Dynamic Origin-Destination estimation in congested networks: Theoretical findings and implications in practice

Rodric Frederix, Francesco Viti, Chris M.J. Tampère
Department of Mechanical Engineering, Katholieke Universiteit Leuven,
Celestijnenlaan 300A - PO Box 2422, 3001 Heverlee, Belgium

Abstract

In this study we analyze the impact of congestion in dynamic origin-destination (OD) estimation. This problem is typically expressed using a bi-level formulation. When solving this problem the relationship between OD flows and link flows is linearized. In the present paper the effect of using two types of linear relationship on the estimation process is analyzed. It is shown that one type of linearization implicitly assumes separability of the link flows, which can lead to biased results when dealing with congested networks. Advantages and disadvantages of adopting non-separable relationships are discussed. Another important source of error attributable to congestion dynamics is the presence of local minima in the objective function. It is illustrated that these local minima are the result of an incorrect interpretation of the information from the detectors. The theoretical findings are cast into a new methodology, which is successfully tested in a proof of concept.

Keywords: Dynamic origin-destination estimation; bi-level optimization; congestion spillback; global minimization

1. Introduction

The importance of dynamic traffic models is nowadays widely acknowledged. They form a central component in a variety of traveler information systems and traffic management systems (e.g. Lam et al. (2005), Papageorgiou (1995)), and they are indispensable to estimate and predict traffic flows in congested networks. Such models require several inputs, e.g. road capacities, route proportions, etc. A critical component among them is the dynamic origin-destination matrix, which captures the spatial and temporal distribution of traffic demand. Errors in this matrix are transferred onto the network flows, leading to errors in the network performance measures. For this reason estimating OD-flows correctly is of paramount importance.

The dynamic OD estimation problem has been tackled by many researchers in the past decades. Their studies focus on a variety of topics, dealing with the stochasticity and underdeterminedness of the problem, solution algorithms, the use of different data sources, and so forth. For a well detailed overview we refer to Lindveld (2003) and Balakrishna (2006). Despite of the progress that has been made in the last decades, there are still many challenges in dynamic OD estimation that remain to be solved, resulting in continuous research. Some recent work includes Kim et al. (2010) and Lee et al. (2011).

In this study we focus on the assignment-based\(^1\) dynamic OD estimation problem using traffic counts, which is a rather traditional approach to this problem. Key issue in the estimation of an OD matrix from traffic counts is the identification of the origin-destination

\(^1\) Corresponding author. E-mail: rodric.frederix@cib.kuleuven.be

\(^2\) It should be pointed out that although the paper deals primarily with these types of OD estimation procedures, the effect of route choice is dealt with marginally in this paper, whereas the effects of congestion dynamics, which have an equally important impact on the assignment matrix, are the main element of analysis.
flows that reproduce the observed traffic data. The assignment-based dynamic OD estimation problem is typically expressed as a bi-level problem:

\[
X^* = \arg \min_X \left[ z_1(X, \hat{X}) + z_2(Y(X), \hat{Y}) \right]
\]

subject to \( Y = A(X)X \)

\[
= B(X)P(X)X
\]

where in the upper level \( z_1 \) and \( z_2 \) are distance measures, \( X = [X_1, X_2, \ldots, X_N] \) denotes the unknown OD vector containing \( N = W \times T \) elements, \( W \) being the number of OD pairs and \( T \) being the number of (discretized) departure time periods. The generic element in \( X \), \( X_j \), is itself a vector of elements \( x_{ij}^j \) with \( j \) denoting any OD pair belonging to the set of OD pairs \( J \); \( X^* \) is the best estimate of the true OD vector given \( \hat{X} \), the target OD vector, and \( \hat{Y} = [Y_1, Y_2, \ldots, Y_K] \), being a vector with traffic counts containing \( M = S \times K \) elements, with \( S \) being the number of sensors and \( K \) being the number of time periods in which traffic counts are collected; \( Y_k \), is itself a vector of elements \( y_{ik} \) denoting the flow measured at link \( i \) during time period \( k \); finally \( Y \) is a vector with the corresponding estimated link flows, given \( X \).

The relationship between the OD flows and the link flows in the lower level is described by the assignment matrix \( A \), which has dimensions \( M \times N \). This matrix can be further subdivided into a crossing fraction matrix \( B \) with dimensions \( M \times R \) (\( R \) being the number of route flows), and a route fraction matrix \( P \), which has dimensions \( R \times N \) (see Cascetta (2001)). The elements of crossing fraction matrix \( B \) express the proportion of a route flow that passes a link, thus describing the spatio-temporal propagation of the route flows throughout the network. The elements of route fraction matrix \( P \) express the proportion of an OD flow choosing a certain route.

Conventional methods to solve problem (1) use an algorithm that iteratively assigns the OD matrix from the upper level into the lower level, and then solves the upper level problem using the assignment matrix from the lower level to approximate the non-linear relationship between the OD flows and the link flows. However, this approach can lead to biased solutions. Yang (1995) indicated that conventional methods for static OD estimation solve problem (1) as if it were a Cournot-Nash game, while in reality it is a Stackelberg game. This finding was later confirmed by Tavana (2001) for the dynamic OD estimation problem. They both argued the importance of including the response of the lower level when solving the upper level. Yang (1995) also showed that different types of response functions have different underlying assumptions. He compared two heuristics with a different response function. The first one is essentially the same as the one in conventional methods, i.e. it actually assumes that link cost functions are independent and thus fully separable. The second one can be used in cases where link cost functions cannot be assumed independent of each other. Also Maher et al. (2001) criticized conventional approaches. They made a distinction between formulating problem (1) as a bi-level program or as an equilibrium program. In a bi-level problem, the response of the lower level is taken into account in the upper level. Conventional approaches treat the two sub-problems in a parallel way and seek a mutually consistent solution by iteratively solving each sub-problem and transferring the latest solution to the other sub-problem. These approaches fall into another type of mathematical programming problem, namely, the equilibrium programming (EP) problem (Garcia and Zangwill, 1981). Maher et al. argued that in general both formulations lead to different solutions.
A first contribution of this paper is the extension of Yang’s analysis on the adoption of different types of response functions to dynamic OD estimation. In particular, we focus on the impact of delay, congestion formation and spillback on the response function. This analysis is described in Section 2.

A major difficulty in solving the bi-level problem (1) is the fact that this problem is inherently non-convex due to the non-linear dependency of the assignment matrix on the OD flows. Therefore it is possible that the estimated OD matrix is a local minimum and not a global minimum of problem (1). While many papers acknowledge this fact, the problem of converging to a local minimum has received little attention. However, in many cases these local minima result in a congestion pattern that differs strongly from reality. An example of this deviation can be seen in (figure 1). This figure was obtained by solving the OD estimation problem (1) on the ring way around Antwerp using only flow measurements. The speed measurements that were available were not used in the estimation process, but were used for validation. The left picture in (figure 1) is the real speed contour plot (note that the measured data was smoothed for purpose of visualization using the method described in Treiber & Helbing (2002)), while the middle picture is the speed contour plot that is obtained by assigning the estimated OD matrix. As one can see the congestion pattern of the estimated OD matrix deviates strongly from the actual congestion pattern, and the validity of the estimated matrix can thus be questioned. The estimation process was then repeated with both flow and density measurements in the goal function, following the methodology described in Frederix et al. (2010). The density measurements were derived by dividing flow measurements by speed measurements. The traffic pattern resulting from the estimated OD matrix can be seen in the right picture in (figure 1). The reason for adding these density measurements is discussed in section 3. This second estimate of the OD matrix also had a better match with the flow measurements than the first estimate (from the middle picture), which means that the first estimation process had converged to a local minimum.

As can be deduced from the assignment relationship in equation (1) there are two main sources of non-linearity leading to the non-convexity of problem (1): one because of the dependence of the crossing fraction matrix $B$ on the OD flows, the other one because of the dependence of the route fraction matrix $P$ on the OD flows. Note that in static OD estimation the crossing fraction matrix $B$ is equal to the link incidence matrix, and is not a function of the OD flows. Therefore this problem arises only in the dynamic OD estimation problem. The second contribution of this paper is the theoretical analysis on the non-linearity of the crossing fraction matrix, and its effect on the risk of converging to a local optimum. Although we acknowledge the importance of route choice and its effect on the occurrence of local optima, we only consider networks without route choice in this paper, to focus exclusively on the effect of the non-linearity of the crossing fraction matrix on OD estimation. The effect of route choice is equally relevant and will be the subject of future research.

Both the analysis on the different types of response function and the theoretical findings on the effect of the non-linearity of the crossing fraction matrix and the occurrence of local optima describe a set of requirements that need to be met for any dynamic OD estimation in congested networks, and thus form a methodology, which is illustrated in a proof of concept in Section 4. This constitutes the final contribution of the present paper.
The paper is organized as follows: the analysis on the underlying assumptions of different types of response functions is presented in section 2. In section 3, we study the non-convexity of the OD estimation problem due to congestion dynamics. The new methodology and its proof of concept are presented in section 4. In the last section we formulate our conclusions along with future research lines.

2. The impact of different response functions

We introduce the following notation:

\( y_i^k \) is the link flow at link \( i \) during time period \( k \)
\( x_j^t \) is the flow between OD pair \( j \) departing in time period \( t \)
\( X \) is a vector containing all OD flows \( x_j^t \)
\( a_{i,k}^{j,t} \) is the fraction of \( x_j^t \) that passes link \( i \) during time period \( k \)
\( J \) is the set containing all OD pairs
\( I \) is the set containing all links
\( J_i \) is the set of OD pairs passing link \( i \)
\( R_j \) is the set of routes between OD pair \( j \)

When solving the upper level problem in (1), a certain relationship between the link flows and OD flows has to be assumed, implicitly or explicitly. Since this relationship is not available analytically, an approximation of this relationship is necessary.

2.1 Linear approximation using separable response functions

Most dynamic OD estimation methods implicitly use a linear relationship between OD flows and link flows described by the assignment matrix. Three very popular examples are Generalized Least Squares (GLS) estimators (e.g. Cascetta & Postorino 2001, Yang et al. 2001), entropy-maximization models (e.g. Cheung et al. (2006)) and state-space models (e.g. Okutani 1987, Ashok & Ben-Akiva 2002). The linear relationship assumed in these approaches at point \( X_0 \) can be written down as follows:

\[
y_j^t = \sum_{i \in J} \sum_{k=1}^K a_{i,k}^{j,t}(X_0)x_j^t
\]  

(2)

The problem with this type of linear relationship is that it assumes that the link flows are separable, or, in other words, it assumes that the flow on a certain link \( i \) during a certain time interval \( k \) can only be changed by changing one of the OD flows that passes link \( i \) in time interval \( t \) when \( X_0 \) is assigned; this can be seen in equation (2) by noting that the assignment fraction \( a_{i,k}^{j,t} \) is zero for all OD flows that do not pass link \( i \) in time interval \( k \) when \( X_0 \) is assigned. This assumption of separability is incompatible with some typical phenomena in congested networks, such as congestion spillback between links, time lags due to the delay during congestion, and interdependencies between crossing (or opposing) flows encountered at intersections. In these cases it is very likely that increasing an OD flow will cause delays to other flows that do not pass that time-space interval, hereby altering the amount of flow passing the link in the considered time interval. This error has already been clearly addressed in past studies (e.g., Yang 1995, Tavana 2001, Lindveld 2003).
2.2 Linear approximation using non-separable response functions

A common manner to specify a linear relationship at a certain point $X_0$ is using a first-order Taylor expansion. We will now make a comparison between the traditional linear relationship (2) that is used in the optimization step of OD estimation methods, and the Taylor expansion approximation. The Taylor approximation of the link flows at point $X_0$ has the following form:

$$x_i^k = \sum_{j=1}^J \sum_{t=1}^T a_{i,j,k}^{r,t}(X_0) x_{o,j}^t + \sum_{j=1}^J \sum_{t=1}^T \frac{dy_i^t(X_0)}{dx_j^t}(x_j^t - x_{o,j}^t)$$

$$= \sum_{j=1}^J \sum_{t=1}^T a_{i,j,k}^{r,t}(X_0) x_{o,j}^t + \sum_{j=1}^J \sum_{t=1}^T \frac{\sum_{i=1}^I \sum_{r=1}^R a_{i,j,k}^{r,t}(X) x_{i,j}^t}{dx_j^t} \left(x_j^t - x_{o,j}^t\right)$$

$$= \sum_{j=1}^J \sum_{t=1}^T a_{i,j,k}^{r,t}(X_0) x_{o,j}^t + \sum_{j=1}^J \sum_{t=1}^T \left[\frac{\sum_{i=1}^I \sum_{r=1}^R a_{i,j,k}^{r,t}(X)}{dx_j^t} \left(x_j^t - x_{o,j}^t\right)\right]$$

The difference between equations (2) and (3) is the inclusion of the second term in equation (3). This term incorporates the sensitivity of the assignment matrix to changes in the OD flows, thereby allowing link flows to have non-separable behavior: the flow on a certain link $i$ during time interval $k$ can now also be changed by changing an OD flow that does not pass link $i$ when $X_0$ is assigned. This can be observed by focusing on equation (3), where the first term is summed over $J$, while the second term is summed over $J$. The sensitivity of the assignment matrix to changes of the OD flows consists of two terms itself, as can be seen in the following equation:

$$\frac{d \left(a_{i,j,k}^{r,t}(X)\right)}{dx_j^t} = \sum_{r \in R_j} \frac{\partial b_{i,j,k}^{r,t}(X_0)}{dx_j^t} \left(p_{i,j}^{r,t}(X_0)\right) + \sum_{r \in R_j} b_{i,j,k}^{r,t}(X_0) \frac{\partial p_{i,j}^{r,t}(X_0)}{dx_j^t}$$

The first term incorporates the dependence of the crossing fraction matrix $B$ on the OD flows keeping the route choice fractions fixed. This dependence is caused by travel time delay and congestion spillback. The second term accounts for the dependence of the route fraction matrix $P$ on the OD flows. The dependence is here caused by rerouting effects. As mentioned previously we focus on cases without route choice, so with $\frac{\partial p_{i,j}^{r,t}(X)}{dx_j^t} = 0$.

While using equation (3) instead of equation (2) might be theoretically more sound, the feasibility of calculating this second term can be questioned. This is an important cause for the widespread use of equation (2). For more complex traffic models that properly describe congestion dynamics no analytical expression is available for the relationship between the assignment fractions and the OD flows. As a result numerical methods are required to assess the derivative of this expression. It is possible to use finite differences to numerically approximate the sensitivity of the link flows to the OD flows, but if we have $N$ OD pairs and $T$ departure time intervals for these OD pairs, it would require performing a Dynamic Network Loading (DNL)/Dynamic Traffic Assignment (DTA) model $N \times T$ times. Since this is usually judged not feasible in terms of computation time, many researchers prefer the use of equation...
(2) to (3). To reduce the calculation effort and thereby allowing the response function to account for non-separable behavior some researchers have suggested the use of heuristic methods. Maher et al. (2001) propose an algorithm that solves the upper level problem with response function (2), and then use this OD matrix to define a search direction and to derive a linear approximation of the relationship between OD flows and link flows along this direction. A line search is then used to determine the step-size. Flötteröd & Bierlaire (2009) propose a linear regression model to approximate the response function. To reduce the dimensions of the model a demand transformation based on Principal Component Analysis is proposed. In Frederix et al. (2011) the method of Marginal Computation is used to derive the sensitivity of the link flows to the OD flows. This method that is described in Corthout et al. (2011) performs a perturbation analysis using kinematic wave theory principles. The difficulty to derive an exact calculation of equation (3) has also attracted attention towards the use of other gradient approximation methods, among which the Simultaneous Perturbation Stochastic Approximation (SPSA, Spall 1998) has been used extensively in dynamic OD estimation problems (e.g. Balakrishna & Koutsopoulos 2008, Cipriani et al. 2010) since it allows one to approximate a descent direction with significantly lower computational resources than through explicit calculation. This descent direction is only a rude approximation of the gradient. However, this descent direction does implicitly take into account the sensitivity of the assignment matrix to the OD flows. So although SPSA implicitly provides only an approximation of the response function (3), this approximation does not have a systematic bias, as opposed to the response function (2). A more thorough analysis of gradient approximation methods and their estimation accuracy is however out of the scope of this paper. Further details on this matter can be found in the above-cited papers.

To summarize, in this section the implicit assumptions behind different types of response function were discussed in detail. It can be concluded that the use of response function (3), whether implicitly or explicitly, is to be recommended when dealing with congested networks. Further research is however necessary to guarantee its applicability in large-scale networks or networks where we relax the assumption the absence of route choice.

3. Non-convexity due to congestion dynamics

To reduce the risk of local minima some papers stressed the importance of having a good initial matrix that does not deviate too much from the actual OD matrix (e.g. Tavana (2001)). However, there are two problems with this advice. Firstly, in many cases no good initial matrix is available. Secondly, it is unclear what kind of distance should be used for this deviation. A natural assumption would be to use Euclidean distance. However, the case study in section 4 will show that this is not always a good measure. In this section we therefore explore whether the condition of an initial matrix close to the actual OD matrix can further be refined.

3.1 Transitions between traffic regimes causing local optima

Looking back at Section 2, the response function (3) describes a local approximation of the relationship between the OD flows and the link flows. The local relationship between OD and link counts might be radically different far away or near the global optimum. Therefore, as long as we are far from the global solution, the local steepest descent direction might point us away from the global optimum rather than bringing us closer, causing convergence to a local minimum. Assuming a monotonic relation between any link flow and any OD flow (so the local relationships would never point in the wrong direction), and using a convex function as a distance function in (1) (e.g. squared deviations in GLS estimators), then the output would be a
unimodal function\(^3\). If we further would assume there were no measurement or modeling errors, the mode of each of these functions would be equal to the solution. Under these assumptions the sum of these functions would be unimodal as well, which means that any local minimum would be a global minimum. Note that we do not believe that the absence of measurement or modeling errors is a realistic assumption, but with this argumentation we try to understand the causes for local optima. The fact that we need the assumption of a perfect possible fit shows that also measurement and modeling errors are a potential source of local minima.

The above argumentation indicates that local relationships that suggest to adjust the OD flows in the wrong direction are a source of local optima. Now, one very important cause of such incorrect search directions is congestion spillback. While in uncongested conditions an increase of OD flow leads to an increase of link flow, spillback can cause this link flow to decrease by increasing the OD flow. To explain this phenomenon let us consider the following example: given a simple bottleneck network with a detector positioned on the upstream end of the link upstream of the bottleneck. The relationship between link flow and density is approximated in this example as the triangular fundamental diagram\(^4\) depicted in (figure 2). Assume that the measured flow at a certain time period is \(q''\) and that there is spillback on this link. This traffic state is shown in the triangular fundamental diagram as point \((q'', k'')\). However, suppose that while solving problem (1) this link is thought to be in free flow (e.g., it has reached point \((q, k)\)) with the current estimate of the demand. The linear relationship of the response function for this measurement will then suggest to decrease the OD flow to reach point \((q'', k')\). Vice versa if in reality the detector is in free flow regime (point \((q, k)\)), but this link is in congestion regime (point \((q'', k'')\)) with the current estimate of the demand, the linear relationship of the response function will suggest that the link flow is insensitive to any change of the OD flows, which of course only holds locally.

This example indicates that the relationship between link flows and OD flows in uncongested conditions is different from the relationship in congested conditions. In the first case the link flows were an increasing function of the passing OD flows. In the congested case the link flows became insensitive to the passing OD flows, and were determined by the bottleneck capacity.

As a consequence the relationship between link flows and OD flows behaves non-monotonic in the transition from free flow regime to congested regime. The sensitivities are different in both regimes, not only in magnitude, but also in sign, thereby leading to local optima. If we want to avoid this non-monotonic behavior, the necessity of this transition should also be avoided. Or in other words, the initial OD matrix should be chosen such that none of the link flows requires a transition from free flow regime to congestion regime (or vice versa). Thus the initial regime on all links needs to be the same as the actual regime. In the remainder of this paper we will refer to this condition as ‘starting from the correct regime’.

3.2 Correct traffic regime does not guarantee monotonicity

The non-convexity issue discussed in subsection 3.1 may seem to suggest that starting from (and staying in) the correct traffic regime leads locally to a monotonic relation between

---

\(^3\) A function \(f(x)\) is unimodal if for some value \(m\) (the mode), it is monotonically increasing for \(x \leq m\) and monotonically decreasing for \(x \geq m\).

\(^4\) Note that such relationship is chosen for illustrative purposes, but this argument would hold with any concave curve as fundamental diagram.
any link flow and any OD flow. However, the following example proves that this assumption does not always hold.

![Figure 3: Example of non-monotonic relation between OD flow and link flow](image)

Consider the network depicted in (figure 3). At the upstream node, suppose that the flow from the right branch (denoted by OD3) has priority over the flow from the left branch (the sum of OD1 and OD2). For this reason a queue starts building on the left branch. As a consequence the flow in this queue is equal to the remaining capacity (represented by point B in the bottom-right diagram). Flow A, which is shown also on the fundamental diagram in the bottom-right picture, is a mixture of OD1 and OD2. Now suppose the flow for OD1 decreases while the one for OD3 increases in such a way that the speed at which the queue spills back (the slope w1 in the bottom-right picture of (figure 3)) remains the same, in accordance with kinematic wave theory. The demand on the left branch then equals A’, and the remaining capacity equals B’. By doing so the speed in the queue is found to drop. This results in the smaller slope of the trajectories within the congested area in the x-t plot (the top-right picture in (figure 3)). Assume now that the flow identified by point A can be decomposed by a time period with only vehicles from OD1 alternated with a time period with only vehicles from OD2. In this scenario the change of the slope in the trajectories alters the composition of the flow passing the detector on the upper left branch during the indicated measurement interval. While originally it detected a flow equal to B (consisting of vehicles from only OD1), now it does not detect any flow (because no vehicles from OD2 are passing). So by decreasing OD1 (and increasing OD3) the link flow is decreased. If now we would further decrease OD1 (and increase OD3, such that the spillback speed remains the same), the slope of the trajectories further decreases, and the detector would measure some OD1 flow passing. Thus, by decreasing OD1, the link flow decreases at first, and afterwards increases again, clearly exhibiting non-monotonic behavior.

Although this scenario may be far-fetched, the example proves that having a correct traffic regime for every measurement point is not a sufficient condition for convergence to the global minimum. The example also illustrates that delays can lead to shifts in the composition of flow, which can cause non-monotonic behavior of the relationship between OD flows and link flows.

### 3.3 The impact of sensor coverage and positioning on unimodality

It should be noted that a non-monotonic relationship does not necessarily lead to local minima in the goal function if detectors are opportunely located in the network. A monotonic relationship of another detector could prevent the incorrect estimates described in the example in subsection 3.2. This is illustrated qualitatively in (figure 4). Consider for instance the flow on detector A to be a non-monotonic function of the OD flow. This relationship can be described generally by $y_A = f_A(x)$, and is illustrated by the black, continuous curve in the left picture of (figure 4). Suppose that this OD flow is also measured by a second detector B, and that the relationship between the OD flow and the flow on detector B is monotonic and described by $y_B = f_B(x)$ (the red, dashed line in the left picture of (figure 4)). Then the squared deviation (in the case of a GLS estimator) $(\hat{y}_A - y_A(x))^2$ in the goal function is not unimodal, but $(\hat{y}_B - y_B(x))^2$ is unimodal (right picture in (figure 4)). In the goal function we take the sum of both quadratic terms, so the sum of a non-unimodal and a unimodal function. As shown in the numerical example by the blue, marked curve representing the sum of squared deviations
from both detectors A and B, the goal function can become unimodal, even though non-
monotonic relationships are present.

Figure 4: Non-monotonic and monotonic relationship between OD flow and link flow (a)
and corresponding non-unimodal and unimodal relationship (b)

This example shows that non-monotonic behavior does not necessarily lead to local
optima; when the deviation from monotonicity in the relationship between the OD flow and the
flow on detector A is small, it is more likely that the goal function (1) is still unimodal. Now,
it can be expected that shifts in the composition of flow caused by delays will be quite small in
practice, and thus also the fluctuation of the link flow caused by the altered composition will
be rather small. The fluctuation of the link flow caused by spillback on the other hand can be
expected to have a larger magnitude. Therefore it seems reasonable to assume that local
minima caused by altered travel times will originate far less frequently than those caused by
congestion spillback.

In summary the position and coverage of sensors influences the occurrence of such local
optima in the goal function (1). To the authors knowledge this problem has not been studied at
present. State-of-the-art sensor positioning problems focus on finding the locations that best
cover the network, or that improve OD trip table estimation in a static environment
(e.g., Chootinan et al. (2005), Chen et al. (2007)). This is understandable, as positioning sensors
traditionally belongs to planning applications. While the influence of sensor positions on the
occurrence of the above described local optima can be an interesting research direction, we
will refrain from this in the remainder of this paper, as it lies outside its scope.

3.4 Identifying the correct traffic regime

A practical question that still arises is how to identify the correct regime. Using extra
information can be an effective way; since a value of traffic flow can correspond either to a
traffic state with low density and high speeds, or to a traffic state with high density and low
speeds. The use of speed or density measurements seems a natural choice. While the use of
speed data in OD estimation is not new (see e.g., Ashok 1996, Balakrishna et al. 2007),
previous studies mainly focus on how to include this extra information. The current paper
shows the theoretical reasons why speed data should be used. On the basis of this finding, the
amount of information that is necessary is actually much smaller: no accurate speed
measurements should be available, only a binary variable that explains whether a link is
congested or not. The extra information, whether speed measurements, density measurements
or binary variables, has to be translated into an OD matrix that produces a correct traffic
regime. Ashok (1996) uses the speed measurements as an empirical alternative to a theoretical
network loading model to derive the actual assignment matrix, which is then used to estimate
the OD matrix. This a viable approach, although accurate speed measurements have to be
available for the entire network. In Tavana (2001) the use of a hybrid model is suggested that
estimates OD flows based on both traffic flow and density. This approach has been tested
successfully in Frederix et al. (2010). Research is still ongoing on more efficient methods.

3.5 Summary

The previous subsections discussed the correlation between the occurrence of local
optima in the goal function and an initial OD matrix that produces link flows that are in a
different traffic regime than in reality. In the first subsection 3.1 it was shown that a monotonic
relationship between the link flows and OD flows is a sufficient condition for having a
unimodal goal function, and thus avoiding local optima. Vice versa a non-monotonic relationship can lead to convergence to a local optimum. The transition between a free flow regime and a congested regime is identified as a major cause for this non-monotonic relationship. To avoid this transition it is necessary to make sure that all links are already in the correct regime when the initial OD matrix is assigned. However, avoiding this transition in the presented manner does not guarantee a monotonic relationship between the link flows and OD flows, as was illustrated with an example in subsection 3.2. Nevertheless a monotonic relationship is not a necessary condition for a unimodal goal function, as was qualitatively shown in subsection 3.3. It is argued that, in theory, convergence to a local minimum is still possible, but in practice this is not expected to occur frequently. It should also be pointed out that while identifying a correct traffic regime for every link is not sufficient to guarantee a global optimum, it is clear that the risk of local minima will be far greater when this condition is not met. Finally in subsection 3.4 we focus on methods to adjust the initial OD matrix such that it produces the correct regime on all links.

4. Methodology and proof of concept

The essence of the theoretical findings elaborated in section 2 and 3 can be summarized in the following methodology. First of all we propose the use of response function (3) to approximate the relationship between OD flows and link flows in the upper level of (1), rather than using the more conventional response function (2). Secondly we suggest to use an initial OD matrix that produces the same congestion pattern as is observed in reality, i.e. that allows one to start with the correct traffic regimes. The fulfillment of both requirements might seem difficult, but existing methods both for the calculation of the gradient and for producing an OD matrix with correct traffic regimes are already available and were presented in the previous sections. Nevertheless, this paper has suggested that there is still need for further research to refine these methods and enhance their computational performance.

The proposed methodology is applied to a synthetic network as a proof of concept. The traffic model that is used to describe the local relationship between the OD flows and the link flows is the Link Transmission Model (Yperman et al. 2006). The sensitivity of the link flows to the OD flows is calculated through numerical approximation.

The synthetic network is depicted in (figure 5(a)). It is a simple merge network. Each branch of the merge is subdivided in 4 links. There are 2 OD pairs, and the real OD flows as a function of time are depicted in (figure 5(b)).

(a)       (b)

Figure 5: Synthetic network (a) and its OD flows (b)

When the OD flows increase after 15 minutes, a bottleneck activates at the merging point, and a queue spills back onto the left branch. Note that there is no queue on the right branch, because we use a capacity proportional node model (see Tampère et al. 2011), and the sending flow of the right branch is smaller than its capacity proportional share. The maximum length of the queue is indicated in (figure 5(a)). When both OD flows decrease again after 95 minutes, the queue starts to decrease, but also the flow in the queue increases because of reduced “competition” of OD2 for the bottleneck capacity. For the OD estimation synthetic measurements are generated for all nodes indicated with an ‘x’. No detectors are assumed available on the right branch. However the flow on this branch can easily be deduced using flow conservation equations. We use a GLS goal function, and as a stop criterion we check whether the maximum change in OD flow over all OD flows between two iterations is smaller than 1 vehicle/hour.
The effect of starting in a different regime and the effect of a different linear model between the OD flows and the link flows is shown. The different linear models that are used are a simplified linear model that only uses the assignment matrix for expressing the relationship (equation (2)), and a more correct linear model that also takes the sensitivity of the assignment matrix into account (equation (3)). A gradient-based optimization method is used: the gradient is calculated at the current point, and we take a step in this direction. Next, the gradient at this new point is calculated, and this process repeats until convergence. Four different OD estimations are performed:

a) Incorrect starting regime, simplified linear model
b) Incorrect starting regime, correct linear model
c) Correct starting regime, simplified linear model
d) Correct starting regime, correct linear model

The initial OD matrix can be seen in (figure 6). The real OD matrix is also depicted in dashed lines as a point of reference. The left OD matrix (a) results in a free flow regime for all links during the whole simulation period, while the right OD matrix (b) produces congestion at the correct place at the correct moment in time, though with incorrect demand values.

![Figure 6: Start OD matrix resulting in a) incorrect regime and b) correct regime](image)

The estimated OD matrices are summarized in (figure 7), and numerical results are provided in (table 1). In all but the last case the OD estimation process converges to an incorrect solution. Note that in case d) the exact solution is not fully reached because the problem is underdetermined for OD flows from 2 to 3 in the time period between the start of the congestion spillback and the moment that this queue spills back over the first detector on the left branch. Apart from this local effect the estimation is quite accurate.

![Figure 7: Estimated OD matrices for case (a), (b), (c) and (d)](image)

**Table 1: Root Mean Square Error in (veh/h) of initial and estimated OD flows**

These results suggest that both conditions, namely a correct starting regime and a correct linear model for the response function, need to be fulfilled to obtain an unbiased OD estimation method. As was mentioned in section 3 some researchers stress the importance of having an initial matrix that does not deviate too much from the actual OD matrix, because otherwise the estimation process might converge to a local optimum. This example illustrates that this condition can be misleading: the average deviation from the actual OD matrix in case a) and b) is 200 veh/h for OD 1-3 and 161.1 veh/h for OD 2-3, while for case c) and d) the average deviation is 584.2 veh/h and 637.8 veh/h respectively. This example therefore illustrates that the initial matrix should yield correct regimes, rather than being “close” in terms of OD flow volumes.

**5. Conclusions**

The present paper deals with the effect of congestion on dynamic OD estimation and focuses on different aspects related to congestion dynamics. First, the analysis of Yang on the impact of different response functions on static OD estimation is extended to the dynamic case, with a
focus on the effect of congestion formation and spillback on these functions. It is shown that conventional methods use solution methods that implicitly assume that link flows are separable, an assumption that is not valid in congested networks. This suggests a first important condition for unbiased OD estimation in congested networks, namely the use of a proper response function. Second, the occurrence of local optima due to congestion spillback is identified as another important source of error. We show that the cause for these local optima is the incorrect interpretation of the measurements due to a different traffic regime on a link assumed in the model as compared to reality. While having a correct traffic regime cannot guarantee convergence to the global minimum, as is illustrated in an example, we argue that the condition of correct traffic regimes will probably suffice in many practical cases, and that not fulfilling this condition will increase the risk of local minima. The use of both conditions is proposed as a new methodology, which is tested in a proof of concept. The experiment indicates that both requirements have to be met for a good OD estimation. They also show that the conventional ‘strategy’ for avoiding local optima, namely having a start OD matrix that is close to the real OD matrix, can be misleading. Because the fulfillment of both requirements is still challenging with currently available methods, there is still need for further research to refine these methods to meet these requirement and still guarantee a feasible computational performance.

References


Table 1: Root Mean Square Error in (veh/h) of initial and estimated OD flows

<table>
<thead>
<tr>
<th>Case</th>
<th>Initial OD 1-3</th>
<th>Initial OD 2-3</th>
<th>Estimated OD 1-3</th>
<th>Estimated OD 2-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case a</td>
<td>200</td>
<td>161.1</td>
<td>100.1</td>
<td>171.7</td>
</tr>
<tr>
<td>Case b</td>
<td>200</td>
<td>161.1</td>
<td>100.2</td>
<td>151.9</td>
</tr>
<tr>
<td>Case c</td>
<td>584.2</td>
<td>637.8</td>
<td>100.1</td>
<td>151.9</td>
</tr>
<tr>
<td>Case d</td>
<td>584.2</td>
<td>637.8</td>
<td>2.0</td>
<td>13.7</td>
</tr>
</tbody>
</table>
Figure 1: (a) Real traffic pattern, (b) traffic pattern with first estimate of OD matrix, and (c) traffic pattern with second estimate of OD matrix

Figure 2: Incorrect traffic state induces errors

Figure 3: Example of non-monotonic relation between OD flow and link flow

Figure 4: Non-monotonic and monotonic relationship between OD flow and link flow (a) and corresponding non-unimodal and unimodal relationship (b)

Figure 5: Synthetic network (a) and its OD flows (b)

Figure 6: Start OD matrix resulting in a) incorrect regime and b) correct regime

Figure 7: Estimated OD matrices for case (a), (b), (c) and (d)