Transport Analytics approaches to the Dynamic Origin-Destination Estimation Problem

Xavier Ros-Roca
PTV Group, Karlsruhe, Germany; Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain

Lidia Montero
Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain

Jaume Barceló
PTV Group, Karlsruhe, Germany

ABSTRACT

Dynamic traffic models require dynamic inputs, and one of the main inputs are the Dynamic Origin-Destinations (OD) matrices describing the variability over time of the trip patterns across the network. The Dynamic OD Matrix Estimation (DODME) is a hard problem since no direct full observations are available, and therefore one should resort to indirect estimation approaches. Among the most efficient approaches, the one that formulates the problem in terms of a bilevel optimization problem has been widely used. This formulation solves at the upper level a nonlinear optimization that minimizes some distance measures between observed and estimated link flow counts at certain counting stations located in a subset of links in the network, and at the lower level a traffic assignment that estimates these link flow counts assigning the current estimated matrix. The variants of this formulation differ in the analytical approaches that estimate the link flows in terms of the assignment and their time dependencies. Since these estimations are based on a traffic assignment at the lower level, these analytical approaches, although numerically efficient, imply a high computational cost. The advent of ICT applications has made available new sets of traffic related measurements enabling new approaches; under certain conditions, the data collected on used paths could be interpreted as an empirical assignment observed de facto. This allows extracting empirically the same information provided by an assignment that is used in the analytical approaches. This research report explores how to extract such information from the recorded data, proposes a new optimization model to solve the DODME problem, and computational results on its performance.

1. INTRODUCTION: ANALYTICAL APPROACHES TO DODME

Trip patterns in terms of Origin to Destination (OD) traffic flows are a key input to traffic assignment models, namely to Dynamic Traffic Assignment models, where they also must be dynamic, or at least time discretized, to properly approximate the time variability of the demand. OD matrices are not yet observable; in the best case, the measurements from Information and Communication Technologies (ICT), as GPS vehicle tracking, or mobile phones Call Detail Records (CDR), allow drawing samples that must be suitably expanded to provide estimates of the whole population. Therefore, their estimation must be done resorting to indirect process, usually based on mathematical models.

One of the most appealing mathematical formulations of the OD estimation problem is in terms of bilevel optimization problems (1), aimed at adjusting an initial target OD, $\boldsymbol{X^H}$, so that it could explain the observed link flow counts $\widehat{\boldsymbol{Y}}$ at counting stations in the network, Ros-Roca et al. (2018).

$$min\, Z(\boldsymbol{X},\boldsymbol{Y}) = w_1 F_1(\boldsymbol{X},\boldsymbol{X^H}) + w_2 F_2\left(\boldsymbol{Y},\widehat{\boldsymbol{Y}}\right) \quad (1)$$

$$s.to\ \ \boldsymbol{Y} = Assignment(\boldsymbol{X})$$

$$\boldsymbol{X} \geq 0$$

where F_1 and F_2 are suitable distance functions between estimated and observed values; while w_1 and w_2 are weighting factors reflecting the uncertainty of the information contained in $\boldsymbol{X^H}$ and $\widehat{\boldsymbol{Y}}$, respectively. The underlying hypothesis is that $\boldsymbol{Y}(\boldsymbol{X})$ are the link flows predicted by assigning the demand matrix $\boldsymbol{X}$ onto the network, which can be expressed by a proportion of the OD demand flows passing through the count location at a certain link. In terms of the assignment matrix $\boldsymbol{A}(\boldsymbol{X})$, which is the proportion of OD flow that contributes to a certain link traffic count, is:

$$\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{X})\boldsymbol{X} \quad (2)$$

Then the resulting bilevel optimization problem solves (at the upper level) the nonlinear optimization problem by substituting the estimated flows $\boldsymbol{Y}$ in the objective function of (1) with the relationship (2):

$$min\, Z(\boldsymbol{X},\boldsymbol{Y}) = w_1 F_1(\boldsymbol{X},\boldsymbol{X^H}) + w_2 F_2\left(\boldsymbol{A}(\boldsymbol{X})\boldsymbol{X},\widehat{\boldsymbol{Y}}\right) \quad (3)$$

$$s.to\ \ \boldsymbol{X} \geq 0$$

To estimate a new OD matrix $\boldsymbol{X}$, while at the lower level, a static user equilibrium assignment is used to solve the assignment problem $\boldsymbol{Y} = Assignment(\boldsymbol{X})$ in order to estimate the assignment matrix $\boldsymbol{A}(\boldsymbol{X})$ induced by the new $\boldsymbol{X}$. Spiess (1990) is a good example of a seminal model based on this approach.

This mathematical model is highly undetermined since the number of variables, OD pairs, is much larger than the number of equations. Link flow counts available at a subset of links in the network, along with the distance functions in the objective function, are the additional information aimed at reducing the degree of indetermination, in the most frequent implementations, using a static traffic assignment to solve the lower level problem, Spiess (1990), Codina and Montero (2006), Lundgren and Peterson (2008). The main reason for these implementations is that they are algorithmically efficient and present nice properties for convergence and stability. However, since static assignment models support them, they cannot properly account for the impacts of traffic dynamics and the induced congestions.

2. EXTENSIONS OF ANALYTICAL FORMULATIONS TO ACCOUNT FOR TIME DEPENDENCIES

Assuming that the functional dependency between the estimated flows $\boldsymbol{Y}$, the assignment

matrix $\boldsymbol{A}(\boldsymbol{X})$ and the estimated matrix $\boldsymbol{X}$, set up in (2), allows a Taylor expansion around the current solution which provides a more detailed insight of the how the path flows contribute to the link flows, which is in essence the information provided by the assignment matrix. This improved approach was explored in Lundgren and Peterson (2008) still using a static assignment. Other researchers, (Frederix et al. 2013; Toledo and Kolechkina 2013; Yang et al. 2017) proposed to use a Dynamic Traffic Assignment at the lower level to account for time dependencies, and therefore for congestion building processes. That allows a richer Taylor expansion also in terms of time, which captures these phenomena. To properly reformulate (3), let's assume that:

- I is the set of Origins, J the set of Destinations and $N \coloneqq I \times J$ the set of OD pairs.
- $\mathcal{T} = \{1, \dots, T\}$ is the set of time intervals.
- L is the set of links in the network. $\hat{L} \subseteq L$ is the subset of links that have sensors.
- $\hat{y}_{lt}$ are the measured flow counts at link l during time period t. y_{lt} are the corresponding simulated flow counts, $\forall l \in \hat{L} \subseteq L$ and $\forall t \in \mathcal{T}$. $\boldsymbol{Y} = (y_{lt})$ and $\hat{\boldsymbol{Y}} = (\hat{y}_{lt})$ are link flow counts in vector form.
- x_{ijr} are the OD flows for (i,j)-th OD pairs departing during time period r, $\forall i \in I, \forall j \in J$ and $\forall r \in \mathcal{T}$. $\boldsymbol{X} = (x_{ijr})$ are the OD flows in vector form.
- a_{ijr}^{lt} is the flow proportion of the (i,j)-OD pair departing at time period $r \in \mathcal{T}$ and captured by link $l \in \hat{L}$ at time period $t \in \mathcal{T}$. $\boldsymbol{A} = [a_{ijr}^{lt}]$ is the assignment matrix.

Then the DODME problem can be reformulated in the following terms: Given a network with a set of links L, a set $N \coloneqq I \times J$ of OD pairs, and the set of time periods $\mathcal{T}$. The goal of the dynamic OD-matrix estimation problem is to find a feasible vector (OD-matrix) $\boldsymbol{X}^* \in G \subseteq \mathbb{R}_+^{N \times \mathcal{T}}$, where $\boldsymbol{X}^* = (x_{ijr}^*),\ (i,j) \in N, r \in \mathcal{T}$, consists of the demands for all OD pairs. It can be assumed that, when assigning the time-sliced OD matrices onto the links of the network, it should be done according to an assignment proportion matrix $\boldsymbol{A} = [a_{ijr}^{lt}], \forall l \in L, \forall (i,j) \in N, \forall r,t \in \mathcal{T}$, where each element in the matrix is defined as the proportion of the OD demand x_{ijr} that uses link l at time period t. The notation in (2) is used to indicate that, in general, these proportions depend on the demand. The linear relationship between the flow count on a link and the given OD pair is in matrix form, which thus sets the vector of detected flows as $\boldsymbol{Y} = (\boldsymbol{Y_1}, \dots, \boldsymbol{Y_T}) = (y_{e_1 1}, \dots, y_{e_L 1}, \dots, y_{e_1 T}, \dots y_{e_L T})$ and the vector of OD flows as $\boldsymbol{X} = (\boldsymbol{X_1}, \dots, \boldsymbol{X_T}) = (x_{i_1 j_1 1}, \dots, x_{i_I j_1 1}, \dots, x_{i_1 j_1 T}, \dots, x_{i_I j_T})$. Then, the relationship (2) can be reformulated in vector form as:

$$y_{lt} = \sum_{(i,j) \in N} \sum_{r=1}^{t} a_{ijr}^{lt} x_{ijr} \tag{4}$$

where the a_{ijr}^{lt} entries of matrix A represents the proportion of OD flow departing at time r, x_{ijr}, passing through link l at time t, y_{lt}.

This linear mapping between the link flows and the OD flows is indeed the first term in the Taylor expansion of the relationship between link flows and OD flows, at $\widetilde{\boldsymbol{X}}$ in the neighbourhood of $\boldsymbol{X}$, the additional terms capture the assignment matrix's sensitivity to changes in the OD flows, path choice and congestion propagation effects (Frederix et al. 2011, 2013; Toledo and Kolechkina 2013).

Details of this approach can be found in Ros-Roca et al. (2019), where the Spiess' approach for the dynamic case by using the first term in the Taylor expansion is defined. It does not account for the propagation effects, but it explicitly considers the time dependencies. The resulting minimization problem is as follows:

$$\min Z(\boldsymbol{X}) = \frac{1}{2}\sum_{t\in\mathcal{T}}\sum_{l\in\hat{L}}\left(\left(\sum_{(i,j)\in N}\sum_{r=1}^{t} a_{ijr}^{lt}x_{ijr}\right)-\hat{y}_{ijt}\right)^2 \qquad (5)$$
$$s.to \quad \boldsymbol{A} = Assignment(\boldsymbol{X})$$
$$x_{ijr} \geq 0$$

where $\boldsymbol{A} = \left[a_{ijr}^{lt}\right]$ is the assignment matrix. Moreover, the Spiess like formulation can be improved adding a second term in the objective function, as in (1), in order to compare the estimated matrix to a historical OD matrix. Assuming a quadratic function to measure the distances between the estimated and the historical or target matrix:

$$\min Z = \frac{1}{2}\sum_{t\in\mathcal{T}}\sum_{l\in\hat{L}}\left(\left(\sum_{(i,j)\in N}\sum_{r=1}^{t} a_{ijr}^{lt}x_{ijr}\right)-\hat{y}_{lt}\right)^2 + \frac{w}{2}\sum_{r\in\mathcal{T}}\sum_{(i,j)\in N}\left(x_{ijr}-x_{ijr}^{H}\right)^2 \qquad (6)$$

And the iterative Dynamic Spiess procedure would be as follows:

$$X_i^{(k+1)} = \begin{cases} X_i^H & \text{for } k = 0 \\ X_i^{(k)}\left(1-\lambda^{(k)}\left[\frac{\partial Z(X)}{\partial X_i}\right]_{X_i^{(k)}}\right) & \text{for } k > 0 \end{cases} \qquad (7)$$

3. A NEW APPROACH – ASSIGNMENT FREE DODME

The analytical approaches to DODME problem discussed so far show that all are based on the availability of the Assignment Matrix $\boldsymbol{A}$ and, in the case of the dynamic extensions, of its expansion a_{ijr}^{lt} for the various time intervals, and the travel times from the origin of the trip to the corresponding link *l*. And the main role of the Dynamic Traffic Assignment at the lower level of (1) is just to provide this estimate at each time interval.

The availability of the GPS tracking data enables us to assume that after a suitable data processing the empirical paths and the path choices can be interpreted in terms of an empirical dynamic assignment. Then if an empirical assignment matrix can also be estimated then it would play a similar role to that of the analytical assignment matrix estimated from the Dynamic Traffic Assignment.

Therefore, the research question addressed in the following sections is:

- Assuming that suitable GPS tacking data (e.g. waypoints) are available for a given period of time, and
- An *ad hoc* data processing generates an empirical assignment matrix of enough quality, and

- Additional traffic information is also available (e.g. link flow counts at a subset of links in the network)

Then, to investigate whether it is possible to use such information to find a new formulation of the DODME problem, in terms of an optimization model, not requiring the execution of any Traffic Assignment procedure.

3.1 From GPS Data to Link Travel Times

Data collected by GPS devices equipping fleets of vehicles when tracking their trajectories across the network. They are usually available in the format of datasets, as the one in Table 1, of trips detailed by an ordered sequence of waypoints, $(ID_k, ts_{kl}, lat_{kl}, long_{kl})$, for each trip k with a Trip Identity ID_k, the recording date, the recording time tag ts_{kl} when the *l-th* observation of trip k was recorded, and the latitude and longitude of the current position of the vehicle when the data were recorded.

ID	DATE	TIME TAG	LATITUDE	LONGITUDE
4261353	2015-11-30	22:43:58	45.445988	9.124048
4261353	2015-11-30	22:44:57	45.445496	9.121952
……..	……	……	……..	……
4261353	2015-11-30	22:50:57	45.444767	9.119217
4261355	2015-11-30	22:43:58	45.44598	9.124048
4261355	2015-11-30	22:44:57	45.445496	9.121952
……	……	……	……..	……..
4262355	2015-11-30	22:50:57	45.444767	9.119217
……	……	……	……	……

Table 1 – GPS Waypoints Data Sample

These data should be map matched onto the map of the scenario being analyzed, the conceptual approach to the procedure developed in this work is depicted in Figure 1.

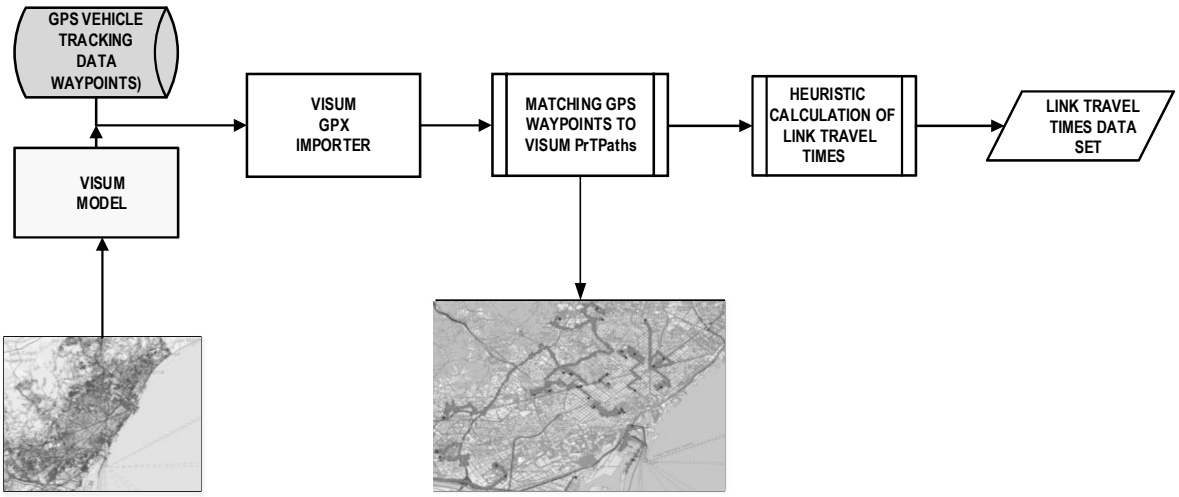
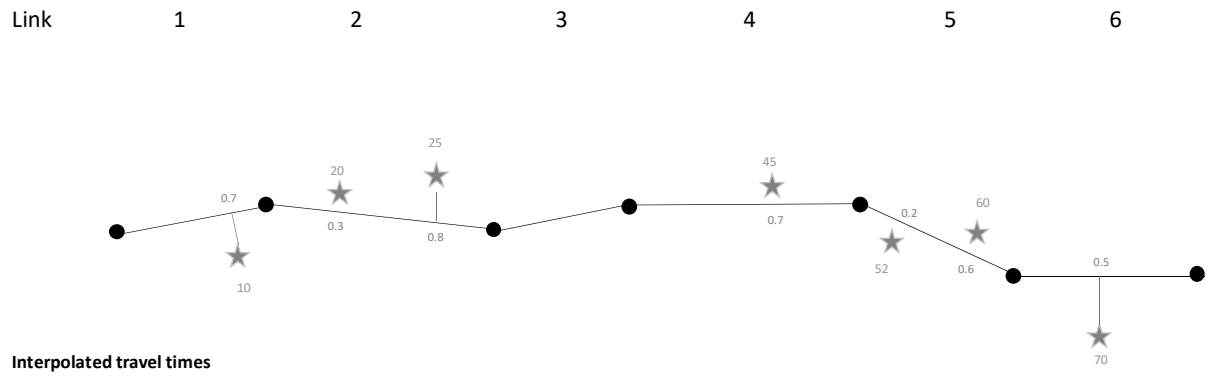


Figure 1 – Conceptual methodological approach to the process of importing waypoints into a Visum model and their use to calculate link travel times

The map matching process transforms sequences of waypoints to paths in VISUM. This is performed by the Map Matching tool of the software. Firstly, it assigns each waypoint to a certain point on the nearest link of the network. After that, the travel times are extrapolated in order to arrive to each waypoint according to the sequence. Figure 2 shows how it works with an example.



Interpolated travel times

tt_1 = tt_6 = not covered completely > Undefined > not part of PrT Path

$$tt_2 = \frac{0.3l_2}{0.3l_1+0.3l_2}10\text{s} + 5\text{s} + \frac{0.2l_2}{0.2l_2+l_3+0.7l_4}20\text{s}$$

$$tt_3 = \frac{l_3}{0.2l_2+l_3+0.7l_4}20\text{s}$$

$$tt_4 = \frac{0.7l_4}{0.2l_2+l_3+0.7l_4}20\text{s} + \frac{0.3l_4}{0.3l_4+0.2l_5}7\text{s} \quad \text{etc.}$$

The interpolated travel time for Link i is stored in the PrT Path Element for the FromNode of Link i.

Figure 2 – Example of the interpolation of travel times according to the waypoints' sequence

Finally, once all the waypoints sequences are converted to several paths, with full details at link level, the travel times by link are averaged. The outcome of this process is the set of observed link travel times at each time period t: $\hat{tt}_{lt}, \forall l \in L, \forall t \in T$ for all links in the network that are used by the GPS traces. These observed link travel times are calculated by averaging all the occurrences for each link at different GPS paths. This is the Data Set of Observed Link Travel Times.

3.2 Route Choice Set Generation

The Route Choice Set is the set of different paths sets, $\mathcal{K} = \left\{K_{ijr}\ , \forall i \in I\ ,\ \ \forall j \in J\ ,\ r \in \{1, \dots, T\}\right\}$, that contain all paths that can be used by all the vehicles, depending on where they come from, where they go to and when they are departing. In simulation, the generation of these sets is usually done by the DTA engine and the volumes are assigned following the equilibrium criteria, which is, indeed, the assignment procedure.

In this case, the set is calculated by using only the first phase of a Stochastic Assignment engine. The Stochastic Assignment in VISUM has two different phases:

- The first phase is called the Path Search and it generates a maximum number (N_{max}) of paths for each OD pair. It generates these paths by calculating several times the

shortest path by perturbing, with a normal distribution perturbation, the links' impedances of the network.

- The second phase assigns the volumes of each OD pair to the different paths, but it is set to 0 iterations, which means that there is no assignment

In this experiment, the Observed Link Travel Times, Section 3.1, for each link of the network l at each time period r, $\widehat{tt}_{lr}$, are used as impedances. When there is no observed link travel time, because the link is not used by the GPS sample, a scaled travel time is used:

$$\widehat{tt}_{l't} = r \cdot tt_{0_{l'}} \quad , \quad r = mean_{l \in GPS}\left(\frac{\widehat{tt}_{lt}}{tt_{0_l}}\right) \tag{8}$$

where tt_{0_l} is the travel time at free flow at each link and r is computed using all observed link travel times and their corresponding travel time at free flow.

3.3. Calculation of CF_k and P_k

The paths in K_{ijr} are noted by $k(i,j,r) \in K_{ijr}$, to show explicitely the dependence to (i,j,r). For a certain path $k(i,j,r)$, the sequence of links that compound it is the set $\Gamma_{k(i,j,r)} = \{e_1, \dots, e_{m_k}\}$.

The path choice for each path on the set K_{ijr} is calculated as a discrete choice model that uses the commonality factor, CF_k, as a penalization factor on travel times, Bovy et al. (2008). That is:

$$CF_{k(i,j,r)} = \frac{1}{\mu_{CF}} \sum_{a \in \Gamma_{k(i,j,r)}} \left(\frac{l_a}{L_{k(i,j,r)}} \log\left(\sum_{h \in K_{ijr}} (\delta_{ahr} + 1) \right) \right)$$

$$P_{k(i,j,r)} = \frac{\exp\left[\mu_{P_K}\left(-\widehat{tt}_{k(i,j,r)} - CF_{k(i,j,r)}\right)\right]}{\sum_{h \in K_{ijr}} \exp\left[\mu_{P_K}\left(-\widehat{tt}_{h(i,j,r)} - CF_{h(i,j,r)}\right)\right]} \tag{9}$$

These calculations permit to obtain the flow distribution for each path, based on observed path travel times. These observed path travel times are the summation of the observed link travel times, considering the arrival time, $t(k)$, at each link a, included in the path $k(i,j,r)$:

$$\widehat{tt}_{k(i,j,r)} = \sum_{a \in \Gamma_{k(i,j,r)}} \widehat{tt}_{at(k)} \tag{10}$$

Regarding parameters in (9), we have used $\mu_{CF} = 1$, as in Bovy et al. (2008) , and for μ_{P_K} a coefficient that depends on the set K_{ijr} and it is calculated as follows:

$$\mu_{P_K} = \frac{1}{mean_{k \in K_{ijr}}\left(\widehat{tt}_{k(i,j,r)}\right)} \tag{11}$$

3.4. Calculation of the Time Dependent Assignment Matrix

Once $\boldsymbol{P_k} = [P_{k(i,j,r)}]$ is calculated from the k-shortest paths calculated after the travel times estimated from the GPS data for all OD pairs, the empirical assignment matrix $\overline{\boldsymbol{A}} = \left[\overline{a}_{ijr}^{lt}\right]$ can be calculated:

$$\overline{a}_{ijr}^{lt} = \sum_{k \in K_{ijr}} \delta_{k(i,j,r)}^{lt} P_{k(i,j,r)} \qquad \forall i,j,r,l,t \tag{12}$$

where $\delta_{k(i,j,r)}^{lt}$ is the empirical incidence indicator:

$$\delta_{k(i,j,r)}^{lt} = \begin{cases} 1 & \text{if path } k(i,j,r) \text{ uses link } l \text{ at time } t \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

3.5 An Assignment Free DODME Approach

The possibility of estimating an empirical assignment (12) enables the reformulation of DODME replacing the assignment matrix provided by the DTA by the empirical one (12) and then (4) can be rewritten as:

$$\overline{y}_{lt} = \sum_{(i,j) \in N} \sum_{r=1}^{t} \overline{a}_{ijr}^{lt} x_{ijr} \tag{14}$$

where $\overline{y}_{lt}$ is the estimated flow in link l at time period t, x_{ijr} is the flow departing origin $i \in I$, with destination $j \in J$, at time interval r, and $\overline{a}_{ijr}^{lt}$, the empirical assignment matrix, is the fraction of trips from origin i with destination j, departing at time r, that reaches link l at time t.

Then, assuming that traffic counts measured by real detectors placed onto the network are available. They are measured for each time interval and denoted by: $\hat{y}_{lt}$, where $l \subseteq \hat{L}$ is the link with the detector of the network and t the time interval when it is measured.

The research question addressed in this paper can be formulated in the following terms: if the data collected from a sample of GPS tracked vehicles provide us with a discretized time estimate of the target OD matrix $\widehat{\boldsymbol{X}} = \left[\hat{x}_{ijr}\right]$, and a suitable processing, (Janmyr and Wadell 2018; Krishnakumari et al. 2019: Nassir 2014), provides a sound empirical estimate of $\overline{a}_{ijr}^{lt}$, then the expansion of the sampled target matrix to estimate the OD matrix can be done in terms of the scaling factors per origins, $\alpha_i, i \in I$, and per destinations $\beta_j, j \in J$, such that:

$$x_{ijr} = \alpha_i \beta_j \hat{x}_{ijr}, \forall i \in I, \forall j \in J, \forall r \in T \tag{15}$$

If $\hat{y}_{lt}, l \in \hat{L} \subset L, t \in T$ are the link flows measured at the counting stations, in a subset $\hat{L} \subset L$ of the network links, the Dynamic Data-Driven OD Matrix Estimation problem can be formulated as the following optimization problem of finding the values of the scaling factors $\alpha_i, i \in I$ and $\beta_j, j \in J$, without the need of conducting the traffic assignment at the lower level of (1), exploiting the empirical assignment matrix $\overline{a}_{ijr}^{lt}$. However, if an historical OD

is available from other sources, then a seed matrix x_{ijr}^0 can be generated combining that historical OD matrix x_{ijr}^H and the observed OD matrix $\hat{x}_{ijr}$ obtained from GPS tracked trips, and it is denoted as $\boldsymbol{X^0} = [x_{ijr}^0]$. Moreover, a third possibility can be studied when both OD matrices, the one that comes from GPS data and the Historical OD matrix, are available. This possibility consists on generating a seed OD matrix as a combination of the two sources:

$$x_{ijr}^0 = \begin{cases} \hat{x}_{ijr} & when\ only\ \hat{x}_{ijr}\ is\ available \\ f(\hat{x}_{ijr}, x_{ijr}^H) & when\ both\ matrices\ are\ available \\ x_{ijr}^H & when\ only\ x_{ijr}^H\ is\ available \end{cases} \quad (16)$$

From (14) and (15), the proposed new formulation of the DODME problem is:

$$\min_{\alpha_i,\beta_j} \left[\sum_{l\in\hat{L}} \sum_{t\in\mathcal{T}} \left(\hat{y}_{lt} - \sum_{(i,j)\in N} \sum_{r=1}^{t} \alpha_i \beta_j \bar{a}_{ijr}^{lt} x_{ijr}^0 \right)^2 \right] \quad (17)$$

$$s.to \quad \alpha_i, \beta_j \geq LB$$

The variables of the problem are multiplicative scaling factors for each origin, α_i, and destination, β_j. They are inspired on gravity models where bi-dimensional constraints for rows and columns are set. The minimization problem is solved iteratively with the L-BFGS-B method appropriated for constrained non-linear problems using the version available in python package *scipy.optimize*.

Although, theoretically *LB* should be a non-negativity constraint for all scaling factors α_i, β_j, from a practical point of view, $\alpha_i = 0$ or $\beta_j = 0$ imply to convert a positive OD flow of the seed OD matrix from a certain origin or certain destination to 0. Therefore, having into account that the seed OD matrix, Equation (16), comes from a reliable former project or from counts of real data, the scaling factors cannot be 0, and therefore, it should be *LB*>0.

The conceptual computational scheme of the proposed Assignment Free DODME approach, powered by the ICT applications capturing GPS data trajectories is depicted in Figure 3.

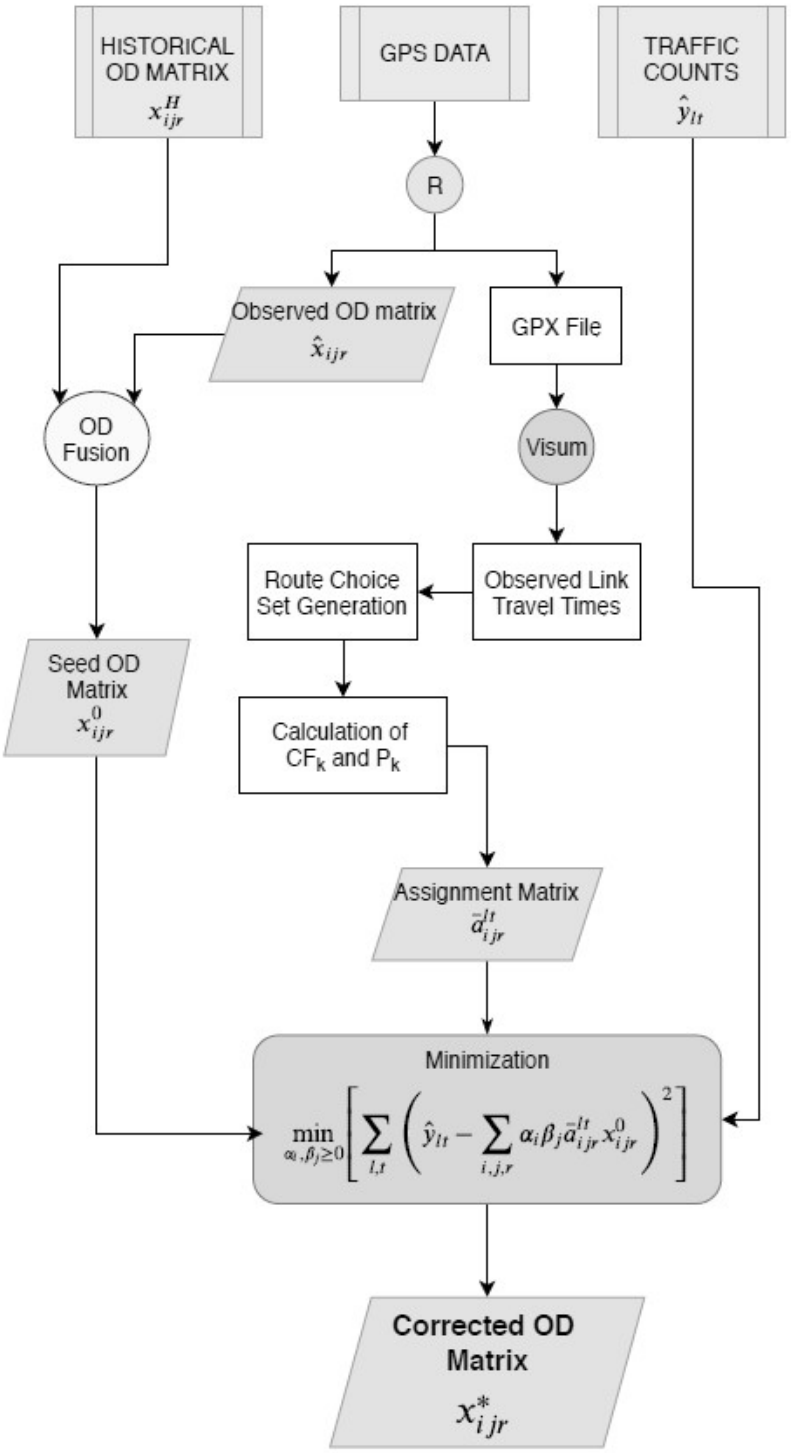


Figure 3 – A Data Driven Assignment Free DODME

4. A COMPUTATIONAL PROOF OF CONCEPT OF THE ASSIGNMENT FREE DATA DRIVEN DODME

A first test with the Torino CDB subnetwork, Figure 4, has been conducted in order to check the functional feasibility of the algorithmic chain, Figure 3, of the new approach.

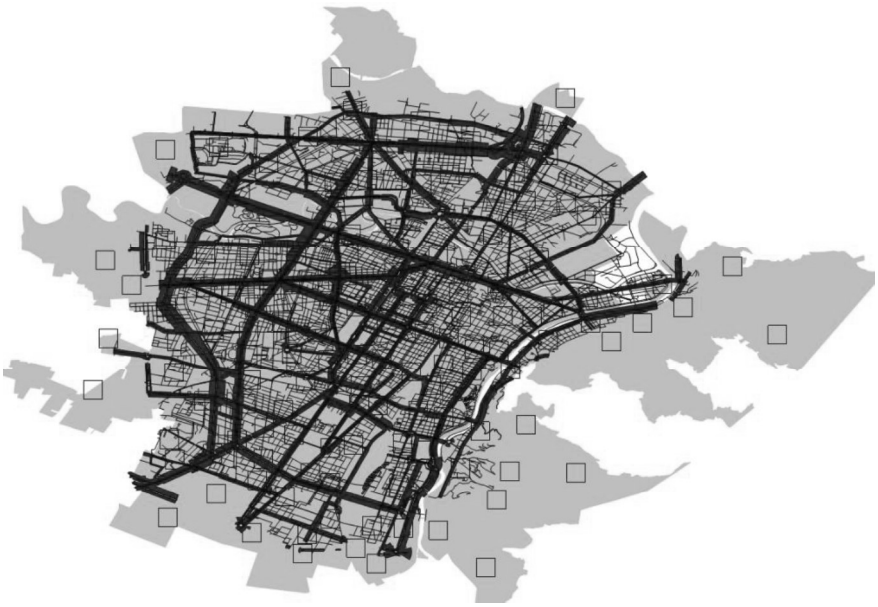


Figure 4 – Torino CBD Subnetwork

A summary of the model and the real inputs is shown in the Table 2:

TIME PERIODS	4
ZONES	285
DETECTORS	561
OD PAIRS X TIME	325K
INRIX WAYPOINTS	2.9M
INRIX TRIPS	166k
OBS OD > 0	32k (12%)
HIST OD > 0	120k (42%)

Table 2 – Torino's CBD Experiment characteristics

4.1 GPS Data: Building the Seed OD matrix

6 months of GPS data coming from INRIX for all Piemonte region are available. These are 220M waypoints for 5.9M trips. There is no further information regarding the type of vehicle (fleet/private) and it points that there was no distinction between them. Consequently, the only filtering that can be done relies on zoning and weekday selection.

After the pre-processing and filtering steps, the final dataset of waypoints (GPX file) contains 2.9M of waypoints for 166k trips (reduction to the 1.31% of the raw data). The observed OD matrix, $\hat{x}_{ijr}$, has 32k positive values, which is the 12% of the matrix.

The GPS data have to be preprocessed in order to filter the useful data for the method proposed in this research paper. In a first step, only weekdays trips (Tuesday, Wednesday and Thursday) in the considered time interval have been selected.

After that, using a SHP file with the information of the Transportation Zoning System (TAZ) of the model, an Origin and a Destination Zone has been assigned to each trip. Those trips that begin and/or end out of the study area, are cut and considered to be started/ended on the first/last zone of the network where they are observed.

There are two outputs of this procedure:

- *GPX file*: This GPX file is generated with the filtered and processed trips. They are coded in a GPX file to be inserted in VISUM GPX import tool.
- *Observed OD matrix*: The filtered and processed trips are counted by Origin, Destination and Departing Time to be translated to an observed OD matrix, $\hat{x}_{ijr}$.

After the Map Matching into VISUM, the total number of PrTPaths is 130k, which means the 78% of the sample. This loss is acceptable due to the short trips.

The observed link travel times per time interval, $\hat{tt}_{lt}$, computed by averaging the PrTPaths' Interpolated travel times at link level after propagating the time, covers the major part of network, resulting a connected network, Figure 5. Furthermore, the number of observations per link and time interval is right-skewed distributed as shown in Table 3.

min	Q1	Median	Mean	Q3	max
1	7	24	50	69	658

Table 3 – Torino's CBD Experiment: statistics for number of observations per link

We observed that the more observations a link has, the lower the standard deviation of its travel times is, which is consistent.

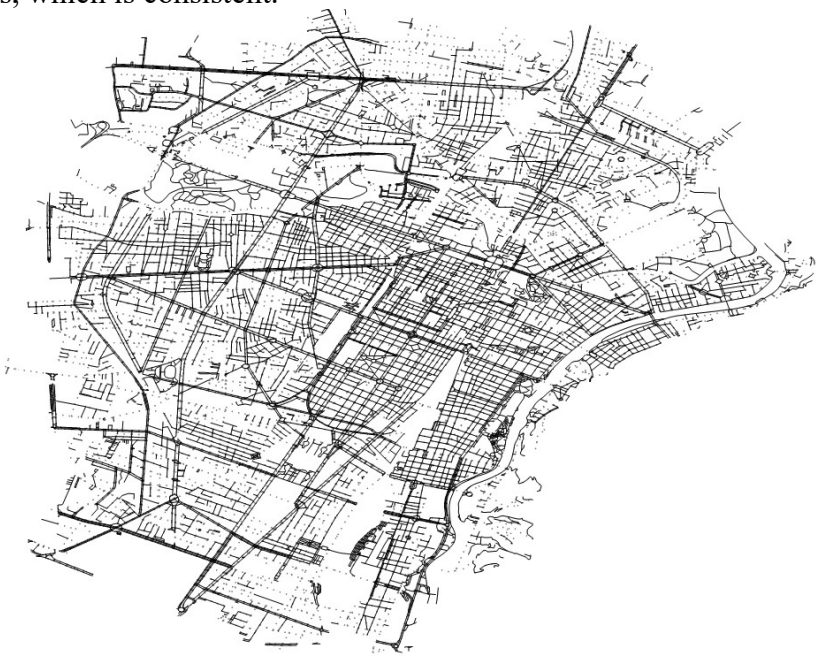


Figure 5 – Links with observed link travel time

As mentioned in Section 3.2, an approximated observed link travel time have been imputed to the missing data (red dotted links in Figure 5), using Equation (8). Despite that fact, a deeper analysis of the observed links shows that the 93% of the total flow uses links that were captured by the GPS sample.

4.2. Stochastic Assignment tool to generate the Route Choice Set

The Stochastic Assignment of Visum is used to generate the Route Choice Set, K_{ijr}, for each

Origin i, Destination j and Departing Time r. N_{max} has been set to 6. In this experiment, a total of 452,123 possible paths have been generated. In order to be efficient in memory, only paths for OD pairs that present a positive flow in the Seed OD matrices are generated, and this avoids more than a half of calculations that are not needed.

4.3. Calculation of CF_k , P_k and $\overline{a}_{ijr}^{lt}$ and Optimization

The calculation of CF_k, P_k and $\overline{a}_{ijr}^{lt}$ has been conducted applying (9) and (12). As mentioned in Section 3.3, the term CF_k acts as a penalization for the discrete choice. It penalizes those paths that are similar to others in the same set K_{ijr} and then, it reallocates the flows accordingly, increasing or decreasing the corresponding flow. In order to visualize the effect of CF_k to the paths flow distribution, an OD flow has been selected, and its corresponding Path Choice Set is depicted in Figure 6, where 6 paths resulted from the Stochastic Assignment. Since there are 3 paths that are very similar, they were clustered by similarity in 4 (I, II, III, IV), to understand better the role of the Commonality Factor.

Figure 6 – Path set (K_{ijr}) for a certain Origin and Destination in the real network

As shown in Figure 6 and Table 4, set I is compound by 3 very similar paths, and without the penalization, i.e. using only travel times, more than one half of the flow would be assigned to I. On the other hand, when the penalization is applied, a 12.13% of the flow is assigned to different paths. In addition, the same happens to path IV, which also loses flow because this path shares the most of its trajectory with the other paths.

	Length	Observed Time	P_k without CF_k	P_k with CF_k	% of gain
I	6.86 km	20 min 19 s	51.16%	39.03%	-12.13%
II	7.28 km	20 min 48 s	16.60%	22.22%	+5.63%
III	7.47 km	22 min 7 s	15.58%	27.05%	+11.47%
IV	7.51 km	20 min 43 s	16.67%	11.70%	-4.96%

Table 4 – Results of Path distribution for a certain OD pair in the real network

The optimization procedure has been launched with 2 different stopping criteria: a maximum of 100 iterations or a threshold for the relative error of the objective function which is $thrsh = 0.005$. It stopped after 16 iterations, satisfying the second stopping criterion. For the plots of Figure 7, the relative error threshold has not been taken into account in order to

show clearer the evolution of the minimization process.

Since there is an available Historical OD matrix, the seed OD matrix used is $x_{ijr}^{0} = x_{ijr}^{H}$. Regarding the constraints of (17), LB has been set to $LB = 0.5$ because:

- As mentioned above, the seed OD matrix is a Historical OD matrix and its total number of trips fits with the real counts of the network.
- In the worst case, when $\alpha_p = \beta_q = 0.5$, the reduction of $x_{pqr}^{*} = 0.5 \cdot 0.5 \cdot x_{pqr}^{H} = 0.25 \cdot x_{pqr}^{H}$ would be acceptable.

The scaling factors, which are the variables of the Objective Function, were initially set to 1. The objective function (17) shows a nice and fast decrease and convergence, see Figure 7a. Moreover, the convergence is clear when Figure 7b is checked.

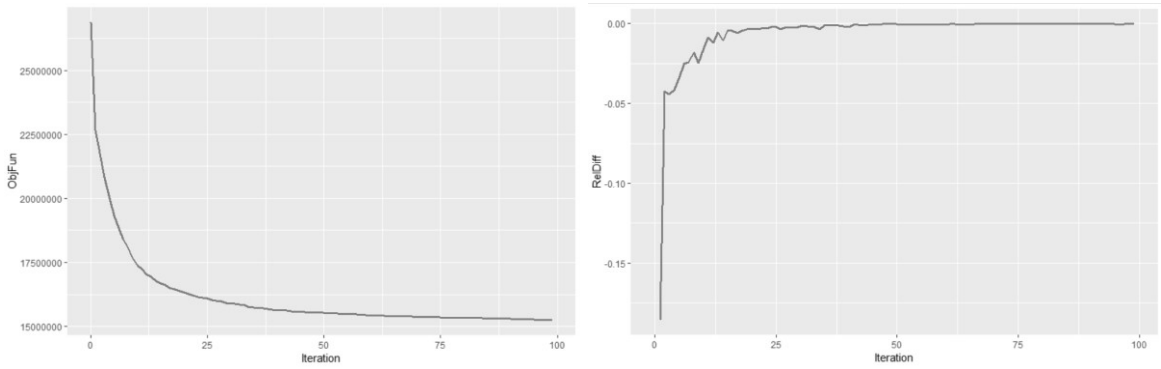


Figure 7 – (a) Descent of the Objective Function. (b) Relative error with previous iteration

A comparison analysis for real traffic counts and estimated counts is shown in Figure 8. On the left, the initial traffic counts, y_{lt}^{0}, using the seed OD matrix, x_{ijr}^{H}, and the calculated assignment matrix, $\bar{a}_{ijr}^{lt}$, are plotted in correspondence with the real traffic counts. On the right, there are plot the traffic counts after the optimization procedure:

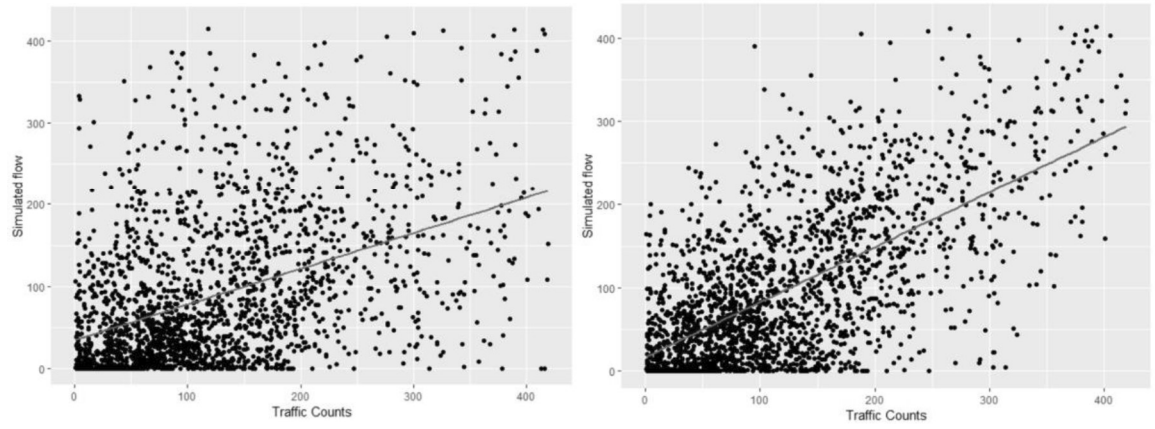


Figure 8 – Traffic Flow counts before (left) and after (right) the optimization procedure

The fit (R^2) of these traffic counts is 0.2385 at the beginning and 0.5634 when the convergence criterion is met. Moreover, after the minimization procedure, the slope of the regression line has become nearer to 1, from 0.49 to 0.74. In both Figure 8, there is a notorious quantity of detectors that capture real traffic counts, but the simulated flows are

almost zero. A deeper analysis, we attribute this phenomenon to the fact that these detectors are placed in locations which are almost not used by the paths produced by the Stochastic Assignment, Section 3.2. Therefore, the calculated assignment matrix does not assign flow to these detectors. This could be solved if the Route Choice Set Generation algorithm will improve, either by incrementing N_{max} and having more paths for each OD pair or maybe by using a more sophisticated method to produce these paths. Moreover, another alternative could be to exclude these detectors of the minimization procedure, when the number of counts is marginal, because they can perturbate other OD flows to compensate them on the objective function.

5. CONCLUSIONS

The identification of the role of the assignment matrix, reported in Ros-Roca el al. 2019, raised the main research question addressed in this paper of whether the empirical traffic measurements enabled by the ICT applications, could be interpreted in terms of an empirical assignment and then an empirical assignment matrix could be generated. This research question has found so far a positive answer as we report, enabling a new formulation of the DODME problem as a nonlinear optimization problem.

This paper provides a proof of concept of the novel approach, that is, it checks the feasibility of the algorithmic chain. The proof of concept looks very promising from the computational point of view but the quality of the results is strongly dependent on the quantity and quality of the available data, since it is a data driven process.

Moreover, we want to highlight another advantage of this new methodology. The classic formulation of the DODME problem (1) is a highly underdetermined problem, with $|N| = |I \times J| = |I| \cdot |J|$ variables and only $|\hat{L}|$ equations, which can carry to different solutions far that lead to similar traffic counts but they show very different mobility patterns. The new formulation (17) reduces the number of variables to $|I| + |J|$ which is nearer to the number of detectors onto the network and, therefore, decreases the underdetermination of the DODME problem.

Since this is an ongoing research, there are many topics we want to investigate further. As for instance replicating it in synthetic experiments that will allow us to explore the sensitivity and the robustness of the described method to the quantity and the quality of the data, the GPS penetration rate, and other factors.

The real GPS data of Torino does not include a split of the GPS tracked vehicles into private vehicles on their daily trips and commercial fleet vehicles used for freight delivering in the city. This must be a very important distinction if the counting OD matrix derived from GPS data is used to define the seed OD matrix in the optimization process. Commercial vehicles are perturbing the OD pattern and affect directly to the estimation process of these matrices, since they are overrepresented in GPS tracking data.

An additional source of concern relies on how the data collection process conducted by the GPS data provider defines a trip. The provider of these data informed us that the data collection process splits long trips into shorter trips, when a random modification of the GPS device identifiers is performed as a routine to preserve privacy issues, which implies that a significant fraction of the origins and destinations do not correspond to the underlying mobility pattern. The effect of the bias introduced by this data collection policy must be

assessed using synthetic experiments.

Another section of the method to investigate further is the Commonality Factor penalization (9). It is interesting to understand the effect of the coefficient μ_{CF} and tune it in order to penalize better and perform a better calculated assignment matrix $\bar{a}_{ijr}^{lt}$.

Improvements can also be expected from the proposed alternative approaches, as well as, from other formulation of the objective function in the optimization problem and more accurate specifications of the bounds on the scaling factors.

ACKNOWLEDGEMENTS

This work has benefited of the technical discussions with Dr. Klaus Nökel, Head of technological Innovation at PTV Group, addressing the data processing aspects as well as the route choice selection; the technical support of Dr. Arne Schnek, Chief Software Engineer at PTV Group, with respect to the Visum issues. The heuristic for the calculation of the link travel times from the waypoints is based on a private information from Frauke Petersen from PTV Group and Alessandro Attanassi from PTV Sistema and, last but not least with Professor Guido Gentile of the Dipartimento di Ingegneria Civile Edile e Ambientale, Sapienza Università di Roma on the seminal ideas about the new optimization approach.

This research was funded by Industrial-PhD-Program 2017-DI-041, TRA2016-76914-C3-1-P Spanish R+D Programs and Secretaria d'Universitats-i-Recerca-Generalitat de Catalunya-2017-SGR-1749.

REFERENCES

BOVY, P., BEKHOR, S. AND PRATO, C. (2008) The factor of revisited path size. T*ransportation Research Record: Journal of the Transportation Research Board***,** No. 2076, pp. 132–140. doi: 10.3141/2076-15.

CODINA, E. AND L. MONTERO (2006). Approximation of the Steepest Descent Direction for the O-D Matrix Adjustment Problem. *Annals of Operations Research*. Vol.144, pp.329–362 in Springer US.

FREDERIX, R., F. VITI, R. CORTHOUT, AND C. TAMPÈRE (2011). New Gradient Approximation Method for Dynamic Origin-Destination Matrix Estimation on Congested Networks. *Transportation Research Record: Journal of the Transportation Research Board* 2263(1), pp.19–25.

FREDERIX, R., VITI F. AND TAMPÈRE C. (2013). Dynamic Origin-Destination Estimation in Congested Networks: Theoretical Findings and Implications in Practice. Transportmetrica *A: Transport Science* 9(6): pp.494–513.

JANMYR, J. AND WADELL D. (2018). Analysis of vehicle route choice during incidents, MsC Thesis, University of Linköping, Department of Science and Technology, LiU-ITN-TEK-A--18/020—SE

KRISHNAKUMARI P., VAN LINT H., DJUKIC T. AND CATS, O. (2019). A data driven method for OD matrix estimation, *Transportation Research C*, doi: 10.1016/j.trc.2019.05.014

LUNDGREN, J .T. AND PETERSON, A. (2008). A Heuristic for the Bilevel Origin-Destination-Matrix Estimation Problem. *Transportation Research Part B: Methodological* 42(4) pp 339–54.

NASSIR N., ZIEBARTH, J., SALL, E., AND ZORN, L. (2014). Choice Set Generation Algorithm Suitable for Measuring Route Choice Accessibility. Transp. Res. Rec. J. Transp. Res. Board, vol. 2430, no. 1, pp. 170–181.

ROS-ROCA, X., MONTERO, L., SCHNECK, A. and BARCELÓ, J. (2018). Investigating the performance of SPSA in Simulation-Optimization approaches to transportation problems. *International Symposium of Transport Simulation (ISTS'18)*, Transportation Research Procedia, Volume 34, pp. 83-90.

ROS-ROCA, X., MONTERO, L. AND BARCELÓ, J. (2019), Investigating the Quality of Spiess-Like and SPSA approaches for Dynamic OD Matrix Estimation, accepted for publication in Transportmetrica. doi: 10.1080/23249935.2020.1722282.

SPIESS, H. (1990). A Gradient Approach for the OD Matrix Adjustment Problem. *Centre for Research on Transportation, University of Montreal, Canada* 693 (Publication No. 693, CRT) pp. 1–11.

TOLEDO, T. AND KOLECHKINA, T. (2013). Estimation of Dynamic Origin-Destination Matrices Using Linear Assignment Matrix Approximations. *IEEE Transactions on Intelligent Transportation Systems* 14(2) pp 618–26.

YANG, Z., LU, Y. AND HAO, W. (2017). Origin-Destination Estimation Using Probe Vehicle Trajectory and Link Counts. *Jorunal of Advanced Transportation,* doi: 10.1155/2017/4341532