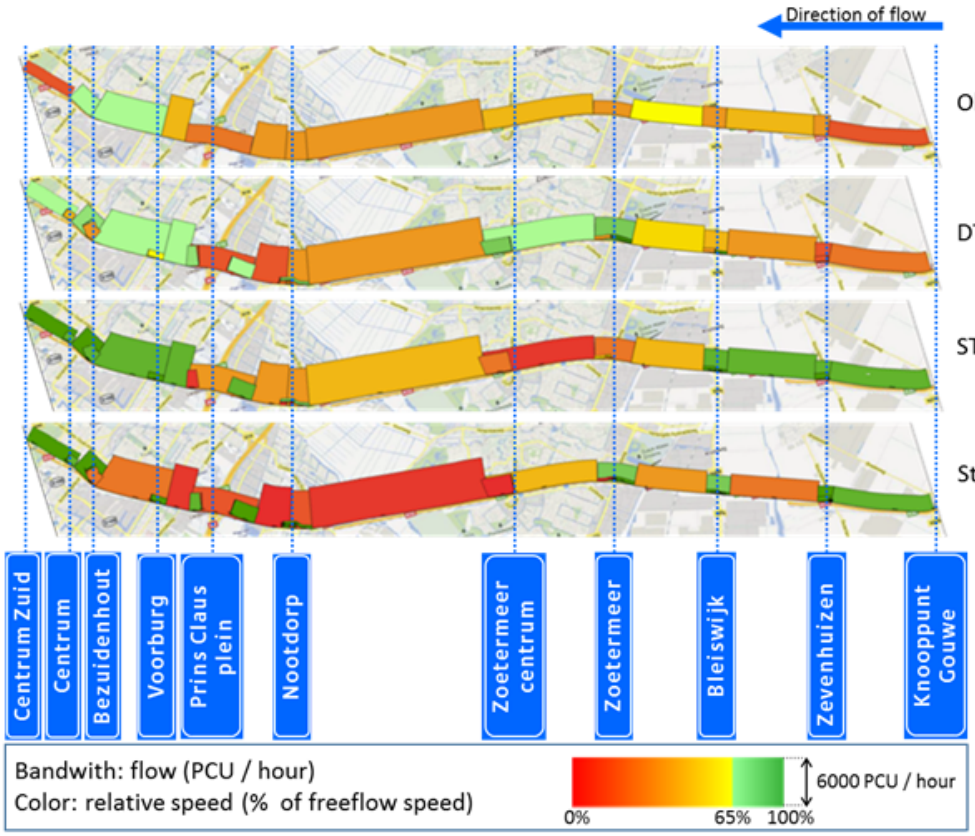


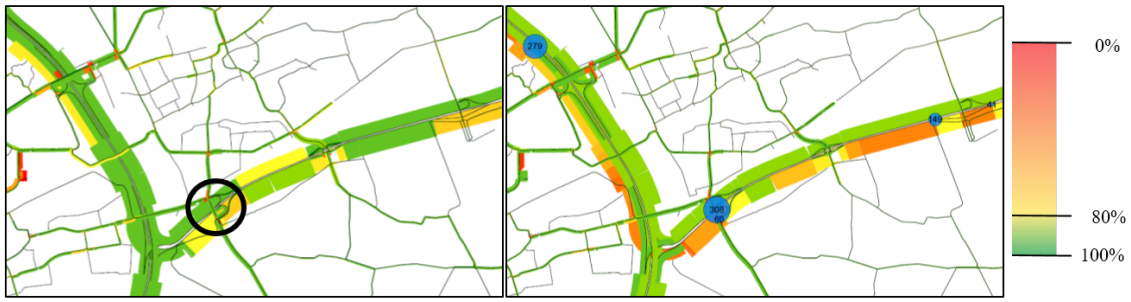
linknr	#lanes	road capacity
3	2	2000
4	4	8000
9	2	3000
10	2	5000
11	2	4000
12	1	2500
13	2	3000
14	1	2000

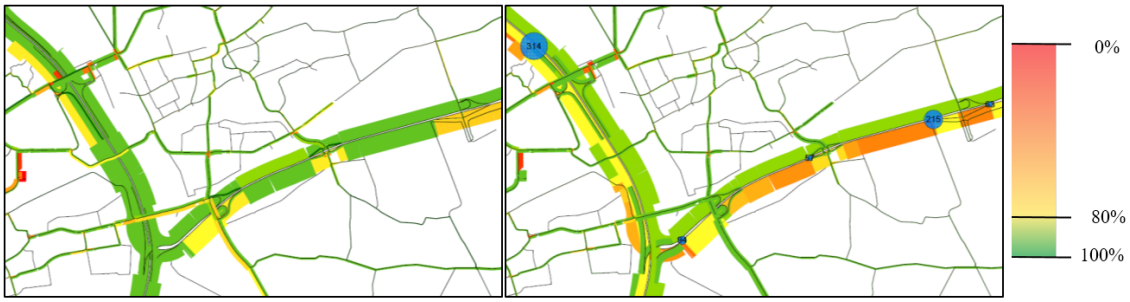
route	Links	cost (freeflow)
1	4,9,12,3	0.08
2	4,9,13,14,3	0.1
3	4,10,11,12,3	0.1
4	4,10,11,13,14,3	0.12

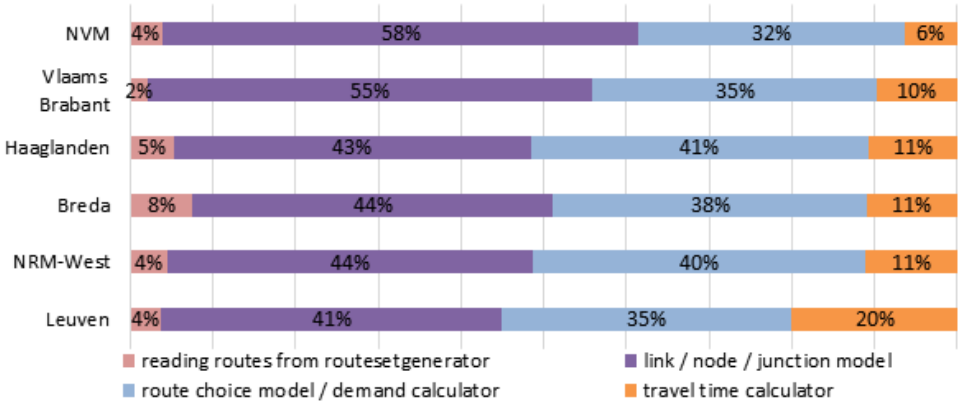


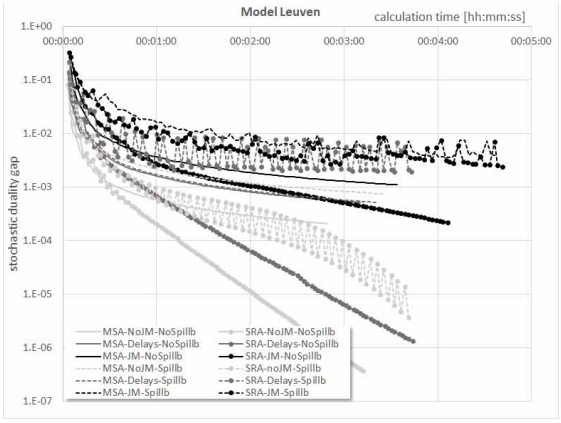


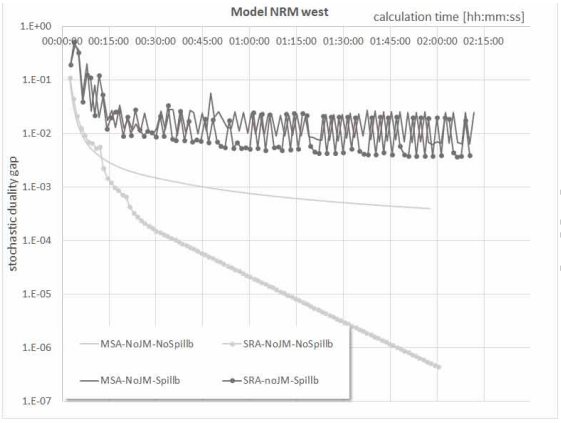


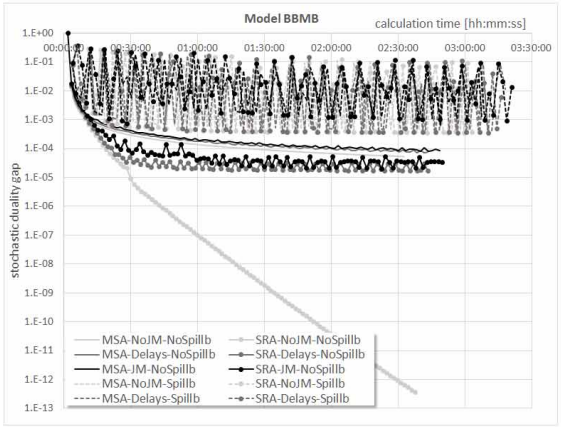


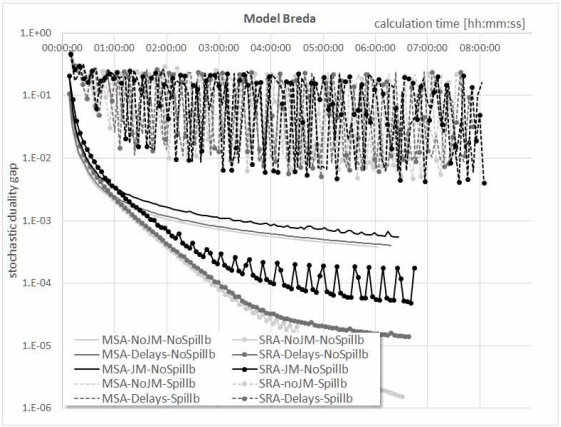


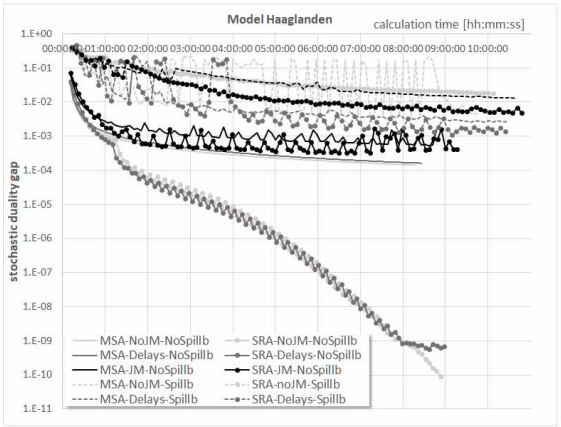


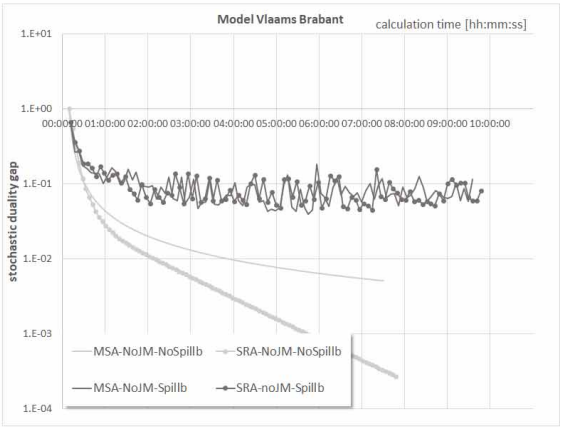

Model Leuven
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-Delays-NoSpillb
SRA-Delays-NoSpillb
MSA-JM-NoSpillb
SRA-JM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb
MSA-Delays-Spillb
SRA-Delays-Spillb
MSA-JM-Spillb
SRA-JM-Spillb

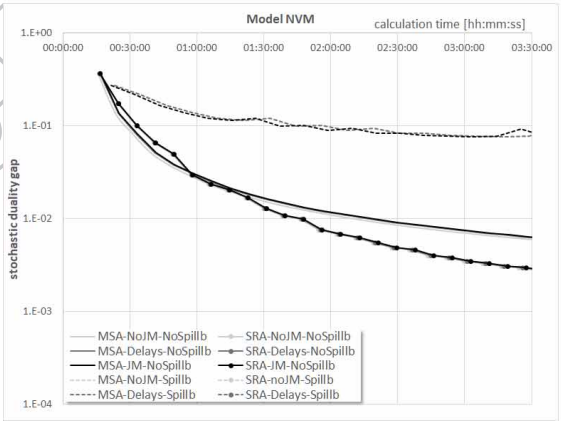

Model NRM west
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb


Model BBMB
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-Delays-NoSpillb
SRA-Delays-NoSpillb
MSA-JM-NoSpillb
SRA-JM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb
MSA-Delays-Spillb
SRA-Delays-Spillb


Model Breda
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-Delays-NoSpillb
SRA-Delays-NoSpillb
MSA-JM-NoSpillb
SRA-JM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb
MSA-Delays-Spillb
SRA-Delays-Spillb


Model Haaglanden
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-Delays-NoSpillb
SRA-Delays-NoSpillb
MSA-JM-NoSpillb
SRA-JM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb
MSA-Delays-Spillb
SRA-Delays-Spillb


Model Vlaams Brabant
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb


Model NVM
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-Delays-NoSpillb
SRA-Delays-NoSpillb
MSA-JM-NoSpillb
SRA-JM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb
MSA-Delays-Spillb
SRA-Delays-Spillb