EasyChair Preprint

№ 87

A Big Data Demand Estimation Framework for Modelling Urban Congested Networks

Francesco Viti and Guido Cantelmo

April 23, 2018

A Big Data Demand Estimation Framework for Modelling Urban Congested Networks

Guido Cantelmo[1] and Francesco Viti[1] [0000-0003-1803-4527]

[1] University of Luxembourg, L-4364 Esch-Sur-Alzette, Luxembourg
{guido.cantelmo;francesco.viti}@uni.lu

Abstract. This paper deals with the problem of estimating daily mobility flows using different sources of data, and in particular from mobile devices, such as mobile phones and floating car data. We show how mobile phone data can be used to better estimate the structure of the demand matrix, both temporally (i.e. the daily generated flows from each zone) and spatially (i.e. distributing the flows on the different OD pairs). Then, floating car data together with traffic counts can be used to further distribute the demand on the available modes and routes. During this phase, a behavioral modelling approach is used, according to traditional dynamic user equilibrium using a joint route and departure time choice model. Floating car data information is used to estimate speed profiles at all links where information is available, and for route travel times, which feed the utility-based models. A two-step approach is then proposed to solve the problem for large scale networks, in which the total demand is first generated, and then equilibrium is calculated through a dynamic traffic assignment model. The effectiveness and reliability of the proposed modelling framework is shown on a realistic case study involving the road network of Luxembourg City and its surroundings, and is compared to the traditional bi-level formulation solved using the Generalized Least Square (GLS) Estimation. The comparison shows how the two-step approach is more robust in generating realistic daily OD flows, and in exploiting the information collected from mobile sensors.

Keywords: Dynamic OD Estimation, Big Mobility Data, Two-steps approach.

1 Introduction

1.1 The demand estimation problem

Dynamic traffic models represent essential tools for assessing properties of robustness and resilience, and for managing transportation networks. These models take as input the demand from each origin and destination and at each time period, and in turn estimate and/or predict route and link flows and travel times.

In order to generate the mobility demand, usually represented in the form of Origin-Destination (OD) matrices, traditional approaches combine survey data and mathematical tools [1]. 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 [2]. Unfortunately, the estimated demand matrix is at most a coarse representation of the systematic component of the demand – such as the typical behavior during a working day. However, daily demand patterns can substantially differ from the systematic ones because of several elements, including weather conditions or road works, as well as because of the inherent stochasticity of the travelers' choices. Deviations between estimated and actual demand patterns can be mitigated by using traffic data, which can be used to update an existing (a-priori) OD matrix. 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 [3-4], the last decades have witnessed to a considerable effort to develop methodologies able to deal with within-day dynamics in order to apply them on (real-time) dynamic traffic management contexts. By limiting ourselves to the widely adopted bi-level optimization formulation, in the upper level, OD flows are updated by minimizing 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 [5]. In order to overcome solution under-determinedness, Balakrishna et al. [6] suggested to use a simulation-based DTA model to generate traffic measures and to include additional information, such as link speed, within the objective function, in order to represent the congested/uncongested network conditions. Following this seminal work, many researchers developed new and more robust algorithms able to properly capture the non-linearity between link-flow propagation and time-varying OD demand [7-10]. Despite this intense effort, the resulting optimization problem remains highly non-linear and non-convex. Hence, the easiest solution 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) [11]. Alternatively, Cascetta et al. [12] 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 constant. Under this assumption the DODE problem is likely to find more robust results. Nevertheless, the authors point out that the resulting matrix will be "intrinsically biased" since this assumption introduces an "intrinsic error" in the spatial distribution of the demand patterns. To partly solve this issue, Cantelmo et al. [13] proposed a generic 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 paper, we show how the Two-Steps approach can be effectively applied in combination with a joint route and departure time choice model to reduce the complexity of the OD estimation problem. Furthermore, by separating the estimation process into a first step that aims at estimating the total number of trips generated by a certain zone, while the second step focuses on the spatial and temporal distribution of the OD flows, we can 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. The next section will only briefly introduces the Two-Step approach. An interested reader can find more details in [13-14].

The Two-Step approach has three characteristics that make it an ideal candidate for applications on large-scale networks. First, as pointed out by Antoniou et al. [15], the starting matrix is still a key input for all state-of-the-art DODE models. The first step of this formulation focuses on improving the historical demand matrix by performing a broad evaluation of the solution space and estimating a "better" updated seed matrix to be used in the second step. Secondly, the proposed model reduces the number of variables in the first step, increasing the overall reliability of the results [14, 16]. On this point, the idea of performing successive iterations and linearizations has been already introduced and validated in [4] for the online DODE, showing that the reliability of the results generally increases.

Driven by these considerations, the contribution of this paper is twofold. First, we show how different big mobility data can be used in a novel estimation framework. Then we apply the new approach to the real network of Luxembourg. The test-network represents most of the country of Luxembourg, including urban roads, motorways and primary roads. Mobile phone antenna density data provided by the largest operator in Luxembourg, Post, is used to create a temporal profile of the demand in and out of Luxembourg City. In addition, real traffic counts extracted from loop detectors are used within the calibration process to further update the demand. Second, as speed profiles on the counting stations were not available, we extend the objective function by including the average speeds over the analysis period, which have been calculated through Floating Car Data (FCD). We show that, when combined with a standard DODE procedure, this information leads to a poor calibration of the demand, as the DODE overfits the data within the objective function. However, as the Two-Step approach over-imposes a linear relation between distribution and generation for a certain traffic zone, it is more likely to capture congestion dynamics at network level, such as the systematic overestimation or underestimating of the demand, thus to avoid this issue.

2 Methodology

2.1 The Two-Steps approach

While for a detailed overview of this model we refer to [13-14], in this section we briefly present its main characteristics.

In the proposed Two-Step procedure, the first step focuses on optimising the generation values of each zone in each time interval, while keeping constant the trip distributions. To achieve this goal, the objective function can be generally written as:

$$\left(\mathbf{E}_1^*,\ldots,\mathbf{E}_n^*\right)=\arg\min\begin{bmatrix} z_1(\mathbf{l}_1,\ldots,\mathbf{l}_{n'},\hat{\mathbf{l}}_1,\ldots,\hat{\mathbf{l}}_{n'}) \\ +z_2(\mathbf{n}_1,\ldots,\mathbf{n}_{n'},\hat{\mathbf{n}}_1,\ldots,\hat{\mathbf{n}}_{n'}) \\ +z_3(\mathbf{x}_1,\ldots,\mathbf{x}_{n'},\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_{n'}) \\ +z_4(\mathbf{r}_1,\ldots,\mathbf{r}_{n'},\hat{\mathbf{r}}_1,\ldots,\hat{\mathbf{r}}_{n'}) \end{bmatrix} \tag{1a}$$

$$\text{s.t.} \qquad x_n^{OD} = E_n^O d_{D|O}^{Seed,n} \qquad \forall O, \forall D, \forall n \tag{1b}$$

Where $E_n{}^O$ is the generation factor of origin zone O at time interval n, $\boldsymbol{E}_n{}^*$ is the generation vector from all origins in time interval n, $X_n{}^*$ is the number of trips originated in O with destination D in time interval n and $d_{D|O}^{Seed,n}$ is the matrix probability distribution between traffic zone D and traffic zone O in time interval n.

2.2 Including mobile network data in the first step:

While the correlation between traffic demand and mobile phone data is well known [17-18], this source of 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 count active connections that are entering or exiting the zones (i.e. the number of *handovers*). Unfortunately, mobile network data is subject to intrinsic errors such as the split of the users between multiple network operators and the degree of activity on the network as well as the general mobile penetration rates. However, this information can be used to estimate the temporal profiles of the generated demand on a certain cluster, as shown in [19]. In [20] 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.

2.3 Including floating car data in the second step:

To consider the relation between the temporal characteristics of road congestion and their impact on the spatial and temporal distribution of the OD flows, a departure time choice model based on the Vickrey/Small [21] formulation has been adopted. 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 e.g. in [22].

3 Case study

3.1 Estimating the generated demand patterns

We show the application of the Two-Step approach on the road network of Luxembourg. The network consists of 3700 links and 1469 nodes. In our case study, we created

two different clusters. One cluster captures the trips generated from the city to the external zones, while the other one captures those entering Luxembourg City, as shown in Figure 1.

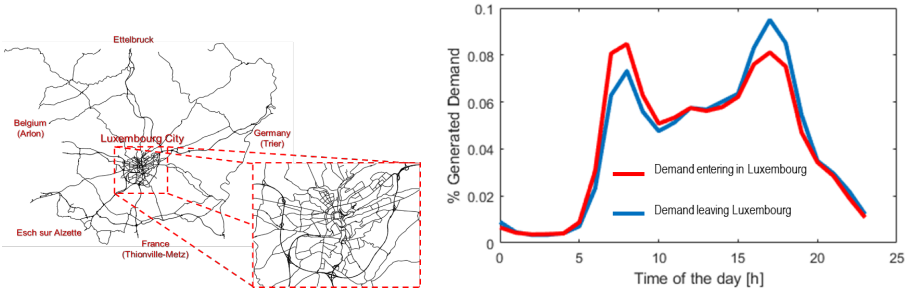

Fig. 1. Road Network of Luxemborg City (left); Antenna densities in and out of the city (right).

This procedure can be easily extended to any urban area, in which mobile connection handovers can be used to calculate the flows exchanged between the study area and the external centroids. Although the profile showed in Figure 1b looks realistic, we do believe that to simply include the emission flows within the goal function may still lead to a biased estimation, since it is equivalent to over-imposing a certain time-dependent profile to the demand. Instead, we propose to use the difference between entering and exiting flows. We use then this information to calibrate the departure time choice model in (1b).

3.2 Estimating the spatial and temporal distribution of OD flows

In this part of the study, we consider the morning peak between 5 AM and noon (8 hours). After some data cleaning, 54 counting stations have been retained, all located on the main arterial roads going to Luxembourg City and on the ring. Unfortunately, these data present two major limitations. The first is that, based on the publicly available data, only three detectors are located inside the ring of Luxembourg. This means that we can expect to have a realistic representation of the demand on the regional network and on the ring, but it is not possible to validate the estimated solution inside the city. The second concerns the time interval aggregation for these data, as traffic counts are aggregated on an hourly basis. This time interval is clearly too large for a network with an average free-flow travel time of 20 minutes since basic congestion dynamics could not be properly captured.

To deal with this lack of information, the company Motion-S provided us average speeds on the ring of Luxembourg for each time interval from Floating Car Data (FCD). The obtained information is based on the average of all available information and does not contain specifications about time and location. Thus, the available average speed broadly captures, in this study, the congestion on the ring-way 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. However, as this information is strongly aggregated, the Single-Step approach has the tendency to over-fit the average speed, while the Two-Step approach manages to provide more reliable results by exploiting the link flows as a constraint within the objective function. This claim is numerically illustrated in the next section.

3.3 Presentation of results

To solve the DODE on the network of Luxembourg, we developed a Matlab package using PTV Visum. The package allows performing assignment-free dynamic or static OD estimation, using a deterministic and/or stochastic approximation of the gradient. While this package has been designed for Luxembourg, it can work with any network in Visum, supporting the idea that the model is ready for practical implementation.

The DODE was solved using both the classical bi-level formulation (referred to as Single-Step, SS) and with the Two-Steps (TS) approach. In both cases, the Simultaneous Perturbation Stochastic Approximation (SPSA) was adopted for the optimisation. In order to reduce the computational time, we adopted the one-sided version of this model. The interested reader can refer to [13] for more details on the solution algorithm. We performed three different sets of experiments: 1) only traffic counts are included within the OF, 2) traffic counts and mobile data are included within the OF, and 3) Using FCD and Traffic counts.

As shown in Fig. 2 (left), results confirm that, when the number of variables is large, the SS model performs a quite 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 [23]. Finally, 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.

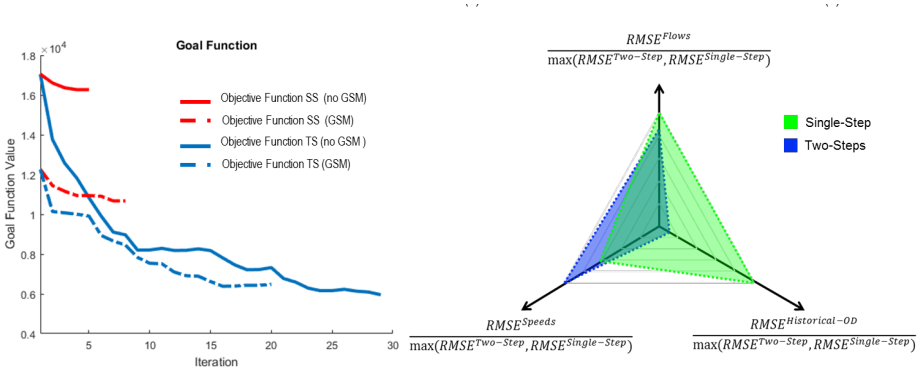

Fig. 2. Performance of the Two-Steps approach with GSM as compared to Single-Step

As shown in Fig. 2 (right), the Two-Steps approach provides a realistic fit for both traffic counts and speeds, although the error on the link flows increases with respect to the starting point. This is because constraint (2b) imposes a constant spatio-temporal structure of the demand remains during the first step of the optimisation. By contrast,

the Single-Step approach over fits the data by strongly changing the original structure of the demand.

4 Conclusions

This paper showed how big mobility data such as mobile phones and probe 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. As reported in [15], 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 Single-Step. Then, we show that by using floating car data and link flows on the second step, the model is capable of 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 (in Luxembourg a real time information system using Automatic Vehicle Location data is currently been tested).

Acknowledgements

The authors acknowledge the FNR for providing the financing grant: AFR-PhD grant 6947587 IDEAS. We also like to thank Motion-S Luxembourg for providing Speed data.

References

1. M. G. McNally, "The Four-Step Model," in *Handbook of Transport Modelling*, vol. 1, 0 vols., Emerald Group Publishing Limited, 2007, pp. 35–53.
2. J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González, "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.
3. E. Cascetta, D. Inaudi, and G. Marquis, "Dynamic Estimators of Origin-Destination Matrices Using Traffic Counts," *Transp. Sci.*, vol. 27, no. 4, pp. 363–373, Nov. 1993.
4. K. Ashok and M. E. Ben-Akiva, "Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows," *Transp. Sci.*, vol. 36, no. 2, pp. 184–198, 2002.
5. H. Tavana (2001). Internally-consistent estimation of dynamic network origin-destination flows from intelligent transportation systems data using bi-level optimization, PhD Thesis.
6. R. Balakrishna, M. Ben-Akiva, and H. Koutsopoulos, (2007). Offline Calibration of Dynamic Traffic Assignment: Simultaneous Demand-and-Supply Estimation. *Transportation Research Record*, vol. 2003.
7. R. Frederix, F. Viti, R. Corthout, and C. M. J. Tampère, "New gradient approximation method for dynamic origin-destination matrix estimation on congested networks," *Transp. Res. Rec.*, no. 2263, pp. 19–25, 2011.

8. E. Cipriani, M. Florian, M. Mahut, and M. Nigro, “A gradient approximation approach for adjusting temporal origin-destination matrices,” *Transp. Res. Part C Emerg. Technol.*, vol. 19, no. 2, pp. 270–282, 2011.
9. C. Antoniou, C. Lima Azevedo, L. Lu, F. Pereira, and M. Ben-Akiva, “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.
10. A. Tympakianaki, H. N. Koutsopoulos, and E. Jenelius, “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,.
11. T. Djukic, J. Van Lint, and S. Hoogendoorn, Application of principal component analysis to predict dynamic origin-destination matrices. *Transportation Res. Rec.* Vol. 2283, 2012.
12. E. Cascetta, A. Papola, V. Marzano, F. Simonelli, and I. Vitiello, “Quasi-dynamic estimation of o–d flows from traffic counts: Formulation, statistical validation and performance analysis on real data,” *Transp. Res. Part B Methodol.*, vol. 55, pp. 171–187, Sep. 2013.
13. Cantelmo, G., Viti, F., Tampère, C.M.J., Cipriani, E., Nigro, M. Two-step approach for the correction of seed matrix in dynamic demand estimation. *Transp. Res. Rec*. 2466, 125-133.
14. G. Cantelmo, F. Viti, E. Cipriani, and N. Marialisa, A Two-Steps Dynamic Demand Estimation Approach Sequentially Adjusting Generations and Distributions. in 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 2015, pp. 1477–1482.
15. C. Antoniou *et al.*, “Towards a generic benchmarking platform for origin–destination flows estimation/updating algorithms: Design, demonstration and validation,” *Transp. Res. Part C Emerg. Technol.*, vol. 66, pp. 79–98, May 2016.
16. V. Marzano, A. Papola, and F. Simonelli, “Limits and perspectives of effective O–D matrix correction using traffic counts,” *Transp. Res. Part C Emerg. Technol.*, vol. 17, no. 2, pp. 120–132, 2009.
17. J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González, “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.
1. Derrmann T., Frank R., Engel T., Viti F. (2015). 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
18. 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.
19. Cantelmo G., Viti F., Derrmann T. (2015). 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.
20. K. A. Small, “The bottleneck model: An assessment and interpretation,” *Econ. Transp.*, vol. 4, no. 1–2, pp. 110–117, Mar. 2015.
21. 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
22. E. Cipriani, M. Florian, M. Mahut, and M. Nigro, “A gradient approximation approach for adjusting temporal origin-destination matrices,” Transp. Res. Part C Emerg. Technol., vol. 19, no. 2, pp. 270–282, 2011.