Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2017) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

23rd International Symposium on Transportation and Traffic Theory, ISTTT 23, 24-26 July 2019, Lausanne, Switzerland

A data driven method for OD matrix estimation

Panchamy Krishnakumari[a], Hans van Lint[a], Tamara Djukic[b], Oded Cats[a]

[a]*Delft University of Technology, The Netherlands*
[b]*Aimsun SL, Barcelona, Spain*

Abstract

The fundamental challenge of the origin-destination (OD) matrix estimation problem is that it is severely under-determined. In this paper we propose a new data driven OD estimation method for cases where a supply pattern in the form of speeds and flows is available. We show that with these input data, we do not require an iterative dynamic network loading procedure that results in an equilibrium assignment, nor do we need an assumption on the kind of equilibrium that emerges from this process. The minimal number of ingredients which *are* needed are (a) a method to estimate / predict production and attraction time series; (b) a method to compute the *N* shortest paths from each OD zone to the next; and (c) two—possibly OD-specific—assumptions on the magnitude of *N*; and on the proportionality of path flows between these origins and destinations, respectively. The latter constitutes the most important behavioral assumption in our method, which relates to how we assume travelers have chosen their routes between OD pairs. We choose a proportionality factor that is inversely proportional to realized travel time, where we incorporate a penalty for path overlap. For large networks, these ingredients may be insufficient to solve the resulting system of equations. We show how additional constraints can be derived directly from the data by using principal component analysis, with which we exploit the fact that temporal patterns of production and attraction are similar across the network. Experimental results on a toy network and a large city network (Santander, Spain) show that our OD estimation method works satisfactorily, given a reasonable choice of *N*, and the use of so-called 3D supply patterns, which provide a compact representation of the supply dynamics over the entire network. Inclusion of topological information makes the method scalable both in terms of network size and for different topologies. Although we use a neural network to predict production and attraction in our experiments (which implies ground-truth OD data were needed), there are straight-forward paths to improve the method using additional data, such as demographic data, household survey data, social media and or movement traces, which could support estimating such ground-truth baseline production and attraction patterns. The proposed framework would fit very nicely in an online traffic modeling and control framework, and we see many paths to further refine and improve the method.

Keywords: traffic demand; origin-destination matrix; OD matrix estimation; 3D supply patterns; data driven; principal component analysis; PCA

* Corresponding author. Tel.: +31-6-48875624
E-mail address: j.w.c.vanlint@tudelft.nl

1. Introduction

Understanding the dynamics of traffic demand over space (from origins to destinations) and time is quintessential for many applications over the entire transportation domain, from operations, control and management; to planning and policy assessment. Since we do not directly observe where everyone is coming from and going to at all times (yet[1]), we have to estimate time-dependent origin-destination (OD) flows $\mathbf{x}_k$ from whatever data and information *are* available. As it stands, the OD estimation problem is (still) one of the toughest problems in transportation to date. The most important difficulty in estimating OD matrices is that, particularly for large congested networks, the problem is severely under-determined, a fact that was already recognized in the early days of the OD estimation literature (e.g. Van Zuylen and Willumsen (1980); Cascetta (1984); Bell (1991)) and is emphasized in virtually all contemporary OD estimation research as well. To put this problem in a practical perspective, the OD matrix of the planning model for the Western part of the Randstad (The Netherlands) contains over two million OD-pairs, whereas literally orders of magnitudes fewer independent equations can be formulated to constrain these OD flows with actual observations, such as OD flow samples, links counts and speeds, that can be related to the OD matrix. Evidently, when we consider dynamics (e.g. with time steps $t + k\Delta t, k = 1, 2, \ldots$, and typically $\Delta t = 15, 30$ or 60 minutes) the situation gets worse. Consequently, solving the OD estimation problem in large networks requires making a large number of assumptions.

There are essentially two paths to estimating OD matrices. One is a "forward path" in which demand is estimated (predicted) using models and data that link activity patterns, land use and zone production and attraction potential to the resulting OD flows. This encompasses a broad research area with modeling approaches ranging from estimating production and attraction potential of zones in an aggregated way to highly detailed disaggregate activity based models (e.g. Kitamura et al. (2000); Bhat and Zhao (2002); Scheffer et al. (2017); Cantelmo et al. (2015); Arentze and Timmermans (2009)). The second path is to "reversely engineer" OD matrices by assimilating a wide variety of data sources into models that describe the assignment of OD matrices onto actual service and infrastructure networks and the resulting route choice and network traffic conditions. This is the common connotation of the term OD estimation and there is a long record of contributions in this area as well (e.g. Van Zuylen and Willumsen (1980); Cascetta (1984); Cremer and Keller (1987); Bell (1991); Yang et al. (1992); Ashok and Ben-Akiva (2000); Zhou and Mahmassani (2007); Castillo et al. (2012); Cascetta et al. (2013); Cantelmo et al. (2014); Cipriani et al. (2014); Antoniou et al. (2016) to name a few in chronological order). Either way, in this data assimilation process, three types of assumptions play a role. First, there are assumptions about the quality and semantics of the available data (the observations, the evidence). Second, a large number of assumptions are incorporated in quantitative models describing the entire chain from activities to traffic flows, which govern how the data presumably relates in mathematical terms to OD flows. These are typically behavioral assumptions and assumptions related to how we represent and abstract transport and traffic flows (micro, meso, macro). Third, there are assumptions involved in the assimilation process itself, and in the supporting algorithms. This assimilation process may have the form of an off-line (bi-level) minimisation or maximisation problem (e.g. Cascetta et al. (2013); Lundgren and Peterson (2008); Cascetta and Postorino (2001)), or of a sequential (recursive) estimation process (e.g. Okutani and Stephanedes (1984); Van Der Zijpp (1997); Zhou and Mahmassani (2007); Djukic et al. (2012a); Carrese et al. (2017)). Below we briefly overview the three categories of assumptions outlined above and how they—in general terms—affect OD estimation quality. We then use this review to motivate the approach we choose in this paper, which we outline in subsequent sections. Note that the scope in this paper is car traffic, and, in terms of the modeling literature we discuss, the "reverse engineering" path to OD estimation.

1.1. OD estimation assumptions: observations

Let us first distinguish between observations that can be directly (i.e. in closed form) related to specific OD flows, and those that can not. The first category of such direct observations are sampled OD matrices obtained from surveys (e.g. Cascetta (1984); Bierlaire and Toint (1995)) or travel diaries (e.g. Scheffer et al. (2017)). Increasingly, slightly larger samples can be obtained through vehicle (re)identification systems (e.g. Kim et al. (2014); Camus et al. (1997); Barcelo et al. (2010); Zhou and Mahmassani (2006)) or data from mobile phone or GPS based movement traces (e.g.

[1] unfortunately from a scientific perspective, but fortunately from a societal perspective

Ge and Fukuda (2016); Alexander et al. (2015); Zin et al. (2018); Gadzinski (2018); Nigro et al. (2018)). Such OD samples $\mathbf{x}_k^{obs}$ are then typically used to construct a prior OD matrix, e.g. $\tilde{\mathbf{x}}_k = \beta\mathbf{x}_k^{obs}$, with β a possibly OD flow specific scaling factor. The standard way of exploiting the information in what are called prior OD matrices is to assume that $\tilde{\mathbf{x}}_k$ is informative of the spatial (in the dynamic case: spatiotemporal) structure of the (unknown) OD matrix $\mathbf{x}_k$. The widely used (static) formulation of the OD estimation problem (e.g. Cascetta and Marquis (1993))

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} f_1(\mathbf{x}, \tilde{\mathbf{x}}) + f_2(\mathbf{y}(\mathbf{x}), \tilde{\mathbf{y}}), \tag{1}$$

exploits this assumption by adding a distance function $f_1(\mathbf{x}, \tilde{\mathbf{x}})$ to the overall objective function, that gently forces the OD estimation solution to minimize the distance between the final estimate $\mathbf{x}$ and the prior matrix $\tilde{\mathbf{x}}$—the second term f_2 penalizes errors on reconstructing the observations $\tilde{\mathbf{y}}$ using the OD estimate, which we discuss below. Clearly, a proper choice of the distance function f_1 is crucial. Standard distance metrics (L2 norm or RMSE) may not steer the estimation in a direction that favors similarity in spatiotemporal structure, for which other metrics such as the structure similarity index (SSI) may be more appropriate (Djukic et al., 2013). Regardless of how f_1 is formulated, the validity of the underlying assumption (that $\tilde{\mathbf{x}}$ has a similar spatiotemporal structure as $\mathbf{x}$) depends on the representativity of the sample for the population, i.e. whether the traveler traits (personal characteristics, income, car ownership, preferences, motives, etc.) and the corresponding destination choices in the sample represent a balanced cross section of the population (Tolouei et al., 2017). With a large enough sample, survey and diary data collection efforts, can be—to a degree—controlled for representativity. This is more difficult in the case of movement traces, which typically come (to the researcher) without these contextual and privacy sensitive data (Ge and Fukuda, 2016). There is a clear and large potential of using both active and passive mobile data for inferring activities, origins, destinations, paths and modes from such movement traces (e.g. Chen et al. (2014); Widhalm et al. (2015); Castro et al. (2013); Yang et al. (2017)). Cheng et al. (2014) convincingly show that mobile traces may improve determining activity patterns, however, it is unclear whether, and under which circumstances, the (much) larger sample sizes obtained with these data *indeed* eradicate a possible representativity bias in the resulting (prior) OD matrix, which would invalidate the assumption that the spatiotemporal structure of a prior OD matrix is fully informative for the population OD matrix.

Under two very strict assumptions, also traffic counts (at cross sections in the network) can be directly related to specific OD flows.

$$y_{d\ell} = \sum_{i,j,k} a_{ijk}^{d\ell} x_{ijk}, \tag{2}$$

where $a_{ijk}^{d\ell}$ depicts an assignment matrix that maps the portion of each OD flow x_{ijk} between i and j departing at period k to the count location d in period ℓ. The fundamental problem is that counts in or downstream of congestion are not informative of demand, but of (discharge) capacity (Frederix et al., 2012). More precisely, equation (2) is valid only and only if (a) $y_{d\ell}$ is observed in uncongested conditions; and (b) no OD path flow x_{ijk}^n experienced a saturated bottleneck upstream before passing cross section d. These assumptions are typically violated in even mildly congested networks. Under such circumstances the counts have no direct linear relation to specific OD flows. Nonetheless, they are still part of an emerging traffic pattern that may be typical for a particular spatiotemporal OD pattern. As a result, there may exist a complex and nonlinear relation between those counts and the overall spatiotemporal OD pattern or perhaps a specific subset of OD flows. The same can be said for many other observations such as densities, speeds, travel times, realized paths, etc. Also these are (highly) non-linearly related to the prevailing dynamic OD pattern or a specific subset of OD flows, however, the form of this relationship is way more complex than a simple linear sum as in equation (2). A possible alternative interpretation is that $a_{ijk}^{d\ell}$ in (2) depicts a probability (Yang et al., 2017) that is governed by a nonlinear process that maps the OD flows to the network. This naturally brings us to modeling assumptions.

1.2. OD estimation assumptions: modeling

The unobserved OD pattern $\mathbf{x}_k$, in combination with the available network (coded in a graph G); prevailing traffic control $\mathbf{u}_k$; route-choice and driving behavior of travelers (ϕ_k, and θ_k); and the external conditions ω_k (e.g. weather, information) affecting these, ultimately result in an (at least to a degree) observable traffic supply pattern $\mathbf{y}_k$, in terms of flows, speeds, travel times, etc. The natural way to model this complex and non-linear mapping $\mathbf{x}_k \rightarrow \mathbf{y}_k : \mathcal{R}^n \rightarrow \mathcal{R}^m$ from OD pattern to supply pattern is to use a traffic simulation model of the form

$$\mathbf{y}_k^{sim} = \mathcal{A}(\mathbf{x}_k, \mathbf{u}_k, \hat{\phi}_k, \hat{\theta}_k, \omega_k) \tag{3}$$

in which $\mathcal{A}$ may represent any micro-, meso, or macroscopic traffic simulation model, and $\hat{\phi}_k, \hat{\theta}_k$ represent assumptions (mechanisms and estimated parameters) related to route choice and driving behaviour. Clearly, the linear mapping between counts and OD flows in equation (2) is a special case of (3), which is valid only, and only if, route choice is fixed and known, and the network is uncongested (i.e. has no capacity constraints), which means that under most conditions more complex simulation models are required. In many cases where such a simulation model is used, an additional assumption is added, which pertains to the existence of a (stochastic) equilibrium state. Whether and under which circumstances (for which purposes) the notion of equilibrium is a valid assumption is highly contested, due to the many open research questions related to the mechanisms with which travelers make choices and how rational these really are (Di and Liu, 2016). To the best of our knowledge, no conclusive empirical evidence has been reported that supports the existence of particular types of equilibria in actual large traffic networks.

As mentioned above, the standard approach to solve the OD estimation problem is to use the same DTA machinery to reversely engineer dynamic OD patterns from traffic supply data in (2), and by extension, *to use the same large amount of assumptions*, e.g. the notion of equilibrium (or not); the assumptions specified in $\hat{\phi}_k, \hat{\theta}_k$; and the assumption that all circumstances ω_k, and control $\mathbf{u}_k$ affecting route choice and driving behavior are known. In this reverse estimation problem, then iteratively an OD pattern is sought that is consistent with (a) the observed data; and (b) all the (behavioral) assumptions discussed here. Finally, there are also (necessary) assumptions pertaining to the data assimilation methodology, that may affect OD estimation quality.

1.3. OD estimation assumptions: solution algorithms

In illustration, we discuss sequential methods. Okutani was one of the first to formulate the time-dependent OD estimation problem (Okutani and Stephanedes, 1984) in state space form such that it can be recursively solved by a Kalman filter (KF), and many authors have followed along a similar path (e.g. Van Der Zijpp (1997); Ashok and Ben-Akiva (2002); Zhou and Mahmassani (2006); Antoniou et al. (2006); Djukic et al. (2012a); Carrese et al. (2017); Lu et al. (2015); Cascetta et al. (2013); Barcelo et al. (2013); Marzano et al. (2018)). The central idea of these approaches is a formulation in state-space form, that is,

$$\mathbf{x}_{k+1} = F\mathbf{x}_k + \mathbf{w}_k, \tag{4}$$

$$\mathbf{y}_k = A\mathbf{x}_k + \mathbf{v}_k, \tag{5}$$

in which the dynamic equation (4) represents a linear (autoregressive) process driven by matrix F, and observation equation (5) refers to the assignment mapping (as in (2)), in which A depicts the assignment matrix. The appeal of the KF approach is that it offers a sequential estimation approach that is optimal in terms of error variance. However, the catch lies in the noise terms $\mathbf{w}_k$ and $\mathbf{v}_k$, which, to guarantee optimality, are assumed to be drawn from independent zero-mean Gaussian noise processes with known co-variance matrices. Put differently, the KF is optimal if both process and observation models are linear and unbiased and the errors they make come from known Gaussian white noise processes. Clearly, these assumptions are typically violated for numerous reasons (biased models, non-negativity flows, autocorrelation in the errors). As a result, many of the "KF approaches" to OD estimation propose ideas to either reformulate the problem, so that a KF approach is better justified—e.g. Ashok and Ben-Akiva (2000, 2002) propose an OD state vector that consists of deviations between estimates and prior OD flows; Djukic et al. (2012a) also propose a reduced state vector (obtained through PCA) and propose a "coloured KF" to address autocorrelated errors. Alternatively, more advanced data assimilation methods are proposed to relax some of the highly restrictive assumptions, e.g. Carrese et al. (2017) propose a variation on the ensemble Kalman filter to alleviate the need for an explicit assignment matrix. In conclusion, whatever the (online or offline) solution methodology chosen, implicit and explicit data assimilation assumptions are necessary to solve the OD estimation problem, and the literature offers many great ideas to this end.

1.4. Motivation and rationale of a new approach

From the brief overview above, it is clear that to solve a severely underdetermined problem like OD estimation, one has to make many assumptions related to using the data, the models that relate these data to the unobserved OD flows, and the assimilation / solution methodologies to derive the OD flows from the combination of those data and models. However, it is not self-evident how many of those assumptions we really need. What is evident, is that the available data (the evidence) should dictate the amount of assumptions needed, and that we should choose methods that are parsimonious, since we generally lack the information to even estimate the magnitude and direction of the possible biases caused by (unnecessary) assumptions.

Now consider the case in which we have speeds either on each link, or in the form of so-called 3D supply patterns—speed averages over regions of the network over space and time (we return to this in the next section). In this case path travel times and consequently also all (or a selection of say N) shortest paths can be derived between any given OD pair. This implies we do not need an iterative network loading process to compute travel times over the available paths between origins and destinations, since we already have those travel times. It doesn't matter for the OD estimation task whether this pattern of path travel times represents an equilibrium state, and what the precise nature of this equilibrium is. What matters is that these were the actual realized travel times, given the OD matrix we seek to find. To do so, we propose an idea that follows the same rationale as Cascetta et al. (2013), in which "*an OD estimator is proposed based on the assumption of constant distribution shares across larger time horizons with respect to the within-day variation of the production profiles, leading to an estimator which improves dramatically the unknowns/equations ratio*". We (can) go a step further, since we have available *all* realized travel times over *all* (shortest) paths. We will show in this paper that the assumptions we need to specify are (a) how many of the shortest paths were actually used for each OD pair; and (b) the proportions of each OD flows over these (N) shortest paths. The latter *is*, of course, a behavioral assumption, but at the macroscopic scale (a path flow proportion), and not in the form of a detailed route choice model with (elaborate) trade offs.

However, these two assumptions about the distribution of traffic over the network are not sufficient to estimate the underlying OD matrix, volumes are needed as well. To this end, we can use flow counts on those locations that meet the two strict requirements mentioned under equation (2). In the remainder of this paper we will show that given production P_{ik} (sum of all outgoing OD flows of zone i during period k) and attraction A_{jk} (sum of all incoming OD flows of zone j during period k) are observable, the two distributional assumptions and counts are sufficient to reliably estimate the full OD matrix in smaller networks. For larger networks, however, we require additional constraints. We will show these can be derived *directly* from the data by using principal component analysis (PCA). Through PCA, we are able to find upper bounds for the so-called "dominant" OD flows, i.e. those OD flows that exhibit the largest dynamics. With these constraints then finally, also the underlying complete OD matrix can be estimated.

1.5. Paper outline

To this end, section 2.1 describes the overall estimation framework; section 2.2 the basic estimation logic from production and attraction totals to OD flows; section 2.3 the PCA method to constrain the solution space for large networks; and section 2.4 the methodology to estimate the attraction and production totals from so-called 3D supply patterns. In section 3 we outline how we assess the method. We do this on two networks: a small network with which we vary extensively with the assumptions in our method; and a larger network to demonstrate the feasibility. These results are presented in section 4. We offer conclusions and a discussion on further research avenues in section 5.

2. Methodology

For convenient reference, the notation used for recurrent variables in the methodology is first presented as follows:

x^n_{ijk}	path flows between $v_i, v_j \in Z$ for travelers departing in period k with $n = 1, \ldots, N_{ijk}$ paths
x_{ijk}	OD flows between $v_i, v_j \in Z$ for travelers departing in period k
P_{ik}	production (sum of all outgoing OD flows) of zone i during period k
A_{jk}	attraction (sum of all incoming OD flows) of zone j during period k
TT^n_{ijk}	travel time for vehicles traversing path n between node i, j departing in period k
β^n_{ijk}	route proportion
y^m_k	link count for link e_m in period k
FV	speed-flow based (SF) feature vector
$\hat{FV}$	speed-flow-topology (SFT) based feature vector

2.1. Framework: OD estimation with minimal assumptions

Figure 1 outlines the main components of our framework. We use two data sources: 3D supply patterns—either in the form of the underlying "raw" link flows and speeds, or in the form of a condensed 3D consensual pattern (Lopez et al., 2017b). The method allows for inclusion of additional data sources that provide more evidence for either the production and attraction patterns, or for the resulting OD matrices and/or path flows. The overall logic now is as follows. Assume that for a given set of zones we can estimate / predict production and attraction time series using the 3D speed and flow patterns, possibly augmented with other data sources and/or models. Figure 1(a) indicates that in this paper we will use machine learning techniques for this purpose, but any other method would suffice as well. With the 3D *speed* patterns we also compute N^* weighted (by travel time) shortest paths (Figure 1(b) and (c)), where N^* is an assumption that reflects our belief into how many route alternatives—on average—were used for each OD pair. This assumption is likely topology and circumstance dependent. The second assumption we make (Figure 1(e)) is that the proportion of each OD flow on each of the N^* paths is inversely proportional to the (realized) travel time along that path, where we do take path overlap into account. We finally constrain the path flow solution space by using all the *admissable* link counts (Figure 1(d))—this will be elaborated further below. We explain our methodology in three parts: (I) the basic rationale of the method; (II) a dimension reduction technique (PCA) required to make the approach scalable for large networks (Figure 1(f)); (III) an example approach to infer production and attraction time series from 3D supply patterns using machine learning.

2.2. Part I: from production and attraction patterns to OD matrices

Consider a directed graph $G(V, E)$ with nodes (vertices) $v_i \in V, i = 1, \ldots, N_v$ and links (edges) $e_m \in E, m = 1, \ldots, N_m$. The set $Z \subset V$ describes the N_Z origin and destination zones in this network, and a dynamic OD matrix x_{ijk} describes the OD flows between $v_i, v_j \in Z$ for travelers departing in period k. These OD flows are distributed over a set P_{ijk} of N_{ijk} paths $p^n_{ijk} = \{\ldots, e_m, \ldots\}$ resulting in path flows x^n_{ijk}, with $n = 1, \ldots, N_{ijk}$. Our OD estimation method is based on utilizing three data sources; these are: production and attraction per zone; realized travel times on all routes; and (a limited number of) link counts. For now assume that we can either directly measure or estimate the production (sum of all outgoing OD flows) of zones i and the attraction (sum of all incoming OD flows) of zones j during period k (in section 2.4 we outline a machine learning method to do so). These data then add two constraints to the path flows

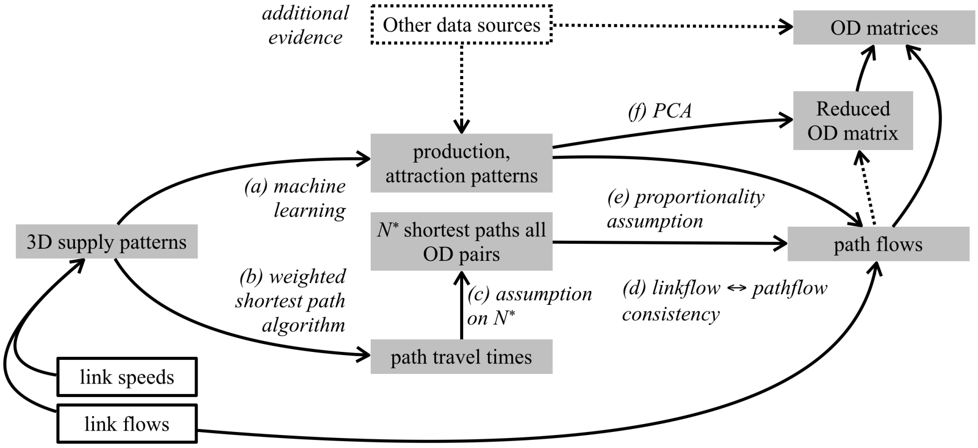


Fig. 1: Framework of the method

$$P_{ik} = \sum_j \sum_n x^n_{ijk} \tag{6}$$

$$A_{jk} = \sum_i \sum_n x^n_{ijk} \tag{7}$$

Secondly, consider we have available realized or estimated link speeds with which we can determine for every path n the travel time TT^n_{ijk} for vehicles traversing path n between node i, j departing in period k. Note that we define the travel time for vehicles traversing path n between node i, j arriving in period k by $\tilde{TT}^n_{ijk}$. We employ the conventional assumption that the distribution of OD flows over these paths follows the Logit model using the realized travel times as the key explanatory variable, that is,

$$\beta^n_{ijk} = \frac{e^{TT^n_{ijk}(1-PS^n)}}{\sum_n e^{TT^n_{ijk}(1-PS^n)}} \tag{8}$$

where β^n_{ijk} is the route proportion; and PS^n the path size factor defined as follows (Ben-Akiva and Bierlaire, 1999):

$$PS^n = \sum_a \left(\frac{l_a}{L^n}\right) \frac{1}{\sum_n \delta^n_a} \tag{9}$$

Where l_a is the length of link a, L^n is the total length of path n and δ^n_a is the link-path incidence variable which equals one if link a is on path n and zero otherwise. The path size factor inclusion in Eq.(8) ensures that paths are penalized based on the extent to which they overlap with other paths included in the set of considered paths. This formulation of the flow distribution assumption is inspired by the path size logit (PSL) model for determining individual route choice probabilities. Accounting for the effect of path overlap is especially important in the case of larger (and particularly in grid or triangular) networks, where the number of available paths between zones i, j increases exponentially. The key idea of our approach is that we cut off the number of considered paths, and compute the path proportions in (8) for a set P^*_{ij} of $N^*_{ijk} \leq N_{ijk}$ shortest paths. This implies we now have an approximation of the OD matrix that reads

$$\hat{x}_{ijk} = \sum_{n=1}^{N^*_{ijk}} \beta^n_{ijk} x^n_{ijk} \tag{10}$$

The quality of this approximation depends on the fraction of the OD flow that was indeed distributed over the cut-off set of paths P^*_{ij}; we return to this point below. By definition, we can also write

$$x^n_{ijk} = \beta^n_{ijk} x_{ijk}, \forall, i, j, k, \text{and } n \in \{1, \dots, N^*_{ijk}\} \tag{11}$$

Finally, consider we have - for a limited portion of the links - counts y^m_k that are informative of traffic demand, or more precisely of those path flows that go through link e_m in period k. This is the case if and only if (a) e_m is not congested in period k, and (b) none of the links upstream of e_m in the set of paths $p^n_{ijl} \in P^m_k, l \leq k - \overline{TT}^{n|m}_{ijk}, \forall i, j, n$ (i.e. all paths that traverse e_m during period k) were congested. $\overline{TT}^{n|m}_{ijk}$ depicts the partial arrival travel time (i.e. up to link e_m) along path p^n_{ijl} when traversing link e_m in period k. These counts now add a fourth constraint on the path flows, that is,

$$y^m_k = \sum_i \sum_j \sum_{p^n_{ijl} \in P^m_k} x^n_{ijl} \tag{12}$$

Equations (6), (7), (11), and (12) constitute a system of equations for the path flows and via (10), they can be defined based on OD flows as formulated in (13), (14), and (15). A full derivation of the system of equations is provided in Appendix A. This system is either underdetermined or overdetermined (or in rare cases full rank), depending on the available link counts, and the choice for the number of shortest paths N^*_{ijk}.

$$x_{i1k} + \cdots + x_{ijk} + \ldots = P_{ik} \tag{13}$$

$$\vdots$$

$$x_{1jk} + \cdots + x_{ijk} + \ldots = A_{jk} \tag{14}$$

$$\vdots$$

$$\beta^n_{11k} x_{11k} + \cdots + \beta^n_{ijk} x_{ijk} + \ldots = y^m_k \tag{15}$$

$$\vdots$$

This system of equations can be solved by reformulating it as a matrix equality

$$C\mathbf{x} = B, \ \ \mathbf{x} \geq 0; \tag{16}$$

where $\mathbf{x}$ is shorthand for the OD matrix x_{ijk}; C contains ones (LHS of equations (6), (7)) and the proportionality assumptions in equation (12); and B the RHS of equations (6), (7); and the LHS of equation (12), respectively. Matrix equality (16) can be solved using constrained linear least squares solution (Altman and Gondzio, 1999) with lower bound set to 0 to ensure a non-negative solution. In case a non-negative solution does not exist, an ordinary least square solution is computed and the negative values of the OD matrix are ignored when computing the estimation error.

2.3. Part II: OD estimation in large networks: reducing the solution space through PCA

Since the number of unknown OD flows grows quadratically with the number of (production and attraction) zones, and the number of rows in matrix equation (16) only grows linearly in the number of zones (6) & (7) and linearly in the number of link flow constraints (12); the linear system in (16) becomes (severely) under determined for large networks with too few link flow constraints. To solve the estimation problem for such large networks, we make use of the results in (Djukic et al., 2012b; Zhou and Mahmassani, 2007) which suggest that, loosely speaking, the largest part of variance in demand flows can be attributed to a few dominant (temporal) patterns, that essentially reflect (daily, weekly) seasonal patterns; large deviations from those patterns, and random fluctuations around these patterns (Djukic et al., 2012b). Our assumption is that a similar phenomenon may hold for the production and attraction flows P_{ik} and A_{jk}, and that—as in Djukic et al. (2012b)—principal component analysis (PCA) can be used to (a) reduce the dimensionality of these time series and (b) as a result help us figure out which production and attraction zones $\tilde{P}_{ik}$ and $\tilde{A}_{jk}$ are sufficient to reconstruct a corresponding reduced set of (dominant) OD flows $\tilde{x}_{ijk}$ from which in turn (c) the complete OD matrix can be inferred. Put simply, because we expect a limited and predictable set of dynamical demand patterns, we may be able to infer these from just a limited set of production and attraction patterns.

We briefly outline the main rationale here; for further details on the PCA method, refer to e.g. Jolliffe (2002). Consider the time series of attractions $[\ldots, A_k^T, \ldots]$, with $A_k = [A_{1k}, \ldots, A_{N_Z k}]$, where N_Z is the number of OD zones and A_k^T is the column vector of A_k. Let μ_A, and Σ_A depict the mean and covariance matrix of A_k, and Γ the matrix of eigenvectors, so that $\Gamma^{-1}\Sigma_A\Gamma = \Lambda$ diagonalizes the covariance matrix; i.e. Λ is a diagonal matrix with the eigenvalues of Σ_A. With these eigenvectors and values, we can now express a set of new orthogonal uncorrelated variables called "principal components". Ranked in order of eigenvalues λ_i (diagonal element i in Λ); these principal components $p_i = A_k\gamma_i$, $i = 1, \ldots, N_Z$, with γ_i the eigenvector (column i in Γ), then depict those components in the data that explain most variance in the data in decreasing order. What this means is that if for example the first 3 principal components explain say 95% of the variance in the A_k time series, we can use just these 3 components to reconstruct the entire N_Z dimensional time series A_k with a loss of only 5% in terms of variance, which would entail a reduction in dimensionality of $3/N_Z$! Note that the columns in Γ represent the eigenvectors of Σ_A, the rows in Γ can be interpreted as the weights with which each original dimension in the data (i.e. the production or attraction of a particular zone) contributes to each principal component. We apply the same (PCA) process to the time series of production $[\ldots, P_k^T, \ldots]$. As a result, we can find the zones that correspond to the largest dynamics in both production and attraction and use the union of these selected zones for constructing a (strongly) reduced OD solution space.

Consider, for example, we find 3 "dominant" production zones and 2 dominant attraction zones. We can then estimate the 6 OD flows between this set of selected zones (we explain how below). These OD flows are *upper bounds*, because in computing them, we disregard path flows (12) between zones that are not in the "selected" set. In turn, these upper bounds can be used to constrain the system of equations in (16), which may now be solvable. We propose the following procedure:

1. Compute the principal components of A_k.
2. Select the n_p principal components that explain X% of the variance, with X an arbitrary threshold (e.g. 95).
3. Use a cutoff threshold $0 \leq \theta \leq 1$ for the correlation coefficient in Γ to determine the selection of original dimensions (zones) in A_k we consider are most relevant (i.e. are associated with most variance in the data).
4. Construct a reduced set of OD flows $\tilde{x}^a_{ijk}$ for $\tilde{A}_k$ using the selected zones.
5. Repeat steps 1-4 for $\tilde{P}_k$ to construct a second reduced set of OD flows $\tilde{x}^p_{ijk}$ for $\tilde{P}_k$
6. The union of $\tilde{x}^p_{ijk}$ and $\tilde{x}^a_{ijk}$ constitutes the final reduced set of OD flows $\tilde{\boldsymbol{x}}$ (shorthand for $\tilde{x}_{ijk}$)

7. Filling in $\tilde{\boldsymbol{x}}$ in (16) now leads to an inequality, since B will also contain OD flows *not* in $\tilde{\boldsymbol{x}}$. As a result we must now solve

$$C\tilde{\boldsymbol{x}} \leq B, \;\; \tilde{\boldsymbol{x}} \geq 0 \tag{17}$$

8. With (17) solved (i.e. $\tilde{\boldsymbol{x}}$ are now known), we may be able to solve the "original" matrix equation in (16) using $\tilde{\boldsymbol{x}}$ as upper bounds for the corresponding set of OD flows in the full OD matrix $\tilde{\boldsymbol{x}}^{+}$, i.e.

$$C\mathbf{x} \leq B, \;\; \begin{cases} \mathbf{x} \geq 0, \\ \tilde{\boldsymbol{x}}^{+} \leq \tilde{\boldsymbol{x}} \;\; where \;\; \tilde{\boldsymbol{x}}^{+} \subset \mathbf{x}, \;\; and \;\; \tilde{\boldsymbol{x}} \subset \mathbf{x} \end{cases} \tag{18}$$

2.4. Part III: estimating production and attraction patterns

Let us emphasize on beforehand that deriving the production and attraction time series is a (replaceable) component in the OD estimation method. We may estimate such production and attraction patterns from demographic data or any other source using any appropriate method. In this section, we propose a machine learning approach to derive the relationship between demand and the (3D) supply patterns (Figure 1(a)). The overall idea for this approach to predict production and attraction patterns has its seeds in previous work on estimating so-called 3D (network time) supply patterns (Lopez et al., 2017b,a) using speed data available in urban networks. In (Lopez et al., 2017a) we show (a) that the network of Amsterdam over 35 days can be synthesized into only 4 such 3D patterns with each 9 clusters (a tremendous reduction in data that nonetheless reveals underlying regularity patterns); and (b) that these patterns successfully predict the travel time of 84% of all trips with an error margin below 25%. The hypothesis here is that such 3D supply patterns (that span (sub)networks and multiple time periods) are also informative of demand patterns.

We now seek a way to represent the traffic dynamics of the network with compact yet insightful feature vectors. We use two different approaches to build these feature vectors. The first is a Speed-flow based (SF) feature vector, in which we simply use the raw speed and flow time series in the network, and concatenate these as a (high dimensional) multivariate time series. The second is a Speed-flow-topology based (SFT) feature vector. In this case we use a condensed 3D pattern we build using the methodology in (Lopez et al., 2017b) and add topological information to it. This is a much lower dimensional feature vector. In the SF based approach, the supply information is formulated as a time series of speeds and flows, that is

$$FV = \{\tilde{u}_{pqk}, y_{pqk}\}; \;\; \forall \; p, q \in V \text{ and } k \in \{1, ..., n\}, \tag{19}$$

where, for a link between node p, q for period k, $\tilde{u}_{pqk} = u_{pqk} - u_{pq}^{max}$ is the speed relative to the speed limit; y_{pqk} the link flow; and n the number of time periods considered for incorporating the time dynamics of the supply information. The relative link speed is used since the network contains links with different functional road classes with different speed limits. Note that the SF based feature vector formulation does not contain any topology information or any relationship between the static OD zones Z in the demand space and the speeds or flows in supply space.

In order to incorporate spatiotemporal dynamics in the supply data, we construct the alternative SFT based feature vector using the 3D dynamic clustering proposed by Lopez et al. (2017b). The 3D clustering provides a compressed form to represent spatiotemporal dynamics in a supply pattern. The 3D clustering can be summarized in the following steps:

1. Spatiotemporal link speeds are clustered using datapoint clustering technique - we use Growing Neural Gas (Martinetz et al., 1991) to do this.

2. The unconnected clusters undergo post treatment to generate connected clusters that minimize the speed variance within the clusters and maximize speed variance between the clusters.
3. Each connected 3D cluster is represented by a single speed (mean speed of the 3D cluster) and flow (mean flow of the 3D cluster) values.

The result is a very compact representation of the same spatiotemporal dynamics contained in the raw speed and flow time series, but now in the form of a few 3D zone variables (average speeds and average flows). The OD zones are conventionally designed by experts with prior knowledge about the infrastructure network and using additional information such as surveys, socio-economic data, etc (Cascetta, 2013). The OD zone is connected to the infrastructure network via unweighted directional virtual links known as connectors which are also defined by experts. A key missing ingredient is the relationship between the OD zones Z and supply space. We incorporate this relationship between these two different spaces using the 3D zones $\hat{Z}$ in supply space with

$$\hat{FV} = \{\delta_{ij}, \hat{u}_i, \hat{y}_i, \sigma^2_{u_i}, \sigma^2_{y_i}\};\ \forall\ i \in \{1, ..., m\} \text{ and } j \in \{1, ..., n\}, \tag{20}$$

where m is the number of 3D zones in the supply space and n is the number of OD zones in the demand space. $\hat{u}_i$ is the average link speed of the 3D zone i for all time periods k, $\hat{y}_i$ is the average flow, $\sigma^2_{u_i}$ is the speed variance within the 3D zone i and $\sigma^2_{y_i}$ is the flow variance. δ_{ij} is the zonal incidence variable that represent the relationship between the 3D supply zone and OD zone given by

$$\delta_{ij} = \sum_{k=1}^{t} \delta_{ijk}\ \forall\ i \in \hat{Z},\ j \in Z, \tag{21}$$

where δ_{ijk} is 1 if the 3D zone i intersects with OD zone j at time period k and 0, otherwise. t is the total number of time periods in the 3D zones. A 3D zone i intersects with OD zone j if a connector (that connect the OD zone to the supply space) directly connects that OD zone to any node or link in that particular 3D zone. This is under the assumption that the OD zones and the connectors between the supply space and OD zones are known. Thus the δ_{ij} represents the topological relationship between the supply and demand space, $\hat{u}_i$ corresponds to the zonal speed, $\hat{y}_i$ the zonal flow, $\sigma^2_{u_i}$ the zonal speed variance and $\sigma^2_{y_i}$ the zonal flow variance. These attributes together define the SFT based feature vector $\hat{FV}$. Additional attributes can easily be incorporated to extend the definition of the $\hat{FV}$.

There is no known relationship between supply patterns and productions and attractions and hence supervised learning is needed for estimating that relationship. Assuming that the feature vectors FV (19) and $\hat{FV}$ (20) that represent the supply space have non-linear relationships with the demand space (productions and attractions), we choose a non-linear machine learning approach to model this relationship - an artificial neural network (ANN). ANN is trained with the feature vectors (FV and $\hat{FV}$) as the input and some ground truth for production and attraction as the output. We choose a standard feed-forward model with three layers: an input, hidden and output layer, with linear output functions and the well known hyperbolic tangent sigmoid function as activation functions for the neurons in the hidden layer (Vogl et al., 1988). Two separate neural network models are trained (with feature vectors FV and $\hat{FV}$ as input vectors respectively) to predict the two demand variables (production and attraction respectively), resulting in a total of four neural networks. In order to minimize model complexity (i.e. keeping the number of weights as small as possible) while retrieving the relevant relationship within the data, we use a relatively small neural net with just 5 hidden neurons. Our assumption is that since traffic knowledge has been incorporated into the feature vector, the correlations with the demand and supply space will be more evident. Since the feature vector contains different types of measurements with different units, we normalize the feature vector in order to avoid biasing the neural net against a particular feature dimension or data source. The neural network models for FV and $\hat{FV}$ are given in Figure 2.

In this work, we use Levenberg-Marquardt optimization (Marquardt, 1963) for updating the weights and bias values of the neural network (Hagan and Menhaj, 1994). To avoid overfitting we choose a (time-consuming but robust) leave-one-out strategy for the training and testing. This strategy works like this: given a dataset of say 100 elements (inputs

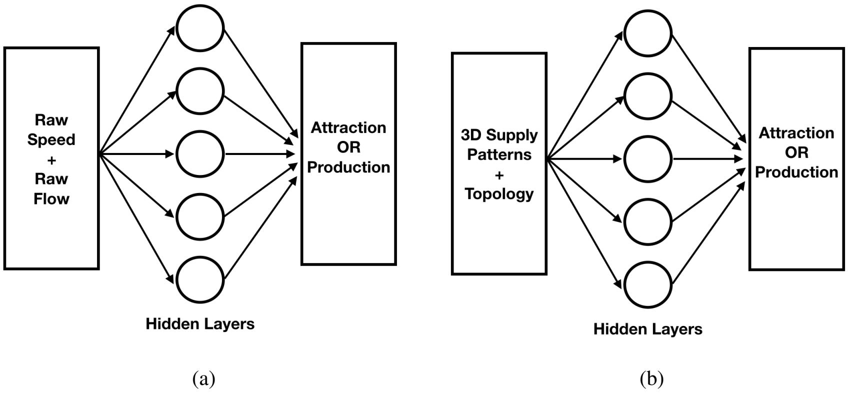


Fig. 2: Neural network model with 5 neurons in the hidden layers for (a) SF (with FV) based predictions (b) SFT (with $\hat{FV}$) based predictions

and targets), 99 are used for training and validation to construct the neural network whereas predictive performance is done on the remaining 1 data element. This is repeated iteratively for all the elements in the data set to achieve a robust prediction accuracy.

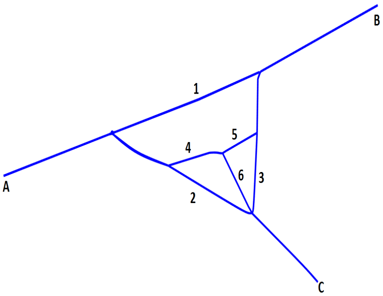
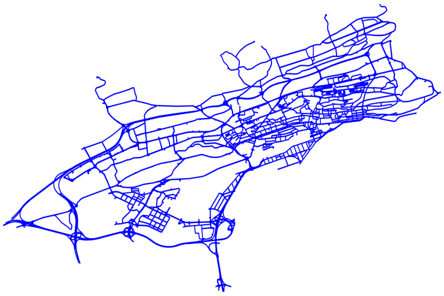
3. Experimental setup

3.1. Data and Networks

As outlined in Figure 1, our method consists of two main components. The first is to predict time series of production and attraction for each demand zone; and the second is to use these predictions to estimate the OD matrix. We demonstrate and validate both components with two networks. First, a relatively simple toy network depicted in Table 1 (middle column), in which nodes A, B and C represent both origins and destinations. This results in 6 OD pairs, each of which has between 2 and 6 (reasonable) routes over which the OD flows are distributed. Note that the graph is directed; however, we use the same link number (1 to 6) to refer to both directions. We simulate this simple network in Aimsun Next (Aimsun, 2017) with a ground-truth OD matrix such that (mild) congestion occurs at the merges (i.e. upstream of some nodes). The entire framework is also applied to a much larger and (with actual data) validated network of Santander, Spain, to demonstrate its scalability in terms of estimation quality and in its ability to reconstruct the correct spatiotemporal structure of the OD matrix. The properties of the network and data availability of the toy network and Santander network are described in Table 1. Note that a zone as defined here may contain multiple origin and destination nodes (sources and sinks) that all connect to different links in the infrastructure network. In both networks the simulated data sufficiently cover supply space, that is, we have a sufficient spread over congested and uncongested conditions, varied over different parts of both networks. This assures the generalization power of the method for different traffic states.

A key assumption in this study is the size N^*_{ijk} of the path set P^*_{ij} between the OD zones for computing the proportionality (equation 8) and the time dependent travel time for the OD estimations. Although this is not necessary for the method, we choose one cut-off value $N^*_{ijk} = N^*$ for all OD pairs and departure time periods within a single scenario and vary with N^* over different scenarios (see below). We emphasize, however, that in principle N^*_{ijk} may be origin, destination, or OD specific and/or even time dependent. We return to this point in the conclusion and discussion section. For the toy network, it is viable to compute *all* the possible paths between each of the OD zone pairs. The paths between two zones i, j are computed by finding the paths between centroid of a given zone (source) to centroid of the other zone (sinks). This computation can easily explode when the network is larger. For example, using a simple Dijkstra algorithm to find the top 10 shortest paths between the 115 OD zone pairs in Santander requires a computation time of approximately 2 weeks on a decent 64-bit windows machine. There are different approaches for speeding up this computation, such as using link elimination to find alternative routes. In this study, for the toy network, we use

Table 1: Data description for the case study - toy network and large scale Santander network.

Properties	**Toy Network**	**Santander Network**
Network Description		
Network Size (Number of links, Number of nodes)	(40, 25)	(4205, 1630)
Number of OD Zones	3	115
Supply Aggregation	5 mins	5 mins
Demand Aggregation	10 mins	5 mins
Supply Time Periods	288 timeslices for 24 hours	48 timeslices for 4 hours of peak period
Demand Time Periods	96 timeslices for 24 hours	48 timeslices for 4 hours of peak period
Data Availability	7 days	7 days

all the possible paths between the OD zones; for the Santander network, we use the top 10 shortest paths between a given zone to the target zone.

3.2. Scenarios and performance measures

We conduct an extensive validation study on the toy network for the two components of our method described in section 2.2 and 2.4 using different scenarios. Firstly, we assess the quality of estimated OD matrix as a function of the accuracy with which we predict production and attraction. Prediction accuracy for all the scenarios in this study is measured using mean absolute percentage error (MAPE) given by

$$M = \frac{1}{n}\sum_{t=1}^{n} |\frac{x_t - \hat{x}_t}{x_t}| * 100\%, \tag{22}$$

where x_t is the ground truth and $\hat{x}_t$ is the predicted value. We consider 11 equally distributed prediction error bins from 0 to 100% to analyze this.

Second, we vary with the time aggregation of the production and attraction time series, and as a consequence of the dynamic OD matrix. For the feature vector that contains topology information, we use 3D zones through clustering. One of the key parameters for building these 3D zones is the number of clusters relative to the size of the network. A third scenario we consider is hence the number of clusters or 3D zones we consider. Fourth, we vary with the number of paths N^* between the OD zones. The optimal choice for N^* will likely heavily depend on the network topology but a thorough sensitivity analysis can provide insight into the number of paths required for fully describing the OD space of that network. We consider six "N^* scenarios": $N^* = 1, 2, 3, 4, 5, 6$ for the toy network and 10 for Santander network.

Finally, link counts are necessary to solve the system of equations for the OD estimates. The link counts are available for all the links in the network for every 5 minutes. However, it is more important to have the counts at the right locations than having more counts at irrelevant locations. For the toy network, we study three link-count availability scenarios:

1. counts on the outer ring (links 1, 2, 3 in Table 1),
2. counts on the inner ring (links 4, 5, 6 in Table 1),
3. counts on both inner and outer ring (links 1 - 6 in Table 1).

This will shed some light on the sensitivity of the OD estimation in the absence of such data. Thus, with the 3 link count scenario and 6 N^* scenarios, we have a total of $3 \times 6 \times 11 = 198$ scenarios for the OD estimation in the toy network. For the Santander network, there are 334 detectors available and we use the link counts at these locations for the OD estimation error analysis. Thus, there are $10 \times 11 = 110$ scenarios for the Santander network with 10 shortest paths. A summary of all the parameters and scenarios used in the study is given in Table 2.

In the next section, we discuss the results of the two methods, that is, production and attraction prediction and OD estimation. To this end we give an extensive sensitivity analysis of different parameters in both the methods for the toy network along the dimensions in Table 2. A thorough sensitivity analysis is not performed on Santander. The aim here is to show that the method is scalable and that the assumptions are valid for different networks as well.

Table 2: Summary of the parameters used in the scenarios

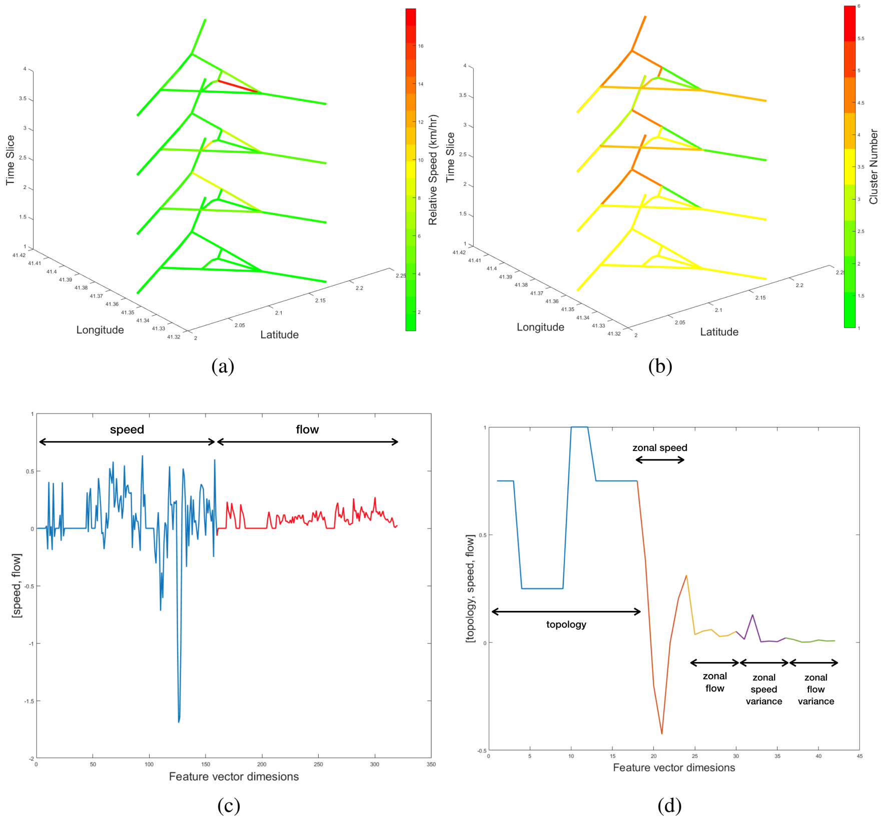
Parameters	**Toy Network**	**Santander Network**
FV	X	
$\hat{FV}$	X	X
Aggregation	X	
Number of clusters	X	
Number of shortest paths	X	X
Link count availability	X	

4. Results and Discussion

4.1. Production and attraction prediction

Before discussing prediction quality, we first discuss the differences between applying the two feature vectors we introduced in the previous section. Figure 3 illustrates the difference between the two feature vectors FV and $\hat{FV}$ (section 2.4) for a relatively short period of 4 time periods, see Figure 3(a). The color represents the relative speed (negative values imply a lower speed). Note that the relative speed of the toy network for the whole data set range from $-60km/hr < u_{pqk} < 30km/hr$. The associated 3D clustering results for the 4 example time periods into 6 3D zones are shown in Figure 3(b). The zones are clearly connected in space and time and each zone can be represented by a single relative speed (and an average flow computed for the same links). Figure 3(c) and (d) show representations of the feature vectors FV and $\hat{FV}$ respectively. The feature vectors are clearly different from each other for the same traffic state. Obviously, $\hat{FV}$ is much more compact than FV—it contains 6 speeds, flows and variances of both, and an incidence vector depicting how the six zones connect to the demand zones, in total 42 dimensions. The dimension of FV is much larger (over 320), and is a function of the number of time periods considered. Whereas the dimension of FV will explode with increasing number of time periods; the dimension of $\hat{FV}$ will remain the same.

Figure 4 shows a single instance of the production and attraction prediction accuracy using the FV and $\hat{FV}$ feature vectors respectively. Both the approaches have a median prediction error of $17 - 18\%$. Closer inspection shows that particularly in predicting production the $\hat{FV}$ results follow the ground-truth more closely (Figure 4)—a result we found consistently over the various scenarios. We can safely conclude that it is advantageous to construct a feature vector using the 3D supply patterns developed in Lopez et al. (2017b). By encoding mean and variance of 3D cluster speed and flow in combination with the topological connection of the 3D clusters to the demand zones, this reduced feature vector $\hat{FV}$ offers a parsimonious abstraction of the entire spatiotemporal supply dynamics of a network. It turns out that the neural net is quite able to correlate $\hat{FV}$ with production and attraction dynamics, and it does so equally well or better than with the much larger feature vector FV, in which simply a time series of all link speeds and flows are used. In the ensuing, we show result with the reduced feature vector only.



(a) (b)

(c) (d)

Fig. 3: Feature vector representation for SF based predictions (a, c) and SFT based predictions (b, d). (a) Speed for 4 time slices. Red represents high speed difference. (b) 3D zones for 4 time slices (c) Corresponding feature vector FV for SF contains normalized {speed, flow} as defined in (19) (d) Corresponding feature vector $\hat{FV}$ for SFT contains normalized vector for representing topology relationship between OD zones and 3D supply zones and supply information for the 3D zones as {topology, zonal speed, zonal flow, zonal speed variance, zonal flow variance} as defined in (20).

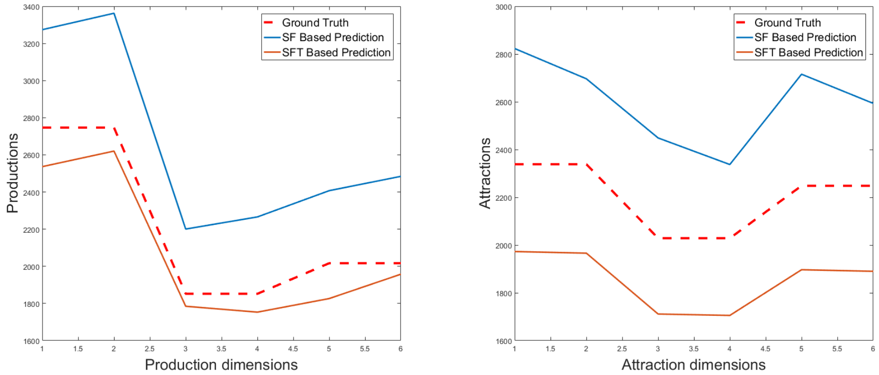


Fig. 4: Single instance of production and attraction prediction for 20 mins time horizon respectively for toy network

Figure 5 shows boxplots of the leave-one-out production and attraction prediction errors using the $\hat{FV}$ feature vectors with 6 3D zones and a time aggregation of 20 mins. The figure shows that 75% of the predictions for both attraction and production have errors less than 35% with a median error of 17% for each production and attraction dimension. This is solid support for the hypothesis that there are strong correlations between supply and demand patterns that can be exploited for OD estimation. We will discuss in the next section whether these errors are small enough to estimate the resulting OD matrix.

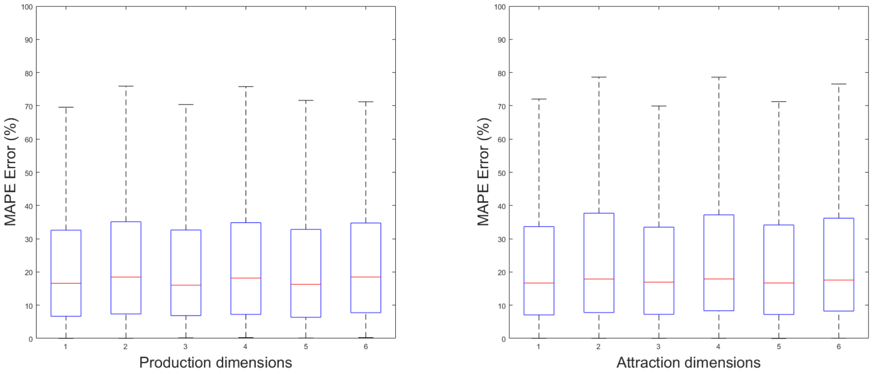
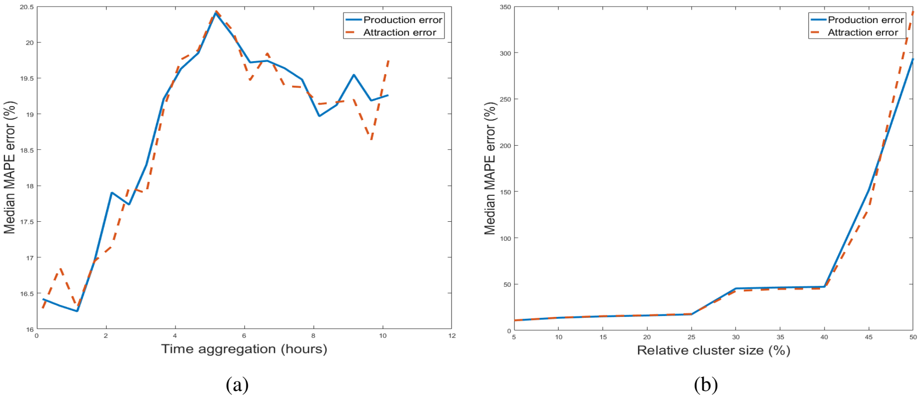


Fig. 5: Leave-one-out production and attraction errors respectively for toy network.

Figure 6(a) shows the relation between level of aggregation and prediction quality. Both attraction and production predictions show similar trends of increasing errors with larger time aggregation. For both we find a median MAPE error of ≈ 16% for aggregation levels of 20 and 60 minutes. This is good news, a 20 minute dynamic production and attraction pattern provides a better basis for estimating dynamic OD matrices in many contexts than a very coarse aggregation that encompasses an entire peak period. The knife cuts both ways here: lower aggregation also provides more training data for the neural network. We return to this point also in the final section of this paper.



(a)

(b)

Fig. 6: Sensitivity analysis of production and attraction prediction accuracy with respect to time aggregation and number of 3D zones respectively.

Figure 6(b) shows the relation between the number of 3D clusters that constitute the $\hat{FV}$ feature vector and the production/attraction prediction quality. Our assumption was that as the number of clusters increases, we can capture more dynamics in the data. However, as can be seen from Figure 6(b), having more clusters in a small network actually impedes learning the patterns. In Figure 6(b), the graph is relatively stable until the number of clusters exceeds about 25% of the network size which is equal to 10 links for the toy network. In hindsight, this inverse proportional relationship between cluster size and prediction error can be easily understood. In the extreme case when the number of clusters is for example 50% relative to the size of the whole network, each cluster contains on average only 1 or 2 links. Such small clusters do not provide much meaningful information about the spatiotemporal traffic dynamics. Hence, the number of clusters have to be chosen relative to the size of the network. We will now look at the OD estimation results on the basis of these predicted production and attraction patterns. For this analysis we consider the optimal time aggregation of 20 mins, and a feature vector with 6 3D clusters.

4.2. OD estimation accuracy

Figure 7 presents the OD estimation errors as a function of (vertically) the errors in production (and attraction—these figures look very similar and are not shown here), and (horizontally) the number of shortest paths (N^*) that is considered. The errors are provided as a discrete contour plot, with the color indicating the error size. Each picture in Figure 7 represents 1 of the 3 detector availability scenarios for the toy network. The figure shows that the OD estimation error gradually increases with increasing production errors. A $0 - 10\%$ mean prediction error in the production results in an OD estimation error of $\approx 20\%$ in the first detector scenario (counts on the outer ring). The error then increases at an almost constant 10% rate with each 10% increase in production error.

The relationship with the number of shortest paths N^* is very different, There appears to be an optimal N^* for a given network to obtain the smallest OD matrix error—in the toy network case we find $N^* = 5$. The optimal N^* will likely depend on the network topology and the prevailing travel patterns in the network. OD estimation accuracy deteriorates when N^* is smaller than the actual average number of routes between OD pairs for obvious reasons: too much flow is distributed over too few routes. The estimation accuracy also deteriorates when N^* is larger than the actual average number of routes between OD pairs, because the proportionality constraint will now redistribute the flow to these additional irrelevant paths. Figure 7 provides clear evidence for this. For all 3 cases, the optimal N^* indeed equals 5. Note that we could have diversified N^*_{ij} for different OD pairs i, j to account for the fact that different OD pairs may have different "optimal" number of paths.

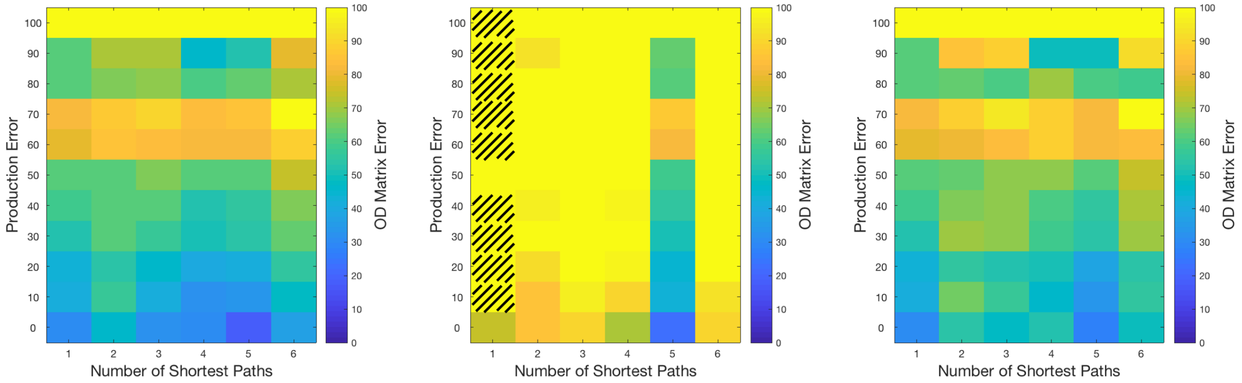


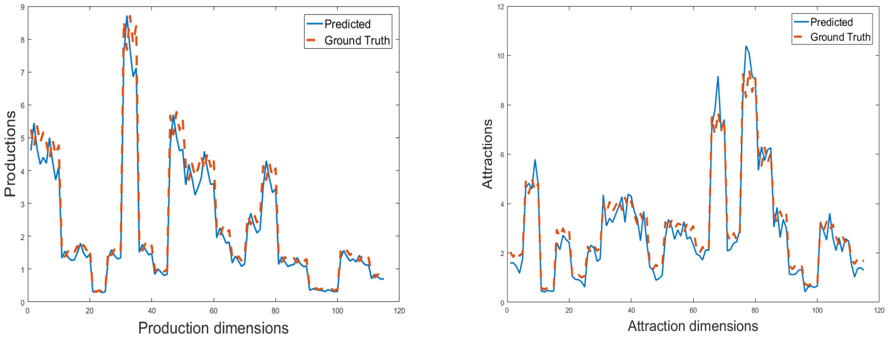
Fig. 7: Mean OD matrix estimation error for the 198 scenarios of the toy network (a) Case 1 - counts on the outer ring, (b) Case 2 - counts on the inner ring, and (c) Case 3 - counts on the inner and outer ring. The shaded region in the case 2 corresponds to the cases where some of the OD matrix estimates returned negative values implying there is no solution for these cases. These estimates were not considered when computing the mean OD matrix errors.

A look at the three different scenarios of link count availability in Figure 7 shows that case 1 and 3 outperform case 2. This is because the top shortest paths between the zones all use the links in the outer ring. This implies that the counts in the inner ring links are not used for solving the equations. However, case 2 also fairs better when $N^* = 5$, since this route set will now include links in the inner ring. When the number of shortest paths is 1 for case 2, some of the scenarios did not provide a non-negative solution. This is depicted by the shaded region in Figure 7. In all the scenarios non-negative solutions were found for the OD estimation. That case 3 (link counts in both inner and outer ring) does not provide better results than case 1 is easily explained. The number of link counts are not as important as having counts at relevant locations. The key lies in the number of linearly independent equations that can be constructed, i.e. in the rank of matrix C in equation (16).

4.3. Santander Case Study

Finally, we apply the framework on a larger network of the city of Santander. We did not perform an extensive sensitivity analysis to estimate the optimal number of 3D zones and aggregation level for production-attraction prediction. We used an aggregation of 20 mins (which was optimal in the Toy network case). An initial attempt to use

10 3D zones was unsuccessful as the neural network was not able to converge. The traffic dynamics were perhaps aggregated too much and the neural net may not have been able to separate the resulting patterns. However, by using 20 clusters, the neural network was able to achieve a median accuracy of $20 - 22\%$ for both production and attraction. A look at one of the (favorable) predictions in Figure 8 shows that the neural network was able to capture the high-dimensional pattern accurately. In the figure we see for one time period the 115 values of the production vector (a) and the attraction vector (b) for one time period.



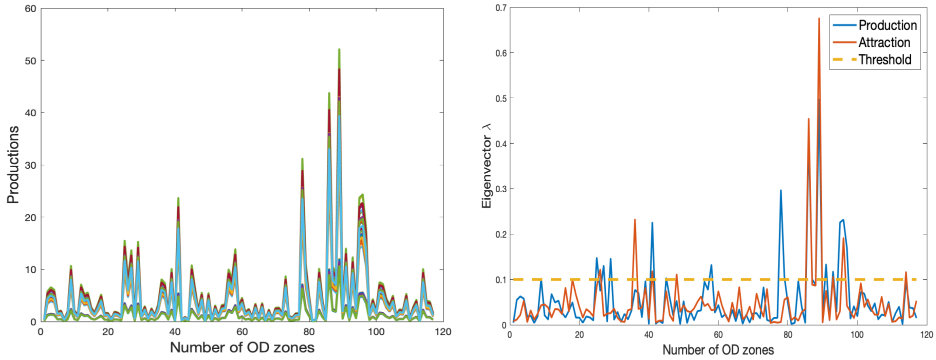
(a)

(b)

Fig. 8: Single instance of production and attraction prediction for 20 mins time horizon respectively

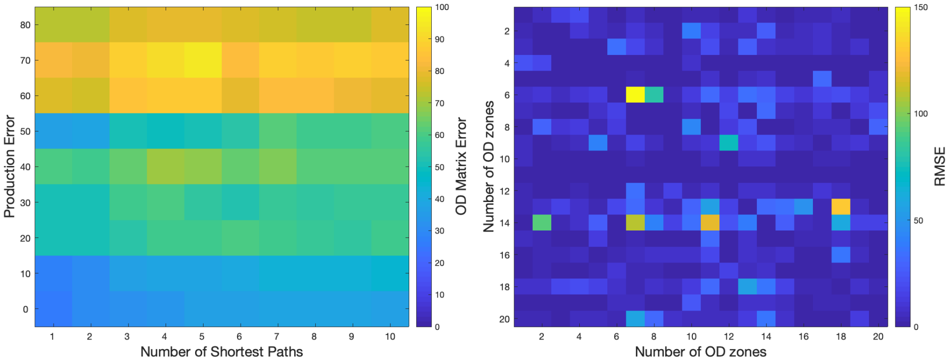
In the Santander case, we apply the PCA method of section 2.3 to construct a reduced solution space. This is indeed necessary, because the full matrix equality (16) could not be solved (was quite severely under determined). It turned out that only a single principal component was needed to explain $\approx 99\%$ of the variance in the production and attraction over a single day. The explanation is that the (calibrated) OD matrix for this network is created by scaling the entire OD matrix by time dependent factor $\tau(t)$ and thus the original dimension varies in the same direction. This is illustrated in Figure 9(a) which shows the production dynamics for all the time periods for a particular day and it is evident that only the magnitude varies for different periods while the demand distribution between zones remains unchanged. Thus, we consider only a single principal component to extract the reduced OD matrix. Figure 9(b) shows the eigenvector λ_1 of the production and attraction time series w.r.t. to the first principal component. Clearly, some of the dimensions (zones) have high correlation with the principal component. We use a cutoff threshold of 0.1 to select a total of 20 zones (as a consequence of the union of 16 production and 9 attraction zones, of which 5 overlap). This leads to a reduced OD space of 20×20 (400 unknowns)—a reduction in dimensionality of 97%(!) compared to the original 115×115 OD matrix (13225 unknowns).

We use the 8 steps in section 2.3 to estimate the full OD matrix, with the reduced OD matrix used as upper bounds for 400 of in total 13,225 unknowns. Figure 10(a) presents the OD estimation error for the Santander network in the same way as in Figure 7(a). The optimal number of shortest paths for the Santander network is $N^* = 1$ which has an overall average error of 26% for $0 - 10\%$ production error. This is justifiable as the traffic congestion in Santander is not high and thus the alternative routes do not provide much additional travel time gain. Therefore, a majority of the OD flows is indeed distributed through the top shortest path. This is also the reason for the marginal influence of the number of shortest paths on the OD estimation error. From figure 10(a), the error only varies from 26% to 37% for $N^* = 1$ to $N^* = 10$. To illustrate the error distribution in the actual OD matrix, Figure 10(b) shows the RMSE of the reduced OD matrix for a single peak period for $N^* = 1$ (Due to the sparsity of the OD matrix, we use the reduced OD matrix to visualise the error in detail) The full OD matrix has a mean error of 22% with a mean RMSE value of 1.5 and a maximum of 150 vehicles/hour.



(a) (b)

Fig. 9: (a) Production dynamics for 48 time periods for a particular day (b) Eigenvector of production and attraction for the first principal component.



(a) (b)

Fig. 10: (a) Mean OD matrix estimation error for the 110 scenarios of Santander network (b) OD estimation error for the reduced matrix for $N^* = 1$.

5. Conclusion and discussion

In this paper we make an argument for a new data driven OD estimation method for cases where a supply pattern in the form of speeds and flows is available. We show that with these input data, an iterative traffic loading procedure with many assumptions on average (or individual) driving behavior; individual route choice trade-offs, and the type of equilibrium network traffic state that emerges *is not needed to solve the OD estimation problem*. The ingredients of this data driven method are (a) a method to estimate / predict production and attraction time series; (b) a method to compute the N^* shortest paths from each OD zone to the next; and (c) assumptions on the magnitude of N^*, and on the proportionality of pathflows between these origins and destinations, respectively. The latter constitutes the only behavioral assumption in our method, and we choose a proportionality factor that is inversely proportional to realized travel time, where we do incorporate a penalty for path overlap. For large networks, these ingredients may not be sufficient to find a unique OD matrix. We show that with a relatively simple and well-known feature space reduction technique (PCA: principal component analysis) we can derive additional constraints directly from the data to solve this underdeterminedness, and as a result a full OD matrix.

From the results on a small toy network and a large real network (of the Spanish city of Santander) we can draw the following conclusions. Most importantly, our simplified OD estimation method works, given a reasonable choice of N^* and a supply pattern (speed in the network) from which all route travel times between all ODs can be estimated. Second, we show that production and attraction time series can be predicted with a coarse 3D supply pattern that includes network topology. This makes the method scalable both in terms of network size and for different topologies, and for cases when the available data is coarse. Third, in the toy network case, we find that OD estimation quality is linearly related to production/attraction prediction quality. With a 10% P/A prediction error, the OD estimation error is about 20% with an additional 10% for every additional 10% P/A prediction error. In the full network we applied the PCA method and were indeed able to estimate the full OD matrix with a mean error of 22% and associated RMSE of 1.5 vehicles.

There are many paths to further explore, refine, and improve the properties of the method. Here we provide just a few main ideas. For large networks, we can refine the "8-step PCA procedure" proposed in computing also lower bounds (instead of just upper bounds). This would necessitate an iterative procedure, but may improve estimation accuracy considerably. Secondly, a key point is that we chose a single assumption on the (average) number of shortest paths N^* for the entire network. Particularly for larger networks, literature (e.g. Viti et al. (2014); Cipriani et al. (2014)) suggests this assumption is likely too crude. The method, however, allows the analyst to diversify different assumptions on N^* for different (groups of) OD pairs, and similarly, add other factors than just travel time that may play a role here (costs, topological considerations, etc). Another avenue of future research is to identify parameters for the method such as the number of clusters, $N*$, for a given transport network, etc. without extensive sensitivity analysis. We believe this strongly depends on the topology of the network, choice set per OD patterns, etc. A fourth direction of further research relates to predicting the production and attraction patterns, which in the method is essentially an exchangeable building block. We see many research avenues here, because there are many aspects with which we may (systematically) vary. For example, what if we can actually (partially) observe production and attraction patterns? Perhaps we do not need supervised prediction models altogether? In this paper, we use a neural network model, which implies a ground-truth data set is required (including production-attraction time series). However, there are straightforward paths for alternative approaches. For example, by using and fusing demographic data with specific cordon counts, household surveys or movement traces. These latter sources could also be used as extra (soft) constraints on the spatial structure of the predicted production and attraction patterns and/or the estimated OD matrix.

Finally, the framework would fit very nicely in an online (DTA) modeling and management framework (with or without assumptions on equilibrium, response to ITS/information, etc), since the OD is now estimated *independently* from any disaggregated route choice or (micro/mes/macro) traffic assumptions. We are excited this idea turned out so fruitful, and hope it will open up many avenues for new research and applications within our field!

Acknowledgements

This research is sponsored by the SETA project funded by the EU's Horizon 2020 Research and Innovation program under the grant agreement No 688082, and by the NWO/TTW project MiRRORS under grant agreement 16270. We thank the anonymous reviewers for the constructive feedback which has greatly improved this paper.

References

Aimsun, 2017. Aimsun Next 8.2 User's Manual. Aimsun SL, Barcelona, Spain.

Alexander, L., Jiang, S., Murga, M., Gonzalez, M.C., 2015. Origin-destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies 58, Part B, 240–250.

Altman, A., Gondzio, J., 1999. Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization. Optimization Methods and Software 11, 275–302.

Antoniou, C., Barcelo, J., Breen, M., Bullejos, M., Casas, J., Cipriani, E., Ciuffo, B., Djukic, T., Hoogendoorn, S., Marzano, V., Montero, L., Nigro, M., Perarnau, J., Punzo, V., Toledo, T., van Lint, H., 2016. Towards a generic benchmarking platform for origin-destination flows estimation/updating algorithms: Design, demonstration and validation. Transportation Research Part C-Emerging Technologies 66, 79–98.

Antoniou, C., Ben-Akiva, M., Koutsopoulos, H.N., 2006. Dynamic traffic demand prediction using conventional and emerging data sources 153, 97–104.

Arentze, T., Timmermans, H., 2009. A need-based model of multi-day, multi-person activity generation. Transportation Research Part B: Methodological 43, 251–265. doi:10.1016/j.trb.2008.05.007.

Ashok, K., Ben-Akiva, M.E., 2000. Alternative approaches for real-time estimation and prediction of time-dependent origin-destination flows. Transportation Science 34, 21–36.

Ashok, K., Ben-Akiva, M.E., 2002. Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows. Transportation Science 36, 184–198.

Barcelo, J., Montero, L., Bullejos, M., Serch, O., Carmona, C., 2013. A kalman filter approach for exploiting bluetooth traffic data when estimating time-dependent od matrices. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations 17, 123–141.

Barcelo, J., Montero, L., Marquos, L., Carmona, C., 2010. Travel time forecasting and dynamic origin-destination estimation for freeways based on Bluetooth traffic monitoring. Transportation Research Record: Journal of the Transportation Research Board 2175, 19–27.

Bell, M.G.H., 1991. The real time estimation of origin-destination flows in the presence of platoon dispersion. Transportation Research Part B: Methodological 25, 115–125. doi:10.1016/0191-2615(91)90018-E.

Bhat, C., Zhao, H., 2002. The spatial analysis of activity stop generation. Transportation Research Part B: Methodological 36, 557–575. doi:10.1016/S0191-2615(01)00019-4.

Bierlaire, M., Toint, P.L., 1995. Meuse: An origin-destination matrix estimator that exploits structure. Transportation Research Part B: Methodological 29, 47–60–.

Camus, R., Cantarella, G.E., Inaudi, D., 1997. Real-time estimation and prediction of origin–destination matrices per time slice. International Journal of Forecasting 13, 13–19–.

Cantelmo, G., Cipriani, E., Gemma, A., Nigro, M., 2014. An adaptive bi-level gradient procedure for the estimation of dynamic traffic demand. IEEE Transactions on Intelligent Transportation Systems 15, 1348–1361.

Cantelmo, G., Viti, F., Cipriani, E., Marialisa, N., 2015. A two-steps dynamic demand estimation approach sequentially adjusting generations and distributions, in: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, pp. 1477–1482.

Carrese, S., Cipriani, E., Mannini, L., Nigro, M., 2017. Dynamic demand estimation and prediction for traffic urban networks adopting new data sources. Transportation Research Part C: Emerging Technologies 81, 83–98. doi:10.1016/j.trc.2017.05.013.

Cascetta, E., 1984. Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator. Transportation Research Part B 18, 289–299. doi:10.1016/0191-2615(84)90012-2.

Cascetta, E., 2013. Transportation systems engineering: theory and methods. volume 49. Springer Science & Business Media.

Cascetta, E., Marquis, G., 1993. Dynamic estimators of origin-destination matrices using traffic counts. Transportation Science 27, 363–373. doi:10.1287/trsc.27.4.363.

Cascetta, E., Papola, A., Marzano, V., Simonelli, F., Vitiello, I., 2013. Quasi-dynamic estimation of od flows from traffic counts: Formulation, statistical validation and performance analysis on real data. Transportation Research Part B: Methodological 55, 171–187.

Cascetta, E., Postorino, M.N., 2001. Fixed point approaches to the estimation of o/d matrices using traffic counts on congested networks. Transportation Science 35, 134–147–.

Castillo, E., Rivas, A., Jimenez, P., Menendez, J.M., 2012. Observability in traffic networks. plate scanning added by counting information. Transportation 39, 1301–1333. doi:10.1007/s11116-012-9390-0.

Castro, P., Zhang, D., Chen, C., Li, S., Pan, G., 2013. From taxi gps traces to social and community dynamics: A survey. ACM Computing Surveys 46. doi:10.1145/2543581.2543584.

Chen, C., Bian, L., Ma, J., 2014. From traces to trajectories: How well can we guess activity locations from mobile phone traces? Transportation Research Part C: Emerging Technologies 46, 326–337. doi:10.1016/j.trc.2014.07.001.

Cheng, L., Zhu, S., Chu, Z., Cheng, J., 2014. A bayesian network model for origin-destination matrices estimation using prior and some observed link flows. Discrete Dynamics in Nature and Society 2014. doi:10.1155/2014/192470.

Cipriani, E., Nigro, M., Fusco, G., Colombaroni, C., 2014. Effectiveness of link and path information on simultaneous adjustment of dynamic o-d demand matrix. European Transport Research Review 6, 139–148.

Cremer, M., Keller, H., 1987. A new class of dynamic methods for the identification of origin-destination flows. Transportation research part B - methodological 21, 117–132.

Di, X., Liu, H., 2016. Boundedly rational route choice behavior: A review of models and methodologies. Transportation Research Part B: Methodological 85, 142–179. doi:10.1016/j.trb.2016.01.002.

Djukic, T., Flotterod, G., Van Lint, H., Hoogendoorn, S., 2012a. Efficient real time od matrix estimation based on principal component analysis, in: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, pp. 115–121. doi:10.1109/ITSC.2012.6338720.

Djukic, T., Lint, J.v., Hoogendoorn, S., 2013. Reliability assessment of dynamic od estimation methods based on structural similarity index., in: Transportation Research Board Annual Meeting, National Academies, Washington D.C.. p. 13.

Djukic, T., van Lint, J.W.C., Hoogendoorn, S.P., 2012b. Application of principal component analysis to predict dynamic origin-destination matrices. Transportation Research Record , 81–89doi:10.3141/2283-09.

Frederix, R., Viti, F., Corthout, R., Tampere, C.M.J., 2012. New gradient approximation method for dynamic origin-destination matrix estimation on congested networks. Transportation Research Record: Journal of the Transportation Research Board 2263, 19–25.

Gadzinski, J., 2018. Perspectives of the use of smartphones in travel behaviour studies: Findings from a literature review and a pilot study. Transportation Research Part C: Emerging Technologies 88, 74–86.

Ge, Q., Fukuda, D., 2016. Updating origin-destination matrices with aggregated data of GPS traces. Transportation Research Part C: Emerging Technologies 69, 291–312. doi:10.1016/j.trc.2016.06.002.

Hagan, M.T., Menhaj, M.B., 1994. Training feedforward networks with the marquardt algorithm. IEEE transactions on Neural Networks 5, 989–993.

Jolliffe, I.T., 2002. Introduction. Springer.

Kim, J., Kurauchi, F., Uno, N., Hagihara, T., Daito, T., 2014. Using electronic toll collection data to understand traffic demand. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations 18, 190–203. doi:10.1080/15472450.2013.806858.

Kitamura, R., Chen, C., Pendyala, R., Narayanan, R., 2000. Micro-simulation of daily activity-travel patterns for travel demand forecasting. Transportation 27, 25–51. doi:10.1023/A:1005259324588.

Lopez, C., Krishnakumari, P., Leclercq, L., Chiabaut, N., Lint, H.v., 2017a. Spatio-temporal partitioning of transportation network using travel time data. Transportation Research Record: Journal of the Transportation Research Board .

Lopez, C., Leclercq, L., Krishnakumari, P., Chiabaut, N., van Lint, H., 2017b. Revealing the day-to-day regularity of urban congestion patterns with 3d speed maps. Scientific Reports 7, 14029.

Lu, Z., Rao, W., Wu, Y.J., Guo, L., Xia, J., 2015. A Kalman filter approach to dynamic od flow estimation for urban road networks using multi-sensor data. Journal of Advanced Transportation 49, 210–227. doi:10.1002/atr.1292.

Lundgren, J.T., Peterson, A., 2008. A heuristic for the bilevel origin-destination-matrix estimation problem. Transportation Research Part B: Methodological 42, 339–354. doi:10.1016/j.trb.2007.09.005.

Marquardt, D.W., 1963. An algorithm for least-squares estimation of nonlinear parameters. Journal of the society for Industrial and Applied Mathematics 11, 431–441.

Martinetz, T., Schulten, K., et al., 1991. A" neural-gas" network learns topologies .

Marzano, V., Papola, A., Simonelli, F., Papageorgiou, M., 2018. A kalman filter for quasi-dynamic od flow estimation/updating. IEEE Transactions on Intelligent Transportation Systems , 1–9.

Nigro, M., Cipriani, E., del Giudice, A., 2018. Exploiting floating car data for time-dependent origindestination matrices estimation. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations 22, 159–174.

Okutani, I., Stephanedes, Y.J., 1984. Dynamic prediction of traffic volume through Kalman filtering theory. Transportation Research Part B: Methodological 18, 1–11. doi:10.1016/0191-2615(84)90002-X.

Scheffer, A., Cantelmo, G., Viti, F., 2017. Generating macroscopic, purpose-dependent trips through monte carlo sampling techniques, in: Transportation Research Procedia, pp. 585–592.

Tolouei, R., Psarras, S., Prince, R., 2017. Origin-destination trip matrix development: Conventional methods versus mobile phone data, in: Transportation Research Procedia, pp. 39–52.

Van Der Zijpp, N., 1997. Dynamic origin-destination matrix estimation from traffic counts and automated vehicle identification data. Transportation Research Record: Journal of the Transportation Research Board 1607, 87–94. doi:10.3141/1607-13.

Van Zuylen, H.J., Willumsen, L.G., 1980. The most likely trip matrix estimated from traffic counts. Transportation Research Part B: Methodological 14, 281–293–.

Viti, F., Rinaldi, M., Corman, F., Tampe're, C.M.J., 2014. Assessing partial observability in network sensor location problems. Transportation Research Part B: Methodological 70, 65–89. doi:10.1016/j.trb.2014.08.002.

Vogl, T.P., Mangis, J., Rigler, A., Zink, W., Alkon, D., 1988. Accelerating the convergence of the back-propagation method. Biological cybernetics 59, 257–263.

Widhalm, P., Yang, Y., Ulm, M., Athavale, S., Gonzlez, M., 2015. Discovering urban activity patterns in cell phone data. Transportation 42, 597–623. doi:10.1007/s11116-015-9598-x.

Yang, H., Sasaki, T., Iida, Y., Asakura, Y., 1992. Estimation of origin-destination matrices from link traffic counts on congested networks. Transportation Research Part B 26, 417–434. doi:10.1016/0191-2615(92)90008-K.

Yang, X., Lu, Y., Hao, W., 2017. Origin-destination estimation using probe vehicle trajectory and link counts. Journal of Advanced Transportation 2017.

Zhou, X., Mahmassani, H.S., 2006. Dynamic origin-destination demand estimation using automatic vehicle identification data. IEEE Transactions on Intelligent Transportation Systems 7, 105–114. doi:10.1109/TITS.2006.869629.

Zhou, X., Mahmassani, H.S., 2007. A structural state space model for real-time traffic origin-destination demand estimation and prediction in a day-to-day learning framework. Transportation Research Part B: Methodological 41, 823–840. doi:10.1016/j.trb.2007.02.004.

Zin, T.A., Kyaing, Lwin, K.K., Sekimoto, Y., 2018. Estimation of originating-destination trips in yangon by using big data source. Journal of Disaster Research 13, 6–13.

Appendix A.

The system of equations given in (13), (14), and (15) is obtained from equations (6), (7), (11), and (12). Here, we give a detailed derivation of the system of equations and formulate it as a matrix equality for solving it. First, we reiterate the equations used:

$$P_{ik} = \sum_{j} \sum_{n} x^{n}_{ijk} \tag{A.1}$$

$$A_{jk} = \sum_{i} \sum_{n} x^{n}_{ijk} \tag{A.2}$$

$$x^{n}_{ijk} = \beta^{n}_{ijk} x_{ijk}, \forall, i, j, k, \text{and } n \in \{1, \ldots, N^{*}_{ijk}\} \tag{A.3}$$

$$y^{m}_{k} = \sum_{i} \sum_{j} \sum_{p^{n}_{ijl} \in P^{m}_{k}} x^{n}_{ijl} \tag{A.4}$$

Equation A.1 can be expanded as follows:

$$P_{ik} = \sum_n x_{i1k}^n + \sum_n x_{i2k}^n + \cdots + \sum_n x_{ijk}^n + \ldots \tag{A.5}$$

Equation A.5 can be redefined as sum of OD flows instead of path flows based on equation A.3 as follows:

$$P_{ik} = \sum_n \beta_{i1k}^n x_{i1k} + \sum_n \beta_{i2k}^n x_{i2k} + \cdots + \sum_n \beta_{ijk}^n x_{ijk} + \ldots \tag{A.6}$$

Since the route proportion β_{i1k}^n is summed across all routes n ranging from 1 to N_{ijk}^*, it can be reformulated as follows:

$$\sum_n \beta_{ijk}^n = \begin{cases} 0 & TT_{ijk}^n = \infty \\ 1 & \text{otherwise} \end{cases} \tag{A.7}$$

where $TT_{ijk}^n = \infty$ implies that no path exists between origin i and destination j. Thus, equation A.6 can be written as:

$$P_{ik} = x_{i1k} + \cdots + x_{ijk} + \ldots \tag{A.8}$$

with example $\sum_n \beta_{i1k}^n = 1$, $\sum_n \beta_{i2k}^n = 0$ and $\sum_n \beta_{ijk}^n = 1$.

Equations A.5 and A.6 are reproduced for equation A.2 as given below:

$$A_{jk} = \sum_n x_{1jk}^n + \sum_n x_{2jk}^n + \cdots + \sum_n x_{ijk}^n + \ldots \tag{A.9}$$

$$A_{jk} = \sum_n \beta_{1jk}^n x_{1jk} + \sum_n \beta_{2jk}^n x_{2jk} + \cdots + \sum_n \beta_{ijk}^n x_{ijk} + \ldots \tag{A.10}$$

Using equation A.7, A.10 can rewritten as:

$$A_{jk} = x_{1jk} + \cdots + x_{ijk} + \ldots \tag{A.11}$$

with example $\sum_n \beta_{1jk}^n = 1$, $\sum_n \beta_{2jk}^n = 0$ and $\sum_n \beta_{ijk}^n = 1$.

The link count given in A.4 can be expanded as:

$$y_k^m = \sum_{p_{11l}^n \in P_k^m} x_{11l}^n + \sum_{p_{12l}^n \in P_k^m} x_{12l}^n + \cdots + \sum_{p_{ijl}^n \in P_k^m} x_{ijl}^n + \ldots \tag{A.12}$$

A.12 can rewritten in terms of OD flows using equation A.3 as follows:

$$y_k^m = \sum_{p_{11l}^n \in P_k^m} \beta_{11l}^n x_{11l} + \sum_{p_{12l}^n \in P_k^m} \beta_{12l}^n x_{12l} + \cdots + \sum_{p_{ijl}^n \in P_k^m} \beta_{ijl}^n x_{ijl} + \ldots \tag{A.13}$$

This term is considered if and only if (a) link e_m is not congested in period k, and (b) none of the links upstream of e_m in the set of paths $p_{ijl}^n \in P_k^m, l \leq k - \overline{TT}_{ijk}^{n|m}, \forall i, j, n$ (i.e. all paths that traverse e_m during period k) were congested. $\overline{TT}_{ijk}^{n|m}$ depicts the partial arrival travel time (i.e. up to link e_m) along path p_{ijl}^n when traversing link e_m in period k. Thus,

$$\sum_{p_{ijl}^n \in P_k^m} \beta_{ijl}^n x_{ijl} = \begin{cases} 0 & \text{links upstream of } e_m \in p_{ijl}^n \text{ is congested} \\ \beta_{ijk}^n x_{ijk} & \text{otherwise} \end{cases} \tag{A.14}$$

Based on A.14, A.13 can be rewritten as:

$$y_k^m = \beta_{11k}^n x_{11k} + \cdots + \beta_{ijk}^n x_{ijk} + \ldots \tag{A.15}$$

assuming $\sum_{p_{12l}^n \in P_k^m} \beta_{12l}^n x_{12l} = 0$

Combining the expanded form of production, attraction and link counts given in A.8, A.11 and A.15 respectively, the system of equations can be formulated as:

$$x_{i1k} + \cdots + x_{ijk} + \ldots = P_{ik} \tag{A.16}$$

$$\vdots$$

$$x_{1jk} + \cdots + x_{ijk} + \ldots = A_{jk} \tag{A.17}$$

$$\vdots$$

$$\beta_{11k}^n x_{11k} + \cdots + \beta_{ijk}^n x_{ijk} + \ldots = y_k^m \tag{A.18}$$

$$\vdots$$

We can reformulate the system of equations as a matrix equation as:

$$C\mathbf{x} = B, \quad \mathbf{x} \geq 0; \tag{A.19}$$

where

$$C = \begin{bmatrix} \sum_n \beta^n_{11k} & \cdots & 1 & \cdots & \sum_n \beta^n_{j1k} & \cdots & 1 & \cdots \\ \vdots & \ddots & & & & & & \\ \sum_n \beta^n_{11k} & \cdots & \sum_n \beta^n_{i1k} & \cdots & 1 & \cdots & 1 & \cdots \\ \vdots & \ddots & & & & & & \\ \beta^n_{11k} & \cdots & \sum_{p^n_{i1l} \in P^m_k} \beta^n_{i1l} & \cdots & \sum_{p^n_{1jl} \in P^m_k} \beta^n_{1jl} & \cdots & \beta^n_{ijk} & \cdots \\ \vdots & \ddots & & & & & & \end{bmatrix}; \mathbf{x} = \begin{bmatrix} x_{11k} \\ \vdots \\ x_{i1k} \\ \vdots \\ x_{1jk} \\ \vdots \\ x_{ijk} \\ \vdots \end{bmatrix}; \text{and } B = \begin{bmatrix} P_{ik} \\ \vdots \\ A_{jk} \\ \vdots \\ y^m_k \\ \vdots \end{bmatrix} \tag{A.20}$$