sciendo

Transport and Telecommunication, 2020, volume 21, no. 4, 245–254
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia
DOI 10.2478/ttj-2020-0019

A BIG DATA DEMAND ESTIMATION MODEL FOR URBAN CONGESTED NETWORKS

Guido Cantelmo[1], Francesco Viti [2]

[1] Department of Civil, Geo and Environmental Engineering Chair of Transportation Systems Engineering
Technical University of Munich, (TSE) Arcisstraße 21, 80333 Munich, Germany
g.cantelmo@tum.de

[2] MobiLab Transport Research Group, Department of Engineering, University of Luxembourg
6, Avenue de la Fonte L-8367, Esch-sur-Alzette, Luxembourg
+352 464445352, francesco.viti@uni.lu

The origin-destination (OD) demand estimation problem is a classical problem in transport planning and management. Traditionally, this problem has been solved using traffic counts, speeds or travel times extracted from location-based sensor data. With the advent of new sensing technologies located on vehicles (GPS) and nomadic devices (mobile and smartphones), new opportunities have emerged to improve the estimation accuracy and reliability, and more importantly to better capture the dynamics of the daily mobility patterns. In this paper we frame this new data in a comprehensive framework which estimates origin-destination flows in two steps: the first step estimates the total generated demand for each traffic zone, while the second step adjusts the spatial and temporal distribution on the different OD pairs. We show how mobile data can be used to obtain OD matrices that reflect the aggregated movements of individuals in complex and large-scale instances, while speed information from floating car data can be used in the second step. We showcase the added value of big data on a realistic network comprising Luxembourg's capital city and its surrounding. We simulate traffic by means of a commercial simulation software, PTV-Visum, and leverage real mobile phone data from the largest telco operator in the country and real speed data from a floating car data service provider. Results show how OD estimation improves both in solution reliability and in convergence speed.

Keywords: Dynamic OD estimation; mobile phone data; bi-level optimisation

1. Introduction

Dynamic traffic models represent essential tools for efficient and cost-effective transport planning, for assessing properties such as robustness and resilience, and for managing traffic in real time. These models take as input the demand flows from each origin and destination and at any time period, and in turn estimate and/or predict route and link flows and travel times.

In order to generate the mobility demand, usually taking the form of Origin-Destination (OD) matrices, traditional approaches combine data and mathematical models translating the individual decision-making process into aggregated sub-models (generation, distribution, mode and route choice assignment), which are based on different assumptions related to mobility principles, and are characterised by a number of parameters that are calibrated using traffic data, the most adopted ones being traffic counts obtained from loop detectors (Cascetta and Nguyen, 1988). Unfortunately, the demand matrix is at best a coarse representation of the individuals' mobility and activity-travel patterns – such as the typical commuting behaviour during a working day. However, daily demand patterns can substantially differ because of several elements, including weather conditions or road works, big events as well as because of the inherent variability, stochasticity and complexity of the travel demand. Deviations between estimated and actual demand patterns can be mitigated by adjusting the results using the traffic measurements. This problem, which is known in the literature as the Dynamic Origin-Destination Estimation (DODE) problem, exploits a properly specified objective function for estimating the time-dependent OD flows.

While the DODE problem has been initially treated as an extension of its static counterpart, the last decades have witnessed a considerable effort in developing methodologies able to deal with the spatial and temporal dynamics of the traffic flows. As traffic models are applied in both offline (medium-long term planning and design) and online (real-time management) contexts, the DODE is commonly

classified between sequential or simultaneous approaches, where usually the first is adopted for online while the second for offline applications (Cascetta *et al*., 1993). By limiting our focus to the offline case (for online OD estimation we refer to e.g. Cantelmo *et al*., 2020), the DODE is usually formulated as a bi-level optimisation problem, where in the upper level the OD flows are updated by minimising the error between simulated and observed traffic data, while in the lower level the DTA solves the combined Route Choice (RC) and Dynamic Network Loading (DNL) problems (Tavana, 2001). A generic formulation can be expressed as:

$$(\boldsymbol{d}_1^*, \dots, \boldsymbol{d}_n^*) = argmin \begin{bmatrix} z_1(\boldsymbol{l}_1, \dots, \boldsymbol{l}_n, \widehat{\boldsymbol{l}_1}, \dots, \widehat{\boldsymbol{l}_n}) + \\ +z_2(\boldsymbol{n}_1, \dots, \boldsymbol{n}_n, \widehat{\boldsymbol{n}_1}, \dots, \widehat{\boldsymbol{n}_n}) + \\ +z_3(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \widehat{\boldsymbol{x}_1}, \dots, \widehat{\boldsymbol{x}_n}) + \\ +z_4(\boldsymbol{r}_1, \dots, \boldsymbol{r}_n, \widehat{\boldsymbol{r}_1}, \dots, \widehat{\boldsymbol{r}_n}) + \end{bmatrix}, \tag{1a}$$

where $\boldsymbol{l}/\hat{\boldsymbol{l}}$ represent, respectively, simulated and measured link performances, $\mathbf{n}/\hat{\mathbf{n}}$ calibrated and observed values on the node, $\mathbf{x}/\hat{\mathbf{x}}$ indicate the estimated and historical values for the OD flows (seed matrix) and $\mathbf{r}/\hat{\mathbf{r}}$ the simulated and observed route flows. Finally, $\boldsymbol{d}_n^*$ designates the estimated demand matrix for time interval n, while $z: \{z_1, z_2, z_3, z_4\}$ is the estimator of the errors between simulated/estimated and measured/a priori values for the corresponding parameters. Each estimator may be specified depending on the available data. The dependence between supply and demand in Equation (1a) is obtained directly by performing a dynamic traffic assignment (DTA), so that:

$$\begin{aligned} \boldsymbol{l}_1, \dots, \boldsymbol{l}_n &= \mathbf{F}_{\boldsymbol{l}}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n), \\ \boldsymbol{n}_1, \dots, \boldsymbol{n}_n &= \mathbf{F}_{\boldsymbol{n}}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n), \\ \boldsymbol{r}_1, \dots, \boldsymbol{r}_n &= \mathbf{F}_{\boldsymbol{r}}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n), \end{aligned} \tag{1b}$$

with $\mathbf{F}_{\boldsymbol{l}}$, $\mathbf{F}_{\boldsymbol{n}}$ and $\mathbf{F}_{\boldsymbol{r}}$ specified by a selected Dynamic Traffic Assignment (DTA) model.

Typically, link flows are available in different points of the network, hence z_1 is usually specified, together with the component controlling the deviation from the seed matrix z_3. Additionally, link speeds and densities have been proved to capture the non-linear relation between demand and supply parameters (Balakrisnha *et al*., 2007; Frederix *et al*., 2010). Moreover, recent works showed how more elaborate information, such as point-to-point data, can also be included in this function and in turn largely improve the overall estimation accuracy (Antoniou *et al*., 2016; Barcelò and Montero, 2015; Mitsakis *et al*., 2013). With the advent of new information, communication and sensing technologies, obtaining node-specific or route-specific data has become easier. Antoniou *et al*. (2011) proposes a new classification based on the functionalities of the sensor – e.g. point-sensors, point-to-point sensors, area-wide sensors. For instance, route (point-to-point) information can be extracted from GPS data installed on probe vehicles to obtain route flow proportions (Nigro *et al*., 2017). Additionally, more recent works have done a significant progress into including new data sources, such as Call Detail Records (CDR), GSM data, sensing data and geospatial data especially for modelling the spatial and temporal dynamics of trips generation and distribution (Di Donna *et al*., 2015; Toole *et al*., 2015; Carrese *et al*., 2019).

An additional advantage of the approach presented in Equation (1) is that all variables are jointly estimated, considering that OD flows over different time intervals are likely to be correlated (Frederix *et al*., 2013). However, for large networks, this approach becomes less reliable since often sufficient traffic data observations are not available (Marzano *et al*., 2009). Therefore, being an underdetermined problem, its solution strongly relies on the prior knowledge of the OD flows and its structure. This prior, or seed matrix, is usually estimated using the so-called *stationary models*, the most popular of which are the four-step model (McNally, 2007), and activity-based travel demand models (Ben-Akiva and Bowman, 1995). However, if the structure of the seed matrix is different from the real one, the estimation can converge to a local optimum that could be very different from the actual solution (Frederix *et al*., 2013). A possible solution to overcome this issue is to reformulate the objective function in order to reduce the number of variables. This can be done, for instance, by using Principal Component Analysis (PCA) (Djukic *et al*., 2012). Alternatively, Cascetta *et al*. (2013) introduced the so-called *quasi-dynamic* assumption, which assumes that the generated demand for a certain OD pair is time dependent, while its spatial distribution remains invariant for a certain time period. Under this assumption, the DODE problem becomes less underdetermined and more likely to find more reliable results. Similarly, Cantelmo *et al*. (2014) proposed a two-step procedure, which separates the DODE in two sub-optimization problems. The first step searches for generation values that best fit the traffic data while keeping spatial and temporal distributions constant. In the second step, the standard bi-level procedure searches for a more reliable demand matrix.

In this work, we adopt the two-step estimation framework developed by Cantelmo *et al*. (2014) and leverage big mobility data especially for obtaining useful information to construct the seed matrix. We use mobile phone data in the first step, i.e. in estimating the total number of trips produced or attracted to each zone and speeds from Floating Car Data (FCD) to improve the results in the second step. Section 2 will present the estimation framework together with the adopted gradient approximation method used to explore the solution space. Section 3 demonstrates the effectiveness of using GSM and FCD data for improving the estimation results and the convergence speed of the results in large scale networks. To support the claim that the model is ready for practical implementation, it is interfaced with PTV-Visum, a widely adopted software tool for traffic analysis. Finally, section 4 provides the main conclusions of the work.

2. Methodology

As pointed out by Antoniou *et al*. (2016), the seed matrix is a key input for all state-of-the-art DODE models. By separating the estimation process into a first step that aims at estimating the total number of trips generated by each zone and a second step that focuses on the spatial and temporal distribution of the OD flows, we show how to more effectively use different (big) mobility data sources, such as mobile phone data (which is a more reliable source for capturing the temporal profile of the demand for all modes of transport) and GPS/floating car data, which is more indicated to capture the spatial and temporal variations of the supply by providing speed profiles at link at route levels.

2.1. Problem framework: Two-Steps approach

This section will only briefly introduce the Two-Step approach. An interested reader can find more details in (Cantelmo *et al*., 2014; Cantelmo *et al*., 2015a). The Two-Step approach is an ideal tool for applications on large-scale networks and for using different data sources. The first step improves the historical demand matrix by performing a broad evaluation of the solution space and estimating a "good" updated seed matrix to be used in the second step. This way, the proposed model reduces the number of variables to be estimated in the first step. The idea of performing successive iterations and linearisations has been already introduced and validated in Ashok and Ben-Akiva (2001) for online DODE, showing that the reliability of the results generally increases.

Following Cascetta *et al*. (2013), the objective function described in equation (1a) can be enhanced by exploiting information on aggregated socio-demographic data such as generation data by traffic zones. The objective function (1a) can be then reformulated as:

$$(\boldsymbol{E}_1^*, \dots, \boldsymbol{E}_n^*) = argmin \begin{bmatrix} \mathrm{z}_1'(\boldsymbol{l}_1, \dots, \boldsymbol{l}_n, \widehat{\boldsymbol{l}_1}, \dots, \widehat{\boldsymbol{l}_n}) + \\ +\mathrm{z}_2'(\boldsymbol{n}_1, \dots, \boldsymbol{n}_n, \widehat{\boldsymbol{n}_1}, \dots, \widehat{\boldsymbol{n}_n}) + \\ +\mathrm{z}_3'(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \widehat{\boldsymbol{x}_1}, \dots, \widehat{\boldsymbol{x}_n}) + \\ +\mathrm{z}_4'(\boldsymbol{r}_1, \dots, \boldsymbol{r}_n, \widehat{\boldsymbol{r}_1}, \dots, \widehat{\boldsymbol{r}_n}) + \end{bmatrix} \tag{2a}$$

subject to

$$x_n^{OD} = E_n^O d_{D|O}^{Seed,n} \qquad \forall O, \forall D, \forall n\,, \tag{2b}$$

where:

- $E_n^{\ O}$ = generated flow from traffic zone O and time interval n;
- E_n^* = generation vector containing the generated flow from all zones in time interval n;
- x_n^{OD} = demand flow from origin zone O to destination zone D in time interval n;
- $d_{D|O}^{Seed,n}$ = distribution share to traffic zone i from traffic zone O in time interval n.

Constraint (2b) is the main difference with the general quasi-dynamic formulation proposed in Cascetta *et al*. (2013), where $d_{D|O}^{Seed,n}$ is updated during the optimization process, whereas constraint (2b) assumes a constant value of the distribution, bringing two major advantages. First, as the number of unknown variables strongly decreases, the approach can be applied to larger networks. Second, this approach does not necessarily require to explicitly account for historical OD flows within the objective function. As pointed out in Section 1, historical OD flows are usually included within equation (1a) in order to reduce the number of possible solutions. However, this information is already considered within

constraint (2b), that over-impose, to the estimated matrix, the spatial/temporal structure of the historical demand. However, a main drawback of this formulation is that it is likely to provide a relatively poorer fit of the traffic data, as already pointed out in Cantelmo *et al*. (2015a). On the other hand, the goal of the first step is to act on the seed matrix in order to obtain a "right level of demand", to be used in the second step in order to optimise the dynamic distributions OD trips. This leads to the following considerations:

- Total generated trips can limit a demand overestimation during the DODE, which is otherwise likely to occur when dealing with congested networks;
- As generation models are considered the most reliable models in transport engineering applications, total generated trips are more reliable than OD trips;
- Adopting the generation values inside the DODE, as in (2), reduces the number of variables.

2.2. Optimisation method: Simultaneous Perturbation Stochastic Approximation

The optimisation method adopted in this paper is the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm proposed by Spall (2012). The SPSA is a stochastic version of the deterministic finite difference method, which is computationally too expensive for large networks (Frederix *et al*., 2011), and it as has been proven to be very effective for tackling the DODE problem, and many authors proposed enhanced versions of the original algorithm (e.g. Cipriani *et al*., 2011; Cantelmo *et al*., 2014; Antoniou *et al*., 2015; Tympakianaki *et al*., 2015; Qurashi *et al*., 2020). By assuming a one-sided perturbation as in Cipriani *et al*., (2011) we compute the approximated gradient $\boldsymbol{G^i}$ at each iteration as:

$$\widehat{\boldsymbol{g}}_k(\boldsymbol{\theta}^i) = \frac{z(\boldsymbol{\theta}^i + c^i\Delta^k) - z(\boldsymbol{\theta}^i)}{c^i}\begin{bmatrix}(\Delta_1^k)\\ \vdots \\ (\Delta_r^k)\end{bmatrix}, \tag{3a}$$

$$\boldsymbol{G^i} = \overline{\boldsymbol{g}}(\boldsymbol{\theta}^i) = \frac{\sum_{k=1}^{\boldsymbol{Grad\_rep}} \widehat{\boldsymbol{g}}_k(\boldsymbol{\theta}^i)}{Grad\_rep}, \tag{3b}$$

with $\boldsymbol{\theta}^i$ the vector with the estimated variables, $z(\boldsymbol{\theta}^i)$ the objective function value in $\boldsymbol{\theta}^i$, c^i the perturbation step, *Grad_rep* is the number of replications to compute the average gradient and Δ is a vector with elements {-1,1}.

Given a properly specified objective function and a descent direction – the gradient $\boldsymbol{G}^i$ – the parameters are updated at each iteration according to:

$$\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i - \alpha^i \boldsymbol{G}^i, \tag{4}$$

where α^i is the stepsize and $\boldsymbol{\theta}^i$ is again the vector of parameters to be updated, the OD or the Generation flows if we are minimizing, respectively, objective functions (1) or (2). Concerning the value of α^i, we proposed to use a line search to find the optimal value in order to reduce the overall computational time.

Given the stochastic nature of the model, it is recommended to repeat the perturbation multiple times in order to obtain a good approximation. If only one replication is used, then $\boldsymbol{G^i} = \widehat{\boldsymbol{g}}_k$. The main advantage of using this formulation is that it allows to reduce the number of simulations needed while still providing a proper approximation of the gradient.

2.3. Including mobile phone data in the first step

While the correlation between network traffic states and mobile phone data is well known, this area-wide information is hard to implement within the DODE, since it provides at most the geographic position at connected antenna levels, so no direct match on the road network is possible. However, by clustering antennas located on the border of each traffic zone, it is possible to use active connections that are entering or exiting the zones (i.e. the number of handovers) and use them as proxy of the movements in and out of a certain zone (Derrmann *et al*., 2019). Unfortunately, mobile network data is subject to intrinsic errors (e.g. the users are split between multiple network operators, the degree of activity on the network as well as the general mobile penetration rates). However, this information can be used as input to estimate the temporal profiles of the generated demand on a certain area, as shown in Di Donna *et al*. (2015). In Cantelmo *et al*. (2015b) we proposed the following two criteria to exploit demand emission flows estimated through the mobile network data: 1) Antenna clusters need to be large enough to minimize the "ping-pong" effect, i.e. counting the same users 'bouncing' back and forth between two

antennas, and 2) Cluster edges shall be positioned so as to maximise the difference between number of people entering and leaving the study area. We will use the same approach in this paper on a large-scale instance.

2.4. Including floating car data in the second step

To consider the relation between the temporal characteristics of congestion and their impact on the spatial and temporal distribution of the OD flows, a utility-based choice model has been adopted (Cantelmo *et al*., 2018). Concerning the congestion dynamics on the supply side, traffic counts and floating car data can be used in a single estimation process to determine flows and speeds on all measured links. In practice, sensors are placed on a limited number of links, and for privacy concerns, floating car data are often aggregated and only average speeds are shared. This limits most of the application of this data for dynamic demand estimation, but we show in our case study that through the adoption of the Two-Step approach we can still reduce the estimation error systematically. Clearly, the availability of more detailed probe vehicles data such as GPS position would strongly be an asset, as shown for instance in Cipriani *et al*. (2015).

2.5. MAMBA-DEV and MAMBA-DEV-A packages

To be able to solve the DODE problem on large real-sized networks, we developed a Matlab package that uses PTV Visum as DTA model. The package allows performing assignment-free dynamic or static OD estimation, using a deterministic and/or stochastic approximation of the gradient. The model also includes the Two-Steps approach, as well as the possibility to use several data sources, including mobile phone data. While the MAMBA-DEV package has been designed for Luxembourg City, it can work with any network. Similarly, MAMBA-DEV-A package is a Matlab package that allows to use the utility-based DODE approach proposed in Cantelmo *et al*. (2018) in combination with PTV Visum. This package uses static OD matrices as an input and provides dynamic, purpose-dependent OD matrices based on the available traffic data. The departure time choice module of MAMBA-DEV-A has been used to create the dynamic OD matrix for this study.

3. Case study: Luxembourg City

3.1. Estimating the total generated demand

We show the application of the Two-Step approach on the road network of Luxembourg (Figure 1, left). The network consists of 3700 links and 1469 nodes. The network includes all national motorways up to the northern city of Ettelbruck, and from the capital to the east, west and south borders. Additionally, the network includes also primary and secondary roads. We collected socio-demographic data from the national open-data portal (https://data.public.lu). These data shared by the National Institute of Statistics (STATEC) include the growth of the population for each year, the population by province and the number of cross-borders. Based on these statistics, a static matrix for the morning commute has been estimated through a classical Four-Step model. We then employed the utility-based model developed in Cantelmo *et al*. (2018) to derive a within-day dynamic OD matrix. This dynamic matrix accounts for 46 traffic zones and represents the historical demand (Seed Matrix) for the experiments presented in the next sub-sections.

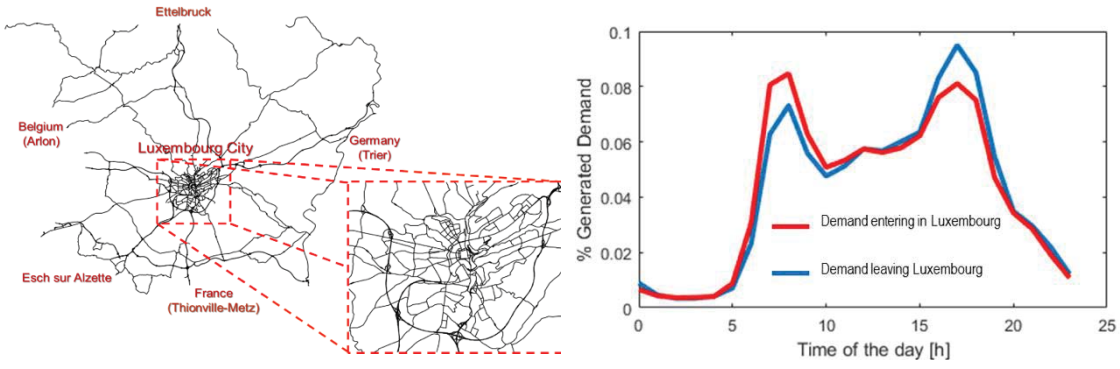

Figure 1. Road Network of Luxemborg City (left); Antenna densities in and out of the city (right)

In our case study, we used antenna densities and handovers from the national mobile phone operator Post, which has the largest number of customers in the country. The data has been aggregated into two large clusters; the first cluster captures the trips generated from the city to the external zones, while the second captures those entering Luxembourg City, as shown in Figure 1 (right). This procedure can be easily extended to any urban area, where connection handovers can be used to calculate the flows exchanged between the study area and the external centroids. In this study, we propose to use the difference between entering and exiting flow, as in Equation (3):

$$\Delta E_n^{GSM} = \frac{E_n^{GSM-IntZones}}{\sum_n E_n^{GSM-IntZones}} - \frac{E_n^{GSM-ExtZones}}{\sum_n E_n^{GSM-ExtZones}} \quad \forall n\,, \tag{5}$$

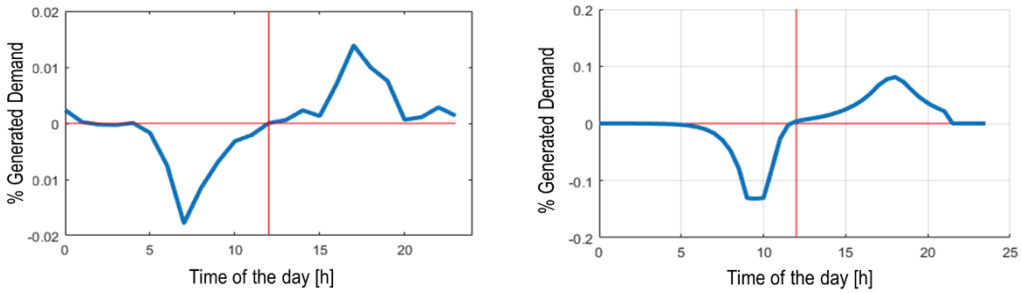

Figure 2. Profile obtained through the real-data (left); Profile obtained through the 4-Step approach and the proposed departure time choice model (right)

where $E_n^{GSM-IntZones}$ and $E_n^{GSM-ExtZones}$ are the aggregated mobile handovers during time period n by the internal or external zones, respectively. Figure 2 shows the profiles of ΔE_n^{GSM} for the real data (left) and the a-priori OD matrix (right). The profile obtained by combining the 4-Step model with a departure time choice model is comparable to the one obtained with the GSM data, indicating that the mobile phone data is a good proxy of the total generated flows. We can also identify two errors within the a-priori OD matrix. First, the average departure time for the morning peak is significantly earlier in the mobile phone data than in the 4-Step model. Second, there is a difference in the y-axis scale. The reason is that, in this application, we calculate the OD flows for the morning and evening commute, thus the demand in the afternoon is highly underestimated. This suggests that, by including Equation (5) within the objective function of the Dynamic Demand Estimation Problem, we can use GSM data as a soft constraint to correct the demand obtained through classical demand generation models.

3.2. Estimating the spatial and temporal distribution of OD flows

The National Open Data portal and the Luxembourgish National Road Administration agency collects and provides traffic counts on most of the motorways and primary roads of the Grand Duchy (Open data portal (https://data.public.lu/en/) and Luxemburgish National Road Administration (https://pch.gouvernement.lu/en.html)). After some data cleaning, 54 counting stations have been retained, all located on the main arterial roads going to Luxembourg City (highways, national roads) and on the ring road. Unfortunately, only three detectors are located inside the City ring. This means that we can expect to have a realistic representation of the demand at regional scale, but not inside the city. Additionally, traffic counts are aggregated on an hourly basis, which is too large for a network with an average free-flow travel time of 20 minutes, i.e. congestion dynamics cannot be properly captured. Moreover, the available data collected from traffic counts did not contain information on mean speeds, which are an essential input when dealing with large congested networks. To compensate for this lack of information, average speeds have been collected from Floating Car Data (FCD). For this study, only data related to the main highways and the ring road of Luxembourg were available. These have been provided by the Luxembourgish company Motion-S. The obtained information is based on the average of all probe vehicles and does not contain specifications about time and location. FCD carry definitely more information than just the average speeds, as demonstrated in Nigro *et al*. (2017), but national privacy laws do not allow sharing sensible data. Nevertheless, the available average speed broadly captures, in this study, the congestion on the ringway at a network level. The downside is that many possible solutions

exist, which can create congestion on the ring. As a consequence, the most logical solution for the DODE should be to keep the demand as close as possible to the historical demand, while at the same time reproducing the speed profile. Since this information is strongly aggregated, we expect the Two-Step approach to provide reliable results by exploiting the link flows and speeds as constraints within the objective function. This claim is numerically illustrated in the next section.

3.3. Results

In this paper we compare the Two-Steps approach (TS) with a more conventional Single-Step (SS) method. The Single-Step OD estimation is formulated in this paper as a single constrained optimisation problem. In both cases, the SPSA was adopted as the optimisation method. We performed two different sets of experiments:

I. Only traffic counts are included within the OF.
II. Traffic counts and GSM/FCD data are included within the OF.

We test two different scenarios. In order to test the effectiveness of the GSM data, we test the model on a synthetic scenario for Luxembourg City. The reason is that GSM data are only available within Luxembourg City but, as mentioned, real data are only available on the regional network. The second experiment uses real data (FCD and traffic counts) on the entire network, as previously discussed.

The Root Mean Squared Error (RMSE) is the chosen as estimator:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{q}_i - q_i)^2}{N}}, \tag{6}$$

where N is the number of observations, $\hat{q}_i$ is the observed value for the measured data and q_i is the simulated one.

3.3.1. Synthetic experiments

The reduced network consists of 2744 links, 1480 nodes and 17 traffic zones. OD flows are estimated over 24 hours assuming a 30-minutes departure interval. Under this assumption, the dynamic matrix contains 13872 variables to be estimated. The real matrix amounts to 239.966 trips.

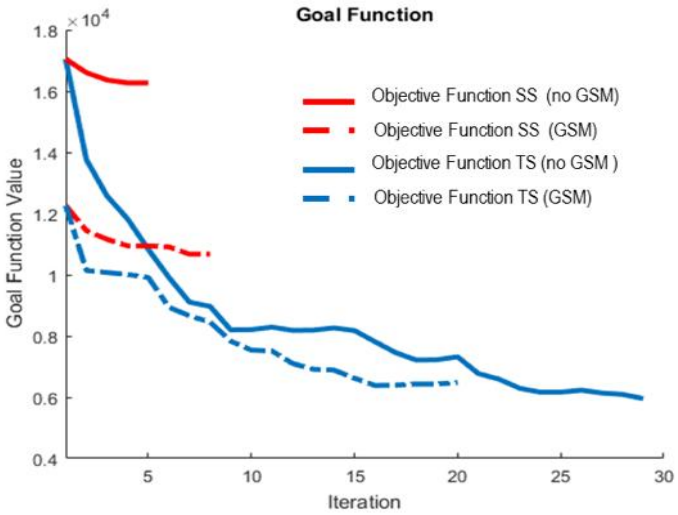

Figure 3. Performance of the Two-Steps approach with GSM as compared to Single-Step

Table 1. Root Mean Square Error of speeds and OD flows

	Seed	***TS***	***SS***
RMSE Speed (Km/h)	3.73	2.98	3.66
RMSE OD (Veh/h)	42.25	40.01	66.02

As shown in Figure 3, when the number of variables is large, the Single Step (SS) model performs only a local adjustment of the OD demand. Specifically, to obtain a reliable estimation of the gradient, the number of stochastic perturbations should be approximately 10% of the number of variables (Cipriani *et al*., 2011). This entails 1382 DTA simulations for each iteration. The Two-Steps (TS) approach achieves a similar error with and without the mobile phone network data. However, as reported in Figure 3, when mobile data are included within the OF, the number of iterations required for solving the DODE strongly decreases. As predictable, the same property is not observed for the SS model, which simply collapses on the closest local minimum. However, when this model is combined with the mobile network data, the error on the link flows decreases with respect to the base case presented in Scenario I (the RMSE is 3% lower). The error for each model is summarized in Table 1.

3.3.2. Experiments with real data

The second experiment includes the entire regional network previously introduced. In this experiment, we consider the morning peak period between 5 AM and noon (8 hours). The seed-matrix accounts for 307.544 trips and 16928 time-dependent OD pairs. We compare the estimation accuracy in terms of Root Mean Square Error (RMSE) on link flows, links speeds and how much the solution deviates from the OD pair when using the FCD. Figure 4 (right) depicts the Spider Chart of the estimation error for speeds, flows and seed-matrix – i.e. the initial point. For each measure, this relative error has been calculated as:

$$Rel_{error} = \frac{RMSE^{Measures}}{\max\left(RMSE^{Two-Step}, RMSE^{Single-Step}\right)}. \quad (7)$$

Figure 4 intuitively shows the dynamics behind the optimization. The Single-Step approach does not manage to move away from the initial point to reduce the error on the link flows. As the Two-Steps approach moves to a new solution during the first phase of the optimization, the distance with respect to the initial matrix increases, while the error on the link flows is twice smaller than the one for the Two-Step. However, the Two-Step also increases the error on the speeds, which was expected as this information has a low weight in the goal function.

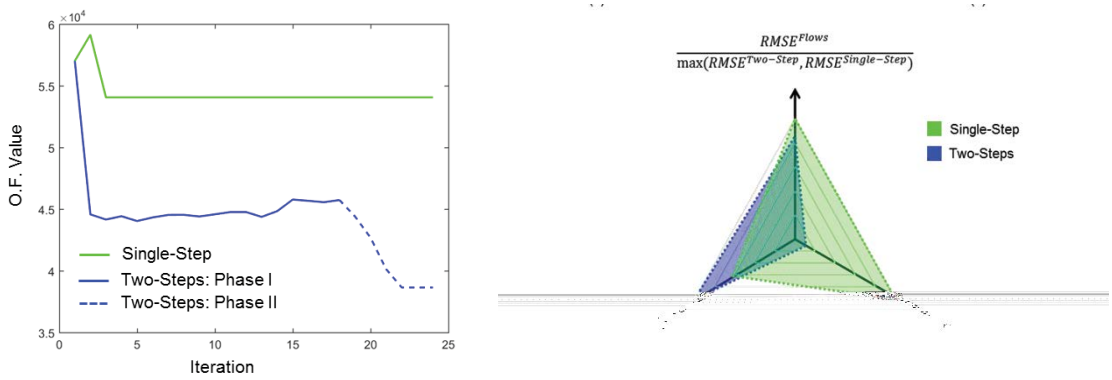

Figure 4. Performance of the Two-Steps approach with GSM as compared to Single-Step

A second result concerns the stability of the estimations. Similar dynamics are observed for the TS approach in both experiments. This means that the Two-Step approach not only manages to have a larger OF improvement in terms of speeds but also to provide more reliable results. These findings are in line with the conclusions already presented in Cantelmo *et al*. (2015b). In general, we can claim that, since the two steps approach sequentially reduces the dimension of the solution space while keeping a lower number of variables with respect to the conventional Single-Step approach, it will provide a more reliable estimation.

4. Conclusions

This paper introduced a novel dynamic demand estimation framework and showed how big mobility data such as mobile phones and floating car data can be adopted into a novel Two-Steps approach on large-scale congested networks.

From a methodological point of view, the proposed approach relaxes the strong limitation of having a good starting demand matrix. The capability of the DODE solution algorithm to correct the biases within the temporal and spatial structure of the demand is a strict requirement for having robust results. Mobile phone data is shown to improve the performances of the estimation. Then, we show that by using floating car data and link flows on the second step, the model is capable of further improving the estimation results, while not affecting significantly the structure of the OD matrix.

Next step in this research will be to extend the study to multimodal networks, an in particular the utility-based model will be extended to include mode choice and transit data.

Acknowledgements

The authors acknowledge the European Commission for providing the financing grant: FEDER MERLIN. This research has been partially sponsored by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 75446. We also thank Post and Motion-S Luxembourg for providing GSM and Speed data.

This paper is a revised and expanded version of the paper "A Big Data Demand Estimation Framework for Multimodal Modelling of Urban Congested Networks" presented at the 4th Conference on Sustainable Urban Mobility, Skiathos Island, Greece, May 24-25, 2018 and published in *Data Analytics: Paving the Way to Sustainable Urban Mobility*" of the book series *Advances in Intelligent Systems and Computing*, v. 879, Springer, 2019.

References

1. McNally, M. G. (2007) The Four-Step Model. In: *Handbook of Transport Modelling*, 1, Emerald Group Publishing Limited, pp. 35–53.
2. Cascetta, E. and S. Nguyen. (1988) A unified framework for estimating or updating origin/destination matrices from traffic counts, *Transportation Research Part B: Methodological*, vol. 22, no. 6, pp. 437–455, Dec. 1988
3. Cascetta, E., Inaudi, D. and G. Marquis. (1993) Dynamic Estimators of Origin-Destination Matrices Using Traffic Counts, *Transportation Science*, 27(4), pp. 363–373, Nov. 1993.
4. Cantelmo, G., Qurashi, M., Prakash, A.A., Antoniou, C., Viti, F. (2020) Incorporating trip chaining within online demand estimation. *Transportation Research Part B: Methodological*, 132, pp. 171-187.
5. Tavana, H. (2001) *Internally-consistent estimation of dynamic network origin-destination flows from intelligent transportation systems data using bi-level optimization*, PhD Thesis, 2001.
6. Balakrishna, R., Ben-Akiva, M., Koutsopoulos, H. (2007) Offline Calibration of Dynamic Traffic Assignment: Simultaneous Demand-and-Supply Estimation. *Transportation Research Record*, vol. 2003.
7. Frederix, R., Viti, F. and C. M. J. Tampère (2010) A density-based dynamic OD estimation method that reproduces within-day congestion dynamics. In: *13th International IEEE Conference on Intelligent Transportation Systems*, 2010, pp. 694–699.
8. Antoniou, C. *et al.* (2016) Towards a generic benchmarking platform for origin–destination flows estimation/updating algorithms: Design, demonstration and validation, *Transportation Research Part C: Emerging Technologies*, vol. 66, pp. 79–98, May 2016.
9. Barceló, J. and L. Montero (2015) A Robust Framework for the Estimation of Dynamic OD Trip Matrices for Reliable Traffic Management, *Transportation Research Procedia*, vol. 10, pp. 134–144.
10. Mitsakis, E., Grau, J.-M. S., Chrysohoou, E. and G. Aifadopoulou (2013) A Robust Method for Real Time Estimation of Travel Times for Dense Urban Road Networks Using Point-to-Point Detectors, *Transportation Research Board 92nd Annual Meeting*, 2013.
11. Antoniou, C., Balakrishna, R., & Koutsopoulos, H. N. (2011). A Synthesis of emerging data collection technologies and their impact on traffic management applications. *European Transport Research Review*, 3(3), 139–148.
12. Nigro, M., Cipriani, E. and A. Del Giudice (2017) Exploiting floating car data for time-dependent O-D matrices estimation, *Journal of Intelligent Transportation Systems: Technology, Planning & Operations*, 22(2), pp. 159-174.
13. Di Donna, S.A., Cantelmo, G., Viti, F. (2015) A Markov Chain dynamic model for trip generation and distribution based on CDR. In: *4th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2015*, pp. 243-250, June 2015.
14. Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A. and M. C. González. (2015) The path most traveled: Travel demand estimation using big data resources, *Transp. Res. Part C Emerg. Technol.*, vol. 58, Part B, pp. 162–177, Sep. 2015.

15. Carrese, F., Cantelmo, G., Fusco, G., Viti, F. (2019) Leveraging GIS data and topological information to infer trip chaining behavior at macroscopic level. In: *6th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems*, MT-ITS 2019, June 2019.
16. Frederix, R., Viti, F. and M. J. Tampère. (2013) Dynamic origin–destination estimation in congested networks: theoretical findings and implications in practice, *Transportmetrica: Transport Science*, 9(6), pp. 494–513, Jul. 2013.
17. Marzano, V., Papola, A. and F. Simonelli. (2009) Limits and perspectives of effective O–D matrix correction using traffic counts, *Transportation Research Part C: Emerging Technologies*, 17(2), pp. 120–132.
18. Ben-Akiva, M.E., Bowman, J.L. (1988) Activity Based Travel Demand Model Systems. In: Marcotte P., Nguyen S. (eds) *Equilibrium and Advanced Transportation Modelling*. Centre for Research on Transportation. Springer, Boston, MA.
19. Djukic, T., Van Lint, J. and S. Hoogendoorn. (2012) Application of principal component analysis to predict dynamic origin-destination matrices. *Transportation Research Record,* vol. 2283.
20. Cascetta, E., Papola, A., Marzano, V., Simonelli, F. and I. Vitiello. (2013) Quasi-dynamic estimation of o–d flows from traffic counts: Formulation, statistical validation and performance analysis on real data, *Transportation Research Part B: Methodological*, vol. 55, pp. 171–187, Sep. 2013.
21. Cantelmo, G., Viti, F., Tampère, C.M.J., Cipriani, E., Nigro, M. (2014) Two-step approach for the correction of seed matrix in dynamic demand estimation. *Transportation Research Record* 2466, 125-133.
22. Cantelmo, G., Viti, F., Cipriani, E. and M. Nigro (2015) A Two-Steps Dynamic Demand Estimation Approach Sequentially Adjusting Generations and Distributions. In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 1477–1482.
23. Ashok, K. and M. E. Ben-Akiva (2002) Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows, *Transp. Sci.*, 36(2), pp. 184–198.
24. Spall, J. C. (2012) Stochastic Optimization. In: *Handbook of Computational Statistics*, J. E. Gentle, W. K. Härdle, and Y. Mori, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 173–201.
25. Frederix, R., Viti, F., Corthout, R. and C. M. J. Tampère (2011) New gradient approximation method for dynamic origin-destination matrix estimation on congested networks, *Transp. Res. Rec.*, no. 2263, pp. 19–25, 2011.
26. Cipriani, E., Florian, M., Mahut, M. and M. Nigro (2011) A gradient approximation approach for adjusting temporal origin-destination matrices, *Transportation Research Part C: Emerging Technologies*, 19(2), pp. 270–282.
27. Antoniou, C., Lima Azevedo, C., Lu, L., Pereira, F. and M. Ben-Akiva (2015) W-SPSA in practice: Approximation of weight matrices and calibration of traffic simulation models, *Transp. Res. Part C Emerg. Technol.*, vol. 59, pp. 129–146, Oct. 2015.
28. Tympakianaki, A., Koutsopoulos, H. N. and E. Jenelius. (2015) c-SPSA: Cluster-wise simultaneous perturbation stochastic approximation algorithm and its application to dynamic origin–destination matrix estimation, *Transp. Res. Part C Emerg. Technol.*, vol. 55, pp. 231–245.
29. Qurashi, M., Ma, T., Chaniotakis, E., Antoniou, C. (2020) PC–SPSA: Employing Dimensionality Reduction to Limit SPSA Search Noise in DTA Model Calibration. *IEEE Transactions on ITS*, 21(4), pp. 1635-1645.
30. Derrmann, T., Frank, R., Engel, T., Viti, F. (2017). How mobile handovers reflect urban mobility: a simulation study. In: *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2017*, pp. 486-491.
31. Cantelmo, G., Viti, F., Derrmann, T. (2017). Effectiveness of the two-step dynamic demand estimation model on large networks. In: *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2017*, pp. 356-361.
32. Cantelmo, G., Viti, F., Cipriani, E., Nigro, M. (2018) A utility-based dynamic demand estimation model that explicitly accounts for activity scheduling and duration". *Transportation Research Part A: Policy and Practice*, Vol. 114, pp. 303-320.
33. Cipriani, E., Del Giudice, A., Nigro, M., Viti, F., Cantelmo, G. (2015). The impact of route choice modeling on dynamic OD estimation. *IEEE Conference on Intelligent Transportation Systems, Proceedings*, ITSC2015, pp. 1483-1488.