OPTIMAL LINEAR ESTIMATOR OF ORIGIN-DESTINATION FLOWS WITH REDUNDANT DATA

FRÉDÉRIC MEUNIER

ABSTRACT. Suppose given a network endowed with a multiflow. We want to estimate some quantities connected with this multiflow, for instance the value of an s–t flow for one of the sources–sinks pairs s–t, but only measures on some arcs are available, at least on one s–t *cocycle* (set of arcs having exactly one endpoint in a subset X of vertices with $s \in X$ and $t \notin X$). These measures, supposed to be unbiased, are random variables whose variances are known. How can we combine them optimally in order to get the best estimator of the value of the s–t flow ?

This question arises in practical situations when the OD matrix of a transportation network must be estimated. We will give a complete answer for the case when we deal with linear combinations, not only for the value of an s–t flow but also for any quantity depending linearly from the multiflow. Interestingly, we will see that the Laplacian matrix of the network plays a central role.

INTRODUCTION

Consider the simple network given in Figure 1 and suppose that you wish to find the total number of cars (the *o-d traffic*) going from the origin o to the destination d. Suppose now that you have an estimate X_{a_1} of this number of cars on arc a_1 with variance w_1, as well as an estimate X_{a_2} on arc a_2 with variance w_2. How can you combine them to get the best estimate of the o–d traffic?

Assuming that the estimates X_{a_1} and X_{a_2} are unbiased, they are both estimators of the whole o–d traffic, since all cars going from the origin o to the destination d have to use arcs a_1 and a_2. Moreover, one can assume that the measures are independent. Hence, if you restrict yourself to linear combination, it is well-known and easy to prove that the best estimator is

$$\frac{w_2}{w_1+w_2}X_{a_1}+\frac{w_1}{w_1+w_2}X_{a_2}.$$

If the network is different, say as in Figure 2, it is also possible to find the best linear estimator. We assume that all estimates are unbiased, independent and have the same precision. It is simply

$$\frac{1}{2}(X_{a_1}+X_{a_2})+\frac{1}{2}(X_{a_3}+X_{a_4}).$$

Now, the central question of this paper is the following one: take any directed graph $D=(V,A)$ endowed with an s–t flow; suppose that we have estimates of the flow on some arcs; find the linear combination of these estimates that provides the estimator of the value of s–t flow of minimal variance. We will always assume that the estimates are independent random variables. This assumption is very natural, a fact that is not necessarily obvious at first glance. It means that the measures of the flow on the different arcs are led independently. For instance, if the procedure consists in randomly stopping cars (with a given probability p for instance) and then asking the driver what is his OD pair, it may happen that the same car is stopped several times during its travel across the network without violating the independence assumption. The point that matters is that the probability of stopping a given car on an arc a is independent of the probability of stopping it on another arc a'.

Key words and phrases. Laplacian of graph, linear estimator, minimal variance, multiflow, OD matrix.

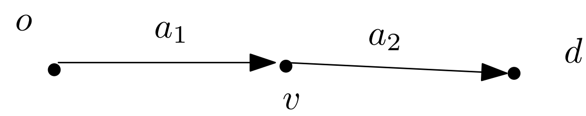

FIGURE 1. A simple case.

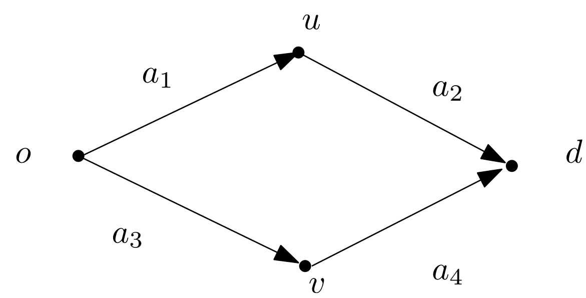

FIGURE 2. Another simple case.

The case with several s–t pairs (multiflow) is also investigated, as well as the estimation of other quantities depending on the flows (for instance, the question of improving the accuracy of the estimate of the s–t flow on a given arc using the other estimates).

We will see that with elementary graph theory and tools from quadratic optimization it is possible to solve this problem in polynomial time. Moreover, we will see that the Laplacian matrix of the network plays a central role in the case of one s–t pair, and also in the case with multiple s–t pairs when estimates are available for all arcs.

The problem modelled and solved in this paper was encountered by engineers of the French Ministry in charge of transportation who had to estimate OD matrices for road traffic. They noticed that they had redundant information and wanted to use it optimally. The problem was exposed in the report [4].

The present work has to be connected with works dealing with similar questions, such as works in [8] or in [6] in a computer network context, of those in [2], [1], or in [7] in a transportation context, among many others. The big difference is that these latters deal with the case when there are not enough data to derive directly an estimator (the quantities are underconstrained) and additional ingredients such as entropy maximization or behavioral assumption are necessary. In the present case, we have in a sense more than enough data and we want to combine them optimally. Some of these results have been presented at the Transport Research Bord Conference 2009 [5].

Outline of the paper. The plan is the following. Section 2 deals with the case detailed above. Only one s–t pair is considered and estimates (more or less precise) of this flow are available for some arcs. We wish to find the best linear estimator of the value of the s–t flow. Theorem 1 summarizes the method for building such an estimator. In Section 3, a somewhat different purpose is considered with the same kind of data, allowing this time that other flows with different origins and destinations take place on the network. We do not wish anymore to estimate the value of an s–t flow, but we aim to estimate any linear combination of the flows. Some potential applications will be briefly described.

1. Notations and basic tools

We denote vectors by bold letters: for instance, if B is a finite set, an element of $\mathbb{R}^B$ is for instance denoted by $\boldsymbol{x}$, while the coordinate corresponding to element b is denoted by x_b. Hence $\boldsymbol{x} = (x_b)_{b\in B}$. Note – as often in mathematics – that we identify the concept of a function $B \to \mathbb{R}$

with that of a vector in $\mathbb{R}^B$. We will do the same kind of identification for matrices: if B and B' are two finite sets, a $B \times B'$ *matrix* is a matrix whose rows are indexed by the elements of B and whose columns are indexed by the elements of B'.

We denote by $\mathcal{ST}$ the set of sources–sinks pairs of our multiflow.

$\mathcal{P}_{st}$ denotes the set of all elementary s–t paths and $\mathcal{C}$ denotes the set of all elementary cycles. Sets of arcs of type $\delta^+(X) \cup \delta^-(X)$ with $X \subseteq V$ (which can also be seen as cuts in the underlying undirected graph) are called *cocycle*. An *s–t cocycle* is a cocycle obtained as above with a subset $X \subseteq V$ such that $s \in X$ and $t \notin X$.

For a subset B of arcs, we denote by χ^B the *indicator vector*:

$$\chi^B_a = \begin{cases} 1 & \text{if } a \in B, \\ 0 & \text{if not.} \end{cases}$$

The sets of arcs leaving (resp. entering) a subset X of vertices is denoted $\delta^+(X)$ (resp. $\delta^-(X)$).

The *incident matrix* M of a directed graph $D = (V, A)$ is a $V \times A$ matrix defined as follows:

$$M = ((m_{v,a}))_{v,a \in V \times A} \text{ where } m_{v,a} = \begin{cases} 0 & \text{if } a \text{ is not incident to } v \\ -1 & \text{if } a \text{ leaves } v \\ 1 & \text{if } a \text{ enters } v \end{cases} \tag{1}$$

We define $F := \text{Ker } M$. The subspace F is the space of *circulations*, which are flows satisfying the Kirchoff law everywhere. $F^\perp$ is the space of *co-circulation*, and it is such that $F^\perp = \text{Im } M^T$, by basic linear algebra. In this context, an element $\boldsymbol{\pi} \in \mathbb{R}^V$ is called a *potential*.

The *(weighted) Laplacian* $L_{\boldsymbol{w}}$ of an undirected weighted graph $G = (V, E)$ with weights $\boldsymbol{w} \in \mathbb{R}^E_+$ is a $V \times V$ matrix defined as follows:

(2)

$$L_{\boldsymbol{w}} = ((w_{u,v}))_{u,v \in V \times V} \text{ where } w_{u,v} = \begin{cases} \sum_{e \in \delta(u)} w_e & \text{if } u = v \\ 0 & \text{if } u \neq v \text{ and } u \text{ and } v \text{ are non adjacent} \\ -\sum_{e=uv} w_e & \text{if } uv \in E. \end{cases}$$

We recall that $\delta(u)$ denotes the set of edges incident to u. The sum $\sum_{e=uv}$ in the equation above means that if there are parallel edges, between vertices u and v, we have to sum their weights.

We can also define the (weighted) Laplacian of a directed weighted graph by simply forgetting the orientations of the arcs. In this case we have

$$M^T \text{Diag}(\boldsymbol{w}) M = L_w, \tag{3}$$

where $\text{Diag}(\boldsymbol{w})$ is the diagonal $A \times A$ matrix with each (a, a) entry equals to w_a.

2. Optimal estimator for the s–t flow

2.1. **Finding a best linear estimator.** The simplest case is when there is only one s–t pair. This is the case detailed in the Introduction. For some arcs a of the network an estimate X_a of the s–t flow going through a is available. Their variances w_a are supposed to be known. We wish to find the linear estimator E of the whole s–t traffic of minimal variance; E has the following form

$$E := \sum_{a \in A} \lambda_a X_a. \tag{4}$$

Our purpose is to find the best estimator of this form, which can be summarized as follows: among all possible values for the λ_a, find those making E an unbiased estimator of the s–t flow and among them, those minimizing the variance of E. Lemma 1 below gives a simple necessary and sufficient condition on the λ_a for E to be an unbiased estimator. Among the λ_a satisfying this condition, we will find those minimizing the variance through a quadratic program (Theorem 1).

Lemma 1. *E is an unbiased estimator of the s–t flow if and only if one has $\sum_{a\in P}\lambda_a = 1$ for each elementary s–t path P and $\sum_{a\in C}\lambda_a = 0$ for each elementary cycle C.*

Proof. Let $\boldsymbol{x}\in\mathbb{R}^A$ be the s–t flow on D. By assumption, one has $\mathbb{E}(X_a) = x_a$.

Write $\boldsymbol{x}$ as a conic combination of elementary s–t paths and elementary cycles (assuming the existence of a linear order, there is no ambiguity in the combination):

$$\boldsymbol{x} = \sum_{P\in\mathcal{P}_{st}} x_P\chi^P + \sum_{C\in\mathcal{C}} x_C\chi^C.$$

Hence, $x_a = \sum_{P\in\mathcal{P}_{st}\colon P\ni a} x_P + \sum_{C\in\mathcal{C}\colon C\ni a} x_C$. The random variable E is an unbiased estimator if and only if $\mathbb{E}(E) = \sum_{P\in\mathcal{P}_{st}} x_P$. Replacing these expression in Equation (4) leads to

$$\sum_{P\in\mathcal{P}_{st}} x_P = \sum_{a\in A}\lambda_a\left(\sum_{P\in\mathcal{P}_{st}\colon P\ni a} x_P + \sum_{C\in\mathcal{C}\colon C\ni a} x_C\right),$$

which can be rewritten

$$\sum_{P\in\mathcal{P}_{st}} x_P = \sum_{P\in\mathcal{P}_{st}}\left(x_P\sum_{a\in P}\lambda_a\right) + \sum_{C\in\mathcal{C}}\left(x_C\sum_{a\in C}\lambda_a\right).$$

By a straightforward identification – this expression shall be correct for all possible values of the variables x_P and x_C – we get the required equality. □

Now, the best estimator is obtained when we minimize the variance.

Theorem 1. *Contract each arc of $D = (V,A)$ with no measure available. Denote by $D' = (V',A')$ the graph obtained by this process, and by $L_{\boldsymbol{w}}$ its Laplacian. If $s = t$ in D', then there is no unbiased linear estimator. If $s \neq t$ in D', $E := \sum_{a\in A}\lambda_a X_a$ is the unbiased linear estimator of minimal variance if and only if $\lambda_a = 0$ for $A\setminus A'$ and $\lambda_{(u,v)} = \pi_v - \pi_u$ for each arc $(u,v)\in A'$, where $\boldsymbol{\pi}\in\mathbb{R}^{V'}$ is solution of the system*

$$\begin{array}{rcll}(L_{\boldsymbol{w}}\boldsymbol{\pi})_u & = & 0 & \text{for } u\in V'\setminus\{s,t\}\\ \pi_s & = & 0 & \\ \pi_t & = & 1. & \end{array}\tag{5}$$

The system (5) has $|V'|$ constraints and $|V'|$ unknowns. Hence, it is easy to compute a linear estimator of best variance. Note that $s = t$ in D' if and only if no s–t cocycle of D has been integrally measured.

Proof. Let us first assume that there is an available measure for each arc. The λ_a are solutions of the following quadratic program

$$\begin{array}{ll}\displaystyle\operatorname*{Min}_{\boldsymbol{\lambda}\in\mathbb{R}^A} & \displaystyle\sum_{a\in A} w_a\lambda_a^2\\[2ex] \text{s.t.} & \left|\begin{array}{ll}\displaystyle\sum_{a\in P}\lambda_a = 1 & \text{for each elementary } s\text{–}t \text{ path } P\\[2ex] \displaystyle\sum_{a\in C}\lambda_a = 0 & \text{for each elementary cycle } C.\end{array}\right.\end{array}\tag{6}$$

In the present form, the set of constraints has an exponential description. Add a fictitious arc (t,s) and put $\lambda_{(t,s)} = -1$. The constraints mean precisely that $\boldsymbol{\lambda}$ is a co-circulation, which comes from a potential $\boldsymbol{\pi} \in \mathbb{R}^V$ such that $\pi_s = 0$ and $\pi_t = 1$. Equation (6) can be rewritten

$$\text{(7)} \qquad \operatorname*{Min}_{\boldsymbol{\pi}\in\mathbb{R}^V} \sum_{(u,v)\in A} w_{(u,v)}(\pi_v - \pi_u)^2; \quad \pi_s = 0,\ \pi_t = 1.$$

Note that $\sum_{(u,v)\in A} w_{(u,v)}(\pi_v - \pi_u)^2 = \boldsymbol{\pi}^T M^T \mathrm{Diag}\,(\boldsymbol{w})\, M\boldsymbol{\pi}$.

Hence, by a straightforward derivation, we get that $\left(M^T \mathrm{Diag}\,(\boldsymbol{w})\, M\boldsymbol{\pi}\right)_u = 0$ for all $u \neq s, t$, as required.

It remains to see why we have to contract edges where there are no measure. In this case, we have to add in Equation (6) constraints of the type $\lambda_{(u,v)} = 0$ where (u, v) is an arc with no measure. Everything remains the same in the proof above, except that then we have $\pi_v = \pi_u$ for such arcs (u, v). This is exactly what we obtain when contracting such edges. $\square$

Remarks : Put a resistance 0 on each arc with no available measure, and $1/w_a$ otherwise. Then λ_a is the tension of the electrical current when a potential 0 is applied in s and a potential 1 is applied in t. Indeed, Equation (6) describes then the fact that we are seeking a minimum energy level for our electrical model. Hence, we could build an analogical machine computing the estimator (the existence of such machines had high interest before the appearance of computers...).

Back to the examples : For the first example in the introduction, the matrix $L_{\boldsymbol{w}}$ is

$$\begin{pmatrix} w_1 & -w_1 & 0 \\ -w_1 & w_1 + w_2 & -w_2 \\ 0 & -w_2 & w_2 \end{pmatrix}.$$

System (5) is here

$$\pi_s = 0;\ -w_1\pi_s + (w_1 + w_2)\pi_v - w_2\pi_t = 0;\ \pi_t = 1,$$

i.e.

$$\pi_s = 0;\ \pi_v = \frac{w_2}{w_1 + w_2};\ \pi_t = 1.$$

We get as expected

$$\lambda_{a_1} = \pi_v - \pi_s = \frac{w_2}{w_1 + w_2} \text{ and } \lambda_{a_2} = \pi_t - \pi_v = \frac{w_1}{w_1 + w_2}.$$

For the second exemple, the matrix $L_{\boldsymbol{w}}$ is

$$\begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & 0 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{pmatrix}.$$

System (5) is here

$$\pi_s = 0;\ -\pi_s + 2\pi_u - \pi_t = 0;\ -\pi_s + 2\pi_v - \pi_t = 0;\ \pi_t = 1,$$

i.e.

$$\pi_s = 0;\ \pi_u = \pi_v = 1/2;\ \pi_t = 1.$$

We get as expected

$$\lambda_{a_1} = \pi_u - \pi_s = 1/2;\ \lambda_{a_2} = \pi_t - \pi_u = 1/2;\ \lambda_{a_3} = \pi_v - \pi_s = 1/2;\ \lambda_{a_4} = \pi_t - \pi_v = 1/2.$$

2.2. **Uniqueness of the best linear estimator.** A natural question is whether there is a unique solution for Equation (5). We can give a complete answer.

Proposition 1. *Delete from D' all arcs with a zero variance. We get an new graph D''. There is a unique solution for Equation (5) if and only if we are in one of the following situations:*

- *D'' has exactly one (weakly) connected component.*
- *D'' has exactly two (weakly) connected components, one of them containing s and the other t, in which case E has a zero variance.*

Interestingly, this is a direct consequence of the (all-minors) matrix tree theorem – a beautiful and well-known theorem in graph theory. There is probably a direct proof, but it will be more lengthy.

Proof. The (all-minors) matrix tree theorem [3] states in particular that the determinant of the submatrix $(V \setminus \{s,t\}) \times (V \setminus \{s,t\})$ of the Laplacian matrix $L_{\boldsymbol{w}}$ is $\sum_F \prod_{a\in F} w_a$ where the sum is over all spanning forests F such that

(1) F is the union of 2 disjoint trees
(2) Each tree of F contains either s or t.

(Note that in the reference [3], an oriented version of the Laplacian is used, although here our Laplacian is not oriented). The conclusion is now straightforward. □

2.3. **Sensibility.** In practical situations, to require more precise measures can increase the cost significantly. Hence, to make an optimal choice, it is necessary to know how much the precision of the estimator is improved when the quality of the estimation of the flow is improved on a given arc. In our framework, this can be very precisely evaluated, as shown by the following proposition.

Proposition 2. *Write $W(\boldsymbol{w})$ the optimal variance of the linear estimator, and let $\boldsymbol{\lambda}$ and $\boldsymbol{\pi}$ be solutions of system (5). Fix an arc a. Then*

$$\frac{\partial W(\boldsymbol{w})}{\partial w_a} = \lambda_a^2.$$

Proof. Assume first that system (5) has a unique solution. We see $\boldsymbol{\pi}$ as a function of $\boldsymbol{w}$. Denote $\delta_u := \frac{\partial \pi_u}{\partial w_a}$. We have

$$\frac{\partial W(\boldsymbol{w})}{\partial w_a} = \lambda_a^2 + 2\boldsymbol{\delta} M^T \mathrm{Diag}(\boldsymbol{w}) M \boldsymbol{\pi}.$$

Since $\left(M^T \mathrm{Diag}(\boldsymbol{w}) M\boldsymbol{\pi}\right)_u = 0$ if $u \neq s,t$ and $\delta_s = \delta_t = 0$, the conclusion is straightforward.

To extend the result for $\boldsymbol{w}$ such that system (5) is singular, it is enough to note that such $\boldsymbol{w}$ are at the boundary of the set of $\boldsymbol{w}$ making system (5) non-singular. We get the conclusion by a simple limit argument. □

3. Estimation of any linear combination of flows

3.1. **Starting example.** Let us start with a simple example. Consider the network of Figure 3. The s_1–t_1 flow is to be estimated. There is also a flow going from s_2 to t_2. The following data are available.

$$\begin{aligned}
X_{a_1}^{s_1t_1} &:= \text{ estimate of the } s_1\text{–}t_1 \text{ flow on arc } a_1\text{, variance} = 1\\
X_{a_3}^{s_1t_1} &:= \text{ estimate of the } s_1\text{–}t_1 \text{ flow on arc } a_3\text{, variance} = 1\\
X_{a_2}^{s_2t_2} &:= \text{ estimate of the } s_2\text{–}t_2 \text{ flow on arc } a_2\text{, variance} = 1\\
X_{a_3}^{s_2t_2} &:= \text{ estimate of the } s_2\text{–}t_2 \text{ flow on arc } a_3\text{, variance} = 1\\
S &:= \text{ estimate of the total flow on arc } a_3\text{, variance} = \tfrac{1}{2}
\end{aligned}$$

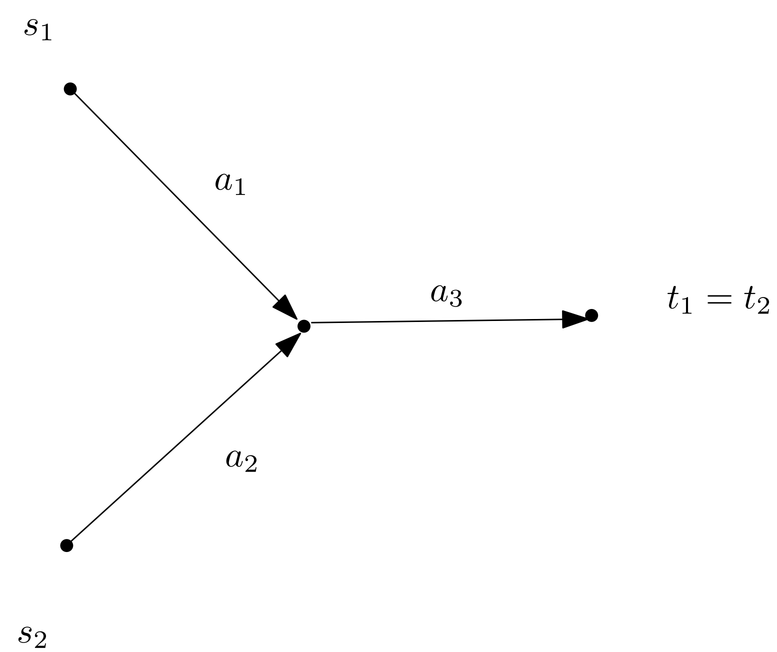

FIGURE 3. A simple case when there are several OD pairs.

Let $E_{s_1t_1}$ be a linear combination of these variables that provides an unbiased estimate of the value of the s_1–t_1 flow. Write

$$E_{s_1t_1} = \lambda_{a_1}^{s_1t_1} X_{a_1}^{s_1t_1} + \lambda_{a_3}^{s_1t_1} X_{a_3}^{s_1t_1} + \lambda_{a_2}^{s_2t_2} X_{a_2}^{s_2t_2} + \lambda_{a_3}^{s_2t_2} X_{a_3}^{s_2t_2} + \lambda^S S.$$

Denote by $x^{s_1t_1}$ (resp. $x^{s_2t_2}$) the real value of the s_1–t_1 flow (resp. s_2–t_2 flow). With this notation, we get

$$\mathbb{E}(E_{s_1t_1}) = (\lambda_{a_1}^{s_1t_1} + \lambda_{a_3}^{s_1t_1} + \lambda^S) x^{s_1t_1} + (\lambda_{a_2}^{s_2t_2} + \lambda_{a_3}^{s_2t_2} + \lambda^S) x^{s_2t_2}.$$

Hence, $E_{s_1t_1}$ is an unbiased estimator of the s_1-t_1 flow if and only if

$$\lambda_{a_1}^{s_1t_1} + \lambda_{a_3}^{s_1t_1} + \lambda^S = 1 \quad \text{and} \quad \lambda_{a_2}^{s_2t_2} + \lambda_{a_3}^{s_2t_2} + \lambda^S = 0.$$

In order to find the best linear estimator, one solves the following program

$$\begin{array}{lcc} & \mathrm{Min}_{\lambda^{s_1t_1},\lambda^{s_2t_2},\lambda^S} & \left(\lambda_{a_1}^{s_1t_1}\right)^2 + \left(\lambda_{a_3}^{s_1t_1}\right)^2 + \left(\lambda_{a_2}^{s_2t_2}\right)^2 + \left(\lambda_{a_3}^{s_2t_2}\right)^2 + \frac{1}{2}(\lambda^S)^2 \\ (8) & \text{s.t.} & \lambda_{a_1}^{s_1t_1} + \lambda_{a_3}^{s_1t_1} + \lambda^S = 1 \\ & & \lambda_{a_2}^{s_2t_2} + \lambda_{a_3}^{s_2t_2} + \lambda^S = 0. \end{array}$$

A direct computation leads to $\lambda_{a_1}^{s_1t_1} = \lambda_{a_3}^{s_1t_1} = \lambda^S = \frac{1}{3}$ and $\lambda_{a_2}^{s_2t_2} = \lambda_{a_3}^{s_2t_2} = -\frac{1}{6}$ and hence a variance of the estimator for the s_1–t_1 flow equal to 1/3. If we used only the measures of the s_1–t_1 flow on the arcs, we would get an estimator with a variance equal to 1/2 (indeed, $\lambda_{a_2}^{s_2t_2} = \lambda_{a_3}^{s_2t_2} = \lambda^S = 0$ in Equation (8)).

We see that it is possible to use the information provided by the estimates of the s_2–t_2 flow to improve the precision of the estimate of the s_1–t_1 flow (in the example, the improvement is 33%).

3.2. **General case.** We have seen in the previous subsection that it is possible to improve the estimation of the value of an s–t flow with the estimates of other flows on the network. We describe now the general method to get the best linear estimator in that case, not only for the value of an s–t flow, but also for any quantity depending linearly from the multiflow.

We suppose that our network is endowed with a multiflow $\left(\boldsymbol{x}^{st}\right)_{st\in\mathcal{ST}}$: for each $st \in \mathcal{ST}$, the vector $\boldsymbol{x}^{st}$ is an s–t flow.

For each pair s–t, we define A^{st} to be the subset of arcs a for which an estimation X_a^{st} of the flow is available. Define similarly A^S to be the subset of arcs a for which an estimation S_a of the total flow on this arc a is available. For these random variables, we denote by w_a^{st} the variance of X_a^{st} and w_a^S the variance of S_a.

This time, we want to find the best linear estimator of any quantity of the type $\sum_{a\in A, st\in\mathcal{ST}} c_a^{st} x_a^{st}$, where $\left(\boldsymbol{c}^{st}\right)_{st\in\mathcal{ST}}$ is a collection of elements in $\mathbb{R}^A$. For instance, putting

$$c_a^{st} = \begin{cases} 0 & \text{if } st \neq \tilde{s}\tilde{t} \text{ or } a \notin \delta^+(\tilde{s}) \cup \delta^-(\tilde{s}) \\ 1 & \text{if } a \in \delta^+(\tilde{s}) \text{ and } st = \tilde{s}\tilde{t} \\ -1 & \text{if } a \in \delta^-(\tilde{s}) \text{ and } st = \tilde{s}\tilde{t} \end{cases}$$

for a given $\tilde{s}\tilde{t} \in \mathcal{ST}$ means that we want to find the best linear estimator of the value of the $\tilde{s}$–$\tilde{t}$ flow, for a given $\tilde{s}$–$\tilde{t}$ pair, using all other measures on the network (as in the previous example).

We can have many other applications. We give two other examples:

$$c_{a'}^{st} = \begin{cases} 0 & \text{if } a' \neq a \text{ or } st \neq \tilde{s}\tilde{t} \\ 1 & \text{if not,} \end{cases}$$

for a given $a \in A$ and $\tilde{s}\tilde{t} \in \mathcal{ST}$ means that we want to find the best linear estimator of the $\tilde{s}$–$\tilde{t}$ flow on arc a, using all other measures on the network.

If c_a^{st} is equal to the length of the arc a, independently of s and t, we will estimate the total distance covered by the users of the network.

Again, let us write down this estimator E in its most general form, assuming that it is a linear estimator. It has necessarily the following form:

$$E = \sum_{a\in A^S} \lambda_a^S S_a + \sum_{st\in\mathcal{ST}} \sum_{a\in A^{st}} \lambda_a^{st} X_a^{st}. \tag{9}$$

As before, it is possible to find a sufficient and necessary condition for E to be an unbiased estimator.

Lemma 2. *E is an unbiased estimator of $\sum_{a\in A, st\in\mathcal{ST}} c_a^{st} x_a^{st}$ if and only if we have for every $st \in \mathcal{ST}$*

$$\sum_{a\in P\cap A^S} \lambda_a^S + \sum_{a\in P\cap A^{st}} \lambda_a^{st} = \sum_{a\in P} c_a^{st}$$

for all $P \in \mathcal{P}_{st}$, and

$$\sum_{a\in C\cap A^S} \lambda_a^S + \sum_{a\in C\cap A^{st}} \lambda_a^{st} = \sum_{a\in C} c_a^{st}$$

for each elementary cycle C.

Proof. Let us write each $\boldsymbol{x}^{st}$ as a conic combination of elementary s–t paths and elementary cycles. As before, by fixing a linear order, there is no ambiguity:

$$\boldsymbol{x}^{st} = \sum_{P\in\mathcal{P}_{st}} x_P^{st} \chi^P + \sum_{C\in\mathcal{C}} x_C^{st} \chi^C.$$

When $P \notin \mathcal{P}_{st}$, we define $x_P^{st} := 0$. With this notation, we have

$$\mathbb{E}(X_a^{st}) = x_a^{st} = \sum_{P\in\mathcal{P}:\, a\in P} x_P^{st} + \sum_{C\in\mathcal{C}:\, a\in C} x_C^{st}$$

and

$$\mathbb{E}(S_a) = \sum_{P\in\mathcal{P}:\, a\in P} \sum_{st\in\mathcal{ST}} x_P^{st} + \sum_{C\in\mathcal{C}:\, a\in C} \sum_{st\in\mathcal{ST}} x_C^{st}.$$

Thus:

$$E \text{ is an unbiased estimator} \quad \iff \quad \mathbb{E}(E) = \sum_{a\in A}\sum_{st\in\mathcal{ST}} c_a^{st}\left(\sum_{P\in\mathcal{P}\colon a\in P} x_P^{st} + \sum_{C\in\mathcal{C}\colon a\in C} x_C^{st}\right).$$

Hence, replacing these expressions in Equation (9), we get

$$\sum_{st\in\mathcal{ST}}\sum_{a\in A} c_a^{st}\left(\sum_{P\in\mathcal{P}\colon a\in P} x_P^{st} + \sum_{C\in\mathcal{C}\colon a\in C} x_C^{st}\right)$$
$$=$$
$$\sum_{st\in\mathcal{ST}}\left(\sum_{a\in A^S}\lambda_a^S\left(\sum_{P\in\mathcal{P}\colon a\in P} x_P^{st} + \sum_{C\in\mathcal{C}\colon a\in C}\sum_{st\in\mathcal{ST}} x_C^{st}\right) + \sum_{a\in A^{st}}\lambda_a^{st}\left(\sum_{P\in\mathcal{P}\colon a\in P} x_P^{st} + \sum_{C\in\mathcal{C}\colon a\in C} x_C^{st}\right)\right)$$

Making an identification of the left and right terms – this expression shall be correct for all possible values of the variables x_P^{st} and x_C^{st} – we get the required equality. $\square$

We can then write down the way for finding the best estimator. With the same notations:

Theorem 2. *Let $\left(\boldsymbol{\lambda}^{st}\right)_{st\in\mathcal{ST}}$ be a collection of elements in $\mathbb{R}^A$ and $\boldsymbol{\lambda}^S$ an element in $\mathbb{R}^A$.*

Then $E := \sum_{a\in A^S}\lambda_a^S S_a + \sum_{st\in\mathcal{ST}}\sum_{a\in A^{st}}\lambda_a^{st}X_a^{st}$ is the unbiased linear estimator of minimal variance if and only if there are $\left(\boldsymbol{\mu}^{st}\right)_{st\in\mathcal{ST}}$ a collection of elements in $\mathbb{R}^A$ and $\left(\boldsymbol{\pi}^{st}\right)_{st\in\mathcal{ST}}$ a collection of elements in $\mathbb{R}^V$ such that

$$\begin{aligned}
\lambda_{(u,v)}^{st} + \lambda_{(u,v)}^S &= \pi_v^{st} - \pi_u^{st} + c_{(u,v)}^{st} \quad \textit{for all } (u,v)\in A \textit{ and all } st\in\mathcal{ST} && \text{(10a)}\\
w_a^S\lambda_a^S + \sum_{st\in\mathcal{ST}}\mu_a^{st} &= 0 \quad \textit{for all } a\in A^S && \text{(10b)}\\
w_a^{st}\lambda_a^{st} + \mu_a^{st} &= 0 \quad \textit{for all } st\in\mathcal{ST} \textit{ and all } a\in A^{st} && \text{(10c)}\\
\sum_{a\in\delta^+(u)}\mu_a^{st} - \sum_{a\in\delta^-(u)}\mu_a^{st} &= 0 \quad \textit{for all } st\in\mathcal{ST} \textit{ and for all } u\in V\setminus\{s,t\} && \text{(10d)}\\
\pi_s^{st} &= 0 \quad \textit{for all } st\in\mathcal{ST} && \text{(10e)}\\
\pi_t^{st} &= 0 \quad \textit{for all } st\in\mathcal{ST} && \text{(10f)}\\
\lambda_a^S &= 0 \quad \textit{for all } a\notin A^S && \text{(10g)}\\
\lambda_a^{st} &= 0 \quad \textit{for all } st\in\mathcal{ST} \textit{ and for all } a\notin A^{st}. && \text{(10h)}
\end{aligned}$$

The system (10) has

$$|A|\times|\mathcal{ST}| + |A| + |A|\times|\mathcal{ST}| + |V|\times|\mathcal{ST}|$$

equations (the first term comes from (10a), the second from (10b) and (10g), the third from (10c) and (10h) and the fourth from (10d), (10e) and (10f)). In total, we have $(|V|+2|A|)\times|\mathcal{ST}|+|A|$ equations. A similar computation leads to $(|V|+2|A|)\times|\mathcal{ST}|+|A|$ unknowns. Hence, it is possible to compute a linear estimator of minimal variance in polynomial time.

Proof of Theorem 2. To find the best linear estimator, we need to find the best choice for the $\boldsymbol{\lambda}^S$ and the $\boldsymbol{\lambda}^{st}$ satisfying the set of constraints provided by Lemma 2. In the present form, the set of constraints has an exponential description. Hence, we have to check if it is possible, as before, to find another description that would be tractable. Lemma 2 means that for each pair $st\in\mathcal{ST}$, there exists a potential $\boldsymbol{\pi}^{st}$ such that

$$\boldsymbol{\lambda}^S + \boldsymbol{\lambda}^{st} - \boldsymbol{c}^{st} = M\boldsymbol{\pi}^{st} \text{ and } \pi_s^{st} = \pi_t^{st} = 0.$$

Our coefficients are therefore solutions of a problem of the form

$$
\begin{aligned}
&\underset{\boldsymbol{\lambda}^{st},\boldsymbol{\lambda}^{S}\in\mathbb{R}^{A},\boldsymbol{\pi}^{st}\in\mathbb{R}^{V}}{\operatorname{Min}} \quad \sum_{a\in A^S} w_a^S\left(\lambda_a^S\right)^2+\sum_{st\in\mathcal{ST}}\sum_{a\in A^{st}} w_a^{st}\left(\lambda_a^{st}\right)^2 \\
&\text{s.t.} \left|\begin{array}{ll}
\boldsymbol{\lambda}^{st}+\boldsymbol{\lambda}^{S}-\boldsymbol{c}^{st}=M\boldsymbol{\pi}^{st} & \text{for each pair } st\in\mathcal{ST} \\
\lambda_a^{st}=0 & \text{for all } a\notin A^{st} \text{ and all } st\in\mathcal{ST} \\
\lambda_a^{S}=0 & \text{for all } a\notin A^{S} \\
\pi_s^{st}=0 & \text{for all } st\in\mathcal{ST} \\
\pi_t^{st}=0 & \text{for all } st\in\mathcal{ST}.
\end{array}\right.
\end{aligned}
\tag{11}
$$

Writing down the Lagrangian of this program, where $2\boldsymbol{\mu}^{st}\in\mathbb{R}^A$ are the Lagrangian multipliers of the first constraint,

$$
\mathcal{L}=\sum_{a\in A^S} w_a^S\left(\lambda_a^S\right)^2+\sum_{st\in\mathcal{ST}}\sum_{a\in A^{st}} w_a^{st}\left(\lambda_a^{st}\right)^2+2\sum_{st\in\mathcal{ST}}\sum_{(u,v)\in A}\mu_{(u,v)}^{st}(\lambda_{(u,v)}^{st}+\lambda_{(u,v)}^{S}-\pi_v^{st}+\pi_u^{st}-c_{(u,v)}^{st}),
$$

a direct differentation according to the different variables leads to the system (10). $\square$

The question of the uniqueness of the optimal estimator is open. As for Theorem 1, there are examples for which the optimal estimator is not unique. Nevertheless, we were neither able to find a necessary and sufficient condition for the uniqueness of the best estimator, nor even to propose a conjecture.

We have a corollary when the flows $\boldsymbol{x}^{st}$ have all estimates on all arcs, in which the Laplacian plays a role again.

Corollary 1. *Assume that $A=A^{st}$ for all $st\in\mathcal{ST}$. Let $\left(\boldsymbol{\lambda}^{st}\right)_{st\in\mathcal{ST}}$ be a collection of elements in $\mathbb{R}^A$, and $\boldsymbol{\lambda}^S$ an element in $\mathbb{R}^A$.*

$E:=\sum_{a\in A^S}\lambda_a^S S_a+\sum_{st\in\mathcal{ST}}\sum_{a\in A^{st}}\lambda_a^{st}X_a^{st}$ is the linear estimator of minimal variance if and only if there is $\left(\boldsymbol{\pi}^{st}\right)_{st\in\mathcal{ST}}$ be a collection of elements in $\mathbb{R}^V$ such that

$$
\begin{aligned}
\left(L_{\boldsymbol{w}^{st}}\boldsymbol{\pi}^{st}\right)_u &= \left(M^T\operatorname{Diag}\left(\boldsymbol{w}^{st}\right)\left(\boldsymbol{\lambda}^S-\boldsymbol{c}^{st}\right)\right)_u \quad \textit{for all } st\in\mathcal{ST} \textit{ and } u\neq s,t && \text{(12a)}\\
w_a^S\lambda_a^S &= \sum_{st\in\mathcal{ST}} w_a^{st}\lambda_a^{st} \quad \textit{for all } a\in A^S && \text{(12b)}\\
\lambda_{(u,v)}^{st}+\lambda_{(u,v)}^{S} &= \pi_v^{st}-\pi_u^{st}+c_{(u,v)}^{st} \quad \textit{for all } (u,v)\in A \textit{ and all } st\in\mathcal{ST} && \text{(12c)}\\
\pi_s^{st} &= 0 \quad \textit{for all } st\in\mathcal{ST} && \text{(12d)}\\
\pi_t^{st} &= 0 \quad \textit{for all } st\in\mathcal{ST} && \text{(12e)}\\
\lambda_a^{S} &= 0 \quad \textit{for all } a\notin A^S && \text{(12f)}
\end{aligned}
$$

Proof. Indeed, according to Equation (10c), if $A=A^{st}$, then $\mu_a^{st}=-w_a^{st}\lambda_a^{st}$ for all $a\in A$. Substituting μ_a^{st} in Equation (10b) leads to Equation (12b). It remains to show that (12a) is satsified. Substituting μ_a^{st} by $-w_a^{st}\lambda_a^{st}$ in Equation (10d) leads to

$$
\left(M^T\operatorname{Diag}\left(\boldsymbol{w}^{st}\right)\boldsymbol{\lambda}^{st}\right)_u=0 \quad \text{for all } st\in\mathcal{ST} \text{ and all } u\in V\setminus\{s,t\}.
$$

Combined with Equation (10a), it leads to (12a).

Conversely, defining $\mu_a^{st}:=-w_a^{st}\lambda_a^{st}$, system (12) leads similarly to system (10): Equation (10d) is obtained by substituting $\boldsymbol{\lambda}^S-\boldsymbol{c}^{st}$ in Equation (12a) by $M\boldsymbol{\pi}^{st}-\boldsymbol{\lambda}^{st}$ (according to Equation (12c)) and then by using the equality $\boldsymbol{\mu}^{st}=-\operatorname{Diag}\left(\boldsymbol{w}^{st}\right)\boldsymbol{\lambda}^{st}$. $\square$

We have another corollary, which can save computations in some cases. It shows that the set of optimal linear estimators of quantities depending linearly from the multiflow is a vector space (of dimension $|A|\times|\mathcal{ST}|$). It can be useful for instance if we know that various quantities depending linearly from the flows have to be estimated. In that case, Corollary 2 below implies that we can compute one and for all the best unbiased linear estimators E_a^{st} of each x_a^{st}, and then get the best linear estimator of these quantities simply by linear combinations of the E_a^{st}.

Corollary 2. *Let E and E' be two unbiased linear estimators of respectively* $\sum_{a\in A, st\in\mathcal{ST}} c_a^{st}x_a^{st}$ *and* $\sum_{a\in A, st\in\mathcal{ST}} c_a'^{st}x_a^{st}$. *Assume that they are both optimal.*

Then for any $\gamma,\gamma'\in\mathbb{R}$, the linear combination $\gamma E+\gamma' E'$ is an optimal linear estimator of $\sum_{a\in A, st\in\mathcal{ST}} (\gamma c_a^{st}+\gamma' c_a'^{st})x_a^{st}$.

Proof. It is a straightforward consequence of the structure of system (10). $\square$

3.3. **The example once again.** Let us know check that we can recover the result of the example above. We want to find the best linear estimator of the value of the s_1–t_1 flow, using all other measures. Using the remark and the notations at the beginning of Subsection 3.2, we have

$$c_{a_1}^{s_1t_1}=1 \text{ and } c_{a_2}^{s_1t_1}=c_{a_3}^{s_1t_1}=c_{a_1}^{s_2t_2}=c_{a_2}^{s_2t_2}=c_{a_3}^{s_2t_2}=0.$$

We also have[1]

$$w_{a_1}^{s_1t_1}=1;\ w_{a_2}^{s_1t_1}=0;\ w_{a_3}^{s_1t_1}=1;\ w_{a_1}^{s_2t_2}=0;\ w_{a_2}^{s_2t_2}=1;\ w_{a_3}^{s_2t_2}=1;\ w_{a_3}^{S}=\frac{1}{2}.$$

Denote the central vertex by u and t_1 and t_2 by t.

Indexing the rows and the columns with s_1,s_2,u,t and a_1,a_2,a_3 in this order, we get

$$L_{\boldsymbol{w}^{s_1t_1}}=\begin{pmatrix}1&0&-1&0\\0&0&0&0\\-1&0&2&-1\\0&0&-1&1\end{pmatrix},\ L_{\boldsymbol{w}^{s_2t_2}}=\begin{pmatrix}0&0&0&0\\0&1&-1&0\\0&-1&2&-1\\0&0&-1&1\end{pmatrix} \text{ and } M=\begin{pmatrix}-1&0&1&0\\0&-1&1&0\\0&0&-1&1\end{pmatrix}.$$

Equation (12a) leads then to

$$2\pi_u^{s_1t}=-\lambda_{a_3}^S-1 \tag{13}$$

and

$$2\pi_u^{s_2t}=-\lambda_{a_3}^S. \tag{14}$$

Equation (12b) leads directly to $\frac{1}{2}\lambda_{a_3}^S=\lambda_{a_3}^{s_1t}+\lambda_{a_3}^{s_2t}$, which can be rewritten with the help of Equation (12c)

$$\frac{5}{2}\lambda_{a_3}^S=-\pi_u^{s_1t}-\pi_u^{s_2t}. \tag{15}$$

Finally, combining Equations (13,14,15), we get

$$\pi_u^{s_1t}=-\frac{2}{3};\ \pi_u^{s_2t}=-\frac{1}{6};\ \lambda_{a_3}^S=\frac{1}{3};\ \pi_{s_1}^{s_1t}=\pi_{s_2}^{s_2t}=\pi_t^{s_1t}=\pi_t^{s_2t}=0.$$

That gives the correct values for $\boldsymbol{\lambda}^{s_1t}$, $\boldsymbol{\lambda}^{s_2t}$ and $\boldsymbol{\lambda}^S$.

[1] If we know for sure that an s–t flow does not use an arc a, we obviously have w_a^{st}=0.

4. Conclusion

On the practical hand, we have described a way to exploit optimally measures of traffic on a network in order to get various informations on this traffic, especially the OD matrix, when the measures are partially redundant. This method appears to be useful in practical cases.

On the theoretical hand, we have discovered a new nice property of the Laplacian of a graph. This adds a new item in a long list of nice properties of this mathematical object. Moreover, a theoretical question remains open, namely the necessary and sufficient condition ensuring the uniqueness of the best estimator in the case when we take into account several origin-destination pairs simultaneously. It would be nice to have something like Proposition 1, in the case when there is a unique OD pair, which is a combinatorial characterization involving the arcs on which estimates are available.

References

1. M. Bierlaire and P. Toint, *An origin-destination matrix estimator that exploits structure*, Transportation Research Part B **29** (1995), 47–60.
2. E. Cascetta and S. Nguyen, *A unified framework for estimating or updating origin/destination matrices from traffic counts*, Transportation Research Part B **22** (1988), 437–455.
3. S. Chaiken, *A combinatorial proof of the all minors matrix tree theorem*, SIAM J. Alg. Disc. Meth. **3** (1982), 319–329.
4. F. Leurent, *Estimer un modèle de distribution spatiale avec des enquêtes en-route*, Modélisation du trafic. Actes INRETS (France) (M. Aron, F. Boillot, and J.-P. Lebacque, eds.), vol. 83, INRETS, 1999, pp. 85–102.
5. F. Leurent and F. Meunier, *Optimal network estimation of origin-destination flow from sufficient link data*, Proc. of the 88th Transportation Research Board Meeting (Washington D.C., USA), January 2009.
6. A. Soule, A. Nucci, R. Cruz, E. Leonardi, and N. Taft, *Estimating dynamic traffic matrices by using viable routing changes*, IEEE/ACM Transactions on Networking (June 2007).
7. H. Spiess, *A maximum likelihhood model for estimating origin-destination matrices*, Transportation Research Part B **21** (1987), 395–412.
8. Y. Vardi, *Estimating source-destination traffic intensities from link data*, Journal of the American Statistical Association **91** (1996), 433.

Université Paris Est, LVMT, ENPC, 6-8 avenue Blaise Pascal, Cité Descartes Champs-sur-Marne, 77455 Marne-la-Valle cedex 2, France.

E-mail address: frederic.meunier@enpc.fr