RESEARCH

Estimation of flow trajectories in a multi-lines transportation network

Guillaume Guex[1*], Romain Loup[2] and François Bavaud[1,2]

*Correspondence: gguex@unil.ch
[1] Department of Language and Information Sciences, University of Lausanne, Lausanne, Switzerland
Full list of author information is available at the end of the article

Abstract

Characterizing a public transportation network, such as an urban multi-lines bus network, requires the origin-destination trip counts during a given period. Yet, if automatic counting makes the embarkment (boarding) and disembarkment (alighting) counts at each bus stop known, it often happens that pedestrian transfers between stops are unknown, and this contribution proposes a three-steps procedure for estimating the missing information, involving maximum entropy and iterative fitting. ** à poursuivre ***

Keywords: multiline bus network; origin-destination flows; boarding and alighting counts; transit flows; maximum entropy estimation

1 Introduction

Transportation networks determine our mobility, require a considerable amount of planning and resources, and elicit much public hopes and critics. They also constitute an endless source of inspiration in formal modeling and optimization, as attested in operations research (classical optimal transportation, maximum flow problem), quantitative geography and spatial econometrics (spatial navigation, multimodality, gravity models for flows), and machine learning (recent developments in regularized optimal transportation, such as color transfer or images interpolation; see e.g. [1]).

This contribution adresses a straightforward, yet central question in public transportation networks: given a network made of many bus lines, how can one estimate the real trips made by the travelers, on the sole basis of the embarkment (boarding) counts and disembarkment (alighting) counts for each bus stop? Although estimating origin-destination flows is a much addressed issue in transportation modeling (see e.g [2] [3] [4] [5] and references therein), the specific problem addressed in this contribution seems, to the best of our knowledge, original.

Pedestrian transfers of travelers between bus lines here constitute the missing information, whose principled evaluation require some methodological reflexion and experimentation. Section 2 introduces the notations and the formalism, as well as the statement of the problem and the iterative solution method, which consist of three consecutive steps: a maximum-entropy computation of the trip distributions, obeying marginal constraints and with a given prior (section 2.4.1); an update of the marginal flows to avoid transfer overflow (section 2.4.3); and an update of the prior distribution (section 2.4.4) by shrinking the components responsible for overflow.

The fist step only is required for solving the single line case (section 2.5), naturally much simpler but yet not trivial, and exhibiting a disembarking probability independent of the embarking stop (Markov property).

Cases studies are presented in section 3 *** à poursuivre ***

2 Notations and formalism

2.1 Lines, stops and junctions

Consider a *transportation network* made of *lines* numbered $\ell = 1,\dots,q$, of respective lengths (number of stops) l_ℓ. Opposite lines, that is parallel lines running in the back and forth directions are considered as distinct.

The $l = \sum_{\ell=1}^{q} l_\ell$ *stops* constitute the nodes of the transportation network. Each stop $i = 1,\dots,l$ belongs to a single line, and defines a unique next or forward stop $F(i)$ (unless i is the line terminus) and a unique backward stop $B(i)$ (unless i is the line start), both on the same line.

Let S_i denotes the *set of stops which can be reached from stop i outside lines connection* (with, e.g., an acceptable walking distance), excluding i itself. A stop i is referred to as an *isolated stop* if $S_i = \emptyset$, and to as a *junction* otherwise.

2.2 Edges, trips, and the incidence matrix

Two sorts of oriented edges are involved in the transportation network:

- *intra-line edges* $(i,j) = (i,F(i))$ belonging to a single line $\ell(i) = \ell(j)$
- *inter-line* or *transfer edges* (i,j) connecting different lines $\ell(i) \neq \ell(j)$, involving walks from junction i to $j \in S_i$. The *set of transfer edges* is denoted by T.

A *st-trip*, noted $[s,t]$, consists of entering into the network at stop s, and leaving the network at t, by following the shortest-path (i.e. achieving the minimum distance, minimum time, or minimum cost), supposed unique, leading to s from t.

The succession of edges (i,j) belonging to the st-trip, noted $(i,j) \in [s,t]$, is unique. Define the *edge-trip incidence matrix* as

$$\chi_{ij}^{st} = \begin{cases} 1 & \text{if } (ij) \in [s,t], \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Note that we can also forbid some aberrant trips across the network, for example, trips $[s,t]$ where (s,t) is a transfer edge (making this trip do not actually use the line network). The *set of permitted trips* across the network is denoted by P, and can be defined regarding some conditions.

2.3 Transportation flows

Let x_{ij} count the *number of travelers using edge* (i,j) in a given period, such as a given hour, day, week or year. The edge flow x_{ij} is denoted by y_{ij} for an intra-line edge (i,j), and by z_{ij} for a transfer edge (i,j). By construction, $x_{ij} = y_{ij} + z_{ij}$, where $y_{ij}\, z_{ij} = 0$.

Let a_i, respectively b_i, the *number of passengers embarking*, respectively *disembarking* at stop i. By construction,

$$\begin{cases} y_{i,F(i)} = a_i \text{ and } b_i = 0 & \text{if } i \text{ is a line start,} \\ y_{B(i),i} = b_i \text{ and } a_i = 0 & \text{if } i \text{ is a line terminus,} \\ y_{i,F(i)} = y_{B(i),i} + a_i - b_i & \text{otherwise.} \end{cases} \tag{2}$$

Also, **a** and **b** must be consistent, in the sense that $A_i \geq B_i$, where A_i (respectively B_i) is the *cumulated number of embarked (resp. disembarked) passengers* on the line under consideration, recursively defined as $A_{F(i)} = A_i + a_i$ (resp. $B_{F(i)} = B_i + b_i$). Moreover, $A_i = B_i$ at a terminal line stop i. This common value yields the total number of passengers transported by the line.

Let the *transportation flow* n_{st} denotes the number of passengers following an st-trip, that is entering the network at s and leaving the network at t by using the shortest-path. One gets from (1)

$$x_{ij} = \sum_{st} \chi_{ij}^{st} n_{st} \tag{3}$$

Among the passengers embarking in i, some transfer from another line, and some others enter into the network:

$$a_i = z_{\bullet i} + n_{i\bullet} \tag{4}$$

where "$\bullet$" denotes the summation over the replaced index, as in $n_{i\bullet} = \sum_{j=1}^{l} n_{ij}$. Similarly, among the passengers disembarking in i, some transfer to another line, and some others leave the network:

$$b_i = z_{i\bullet} + n_{\bullet i} \tag{5}$$

By construction

$$a_\bullet = b_\bullet = z_{\bullet\bullet} + n_{\bullet\bullet}$$

where $n_{\bullet\bullet}$ counts the number of passengers, and $z_{\bullet\bullet}$ counts the number of transfers. $z_{\bullet\bullet}/n_{\bullet\bullet}$ is the average number of transfers per passenger.

As explained in section 2.1, transfers can only occur at junctions, that is $z_{ij} > 0$ implies $(i, j) \in T$. In particular, $z_{ii} = 0$: no traveller is supposed to disembark and re-embark later at the same stop.

2.4 Statement of the problem and solution method

Automatic passenger counters measure the number of passengers entering and leaving lines at each stop [Boyle, 1998], that is **a** and **b**, which provide the basic raw data of the present study, kindly provided by the Lausanne Transportation Agency (tl) for the case study in section 3.3. We will suppose here that this data obeys the necessary consistency condition $a_\bullet^\ell = b_\bullet^\ell$ (where the latter quantities denote the total embarkments and disembarkments on line ℓ), even if, in real case studies, a rescaling must usually be performed to balance in and out-flows on each lines.

Intra-line edge flows $\mathbf{Y} = (y_{ij})$ can be determined by (2), but transfer edge flows $\mathbf{Z} = (z_{ij})$ are, here and typically, unknown. The objective is to estimate the $l \times l$ transportation flow matrix $\mathbf{N} = (n_{st})$. Many consistent solutions coexist in general, even for a single line with no transferts (section 2.5). This issue of incompletely observed data can be tackled by the maximum entropy formalism [6], which has often been the case in transportation modelling researches [7] [8].

Let $f_{st} = n_{st}/n_{\bullet\bullet}$ be the *distribution of st-trips* (empirical distribution) and let g_{st} be some prior guess on its shape (theoretical distribution). Assuming some reasonable initial prior g_{st},

(1) we shall first suppose that the empirical margins $\alpha_s = f_{s\bullet}$ and $\beta_t = f_{\bullet t}$ are known. Then f_{st} can be determined as the maximum entropy solution (section 2.4.1), i.e. as the distribution closest to g_{st} in the Kullback-Leibler divergence sense under the margin constraints, to be calibrated by an iterative fitting inner loop

(2) then (section 2.4.3), the margins will be updated to $\tilde{\alpha}_s$ and $\tilde{\beta}_t$ by requiring a *minimum proportion* $\theta \in (0,1)$ of passengers entering/leaving the network at each stop, as well as avoiding transfer overflow exceeding the embarking and disembarking counts at each stop

(3) finally (section 2.4.4), the prior will be updated to $\tilde{g}_{st}$ by shrinking, if necessary, the priors g_{st} associated to overflows.

With the new prior distribution $\widetilde{g}_{st}$ and the new margin distributions $\widetilde{\alpha}_s$, $\widetilde{\beta}_t$, we can iterate the the above steps, until convergence. The only free parameter is θ, whose effect is studied on toy examples in section 3.2.

The above iterative solution method is somewhat reminiscent of the EM algorithm. As a matter of fact, the first step (maximum entropy) exactly correspond to the "expectation step" of the EM algorithm (see e.g. [9] [10]), but steps two and three, aiming at calibrating parameters α_s, β_t and g_{st}, do not follow the maximum likelihood rationale of the "maximisation step". Pseudocode of the algorithm is shown with Algorithm 1.

2.4.1 Maximum entropy estimate of st-trips

As announced, the proportion of st-trips $f_{st} = n_{st}/n_{\bullet\bullet}$ (empirical distribution) will be estimated from some prior guess g_{st} (theoretical distribution) and margin constraints α_s and β_t for f_{st} by maximum entropy, i.e. by solving the problem

$$\begin{aligned}
&\min_{\mathbf{f}\in\mathscr{F}} \sum_{st} f_{st} \log \frac{f_{st}}{g_{st}}, \\
&s.t. \sum_{t} f_{st} = \alpha_s, \\
&\quad\;\; \sum_{s} f_{st} = \beta_t.
\end{aligned} \tag{6}$$

The Lagragian is

$$L = \sum_{st} f_{st} \log \frac{f_{st}}{g_{st}} - \sum_{s} \lambda_s(\alpha_s - \sum_{t} f_{st}) - \sum_{t} \mu_t(\beta_t - \sum_{s} f_{st}),$$

which gives, after deriving and setting to zero,

$$f_{st} = \phi_s \psi_t g_{st} \qquad \text{with } \phi_s := \exp(-1-\lambda_s),\ \psi_t := \exp(-\mu_t). \tag{7}$$

Using constraints in (6), we find

$$\phi_s = \frac{\alpha_s}{\sum_t \psi_t g_{st}}, \qquad \psi_t = \frac{\beta_t}{\sum_s \phi_s g_{st}}, \tag{8}$$

which yields the following *iterative fitting algorithm*: starting with some $\psi_t^{(0)} > 0$, one performs the iteration

$$\phi_s^{(\iota)} = \frac{\alpha_s}{\sum_t \psi_t^{(\iota)} g_{st}}, \qquad \psi_t^{(\iota+1)} = \frac{\beta_t}{\sum_s \phi_s^{(\iota)} g_{st}}, \tag{9}$$

until convergence to ϕ_s and ψ_t obeying (8).

In view of (4) and (5), the postulated margins must satisfy, for each isolated stop i

$$\alpha_i = \frac{a_i}{n_{\bullet\bullet}} \qquad \beta_i = \frac{b_i}{n_{\bullet\bullet}} \tag{10}$$

permitting to determine the total flow as $n_{\bullet\bullet} = \frac{a_i}{\alpha_i}$, or $n_{\bullet\bullet} = \frac{b_i}{\beta_i}$ for any isolated stop i, and thus the *st*-flow itself as

$$n_{st} = n_{\bullet\bullet} f_{st} = n_{\bullet\bullet} \phi_s \psi_t g_{st} \tag{11}$$

whose plugging into (3) yields the intra-line edge flows $\mathbf{Y} = (y_{ij})$ and the transfer edge flows $\mathbf{Z} = (z_{ij})$.

2.4.2 Initialization of the prior and the margins

The geometry of the network permits to define the set of permitted *st*-trips across the network denoted by P. The initial prior was chosen as the uniform distribution on admissible paths, that is as

$$g_{st} = \begin{cases} \frac{1}{|P|} & [s,t] \in \mathrm{P}, \\ 0 & \text{otherwise.} \end{cases}$$

The initial margins were chosen as those of the flow without transfer, namely $\alpha_i = \frac{a_i}{a_\bullet}$ and $\beta_i = \frac{b_i}{b_\bullet}$ for all stops.

2.4.3 Updating the margin distributions

Let us define the hyperparameter $\theta \in (0,1)$ as the *minimum proportion of passengers (among a_i and b_i) entering/leaving the network at each stop*, that is $n_{s\bullet} \geq \theta a_s$ and $n_{\bullet t} \geq \theta b_t$. Note that we could set a different hyperparameter for each node, and differing for embarkments and disembarkments, but without addition information, we will restrain to this simpler case. Identities (4) and (5) then imply the inequalities

$$z_{\bullet s} \leq (1-\theta) a_s \qquad\qquad z_{t\bullet} \leq (1-\theta) b_t$$

the violation of which constitutes transfer overflow. Hence requiring a maximal transfer yet avoiding overflow can be granted with the following updating of margins

$$\widetilde{\alpha}_s = \frac{\min(\theta a_s, a_s - z_{\bullet s})}{\sum_{s'} \min(\theta a_{s'}, a_{s'} - z_{\bullet s'})} \qquad\qquad \widetilde{\beta}_t = \frac{\min(\theta b_t, b_t - z_{t\bullet})}{\sum_{t'} \min(\theta b_{t'}, b_{t'} - z_{t'\bullet})} \quad . \tag{12}$$

2.4.4 Updating the prior distribution

Overflow occurs in transfer edge (i,j) if $z_{i\bullet} > (1-\theta)b_i$ or $z_{\bullet j} > (1-\theta)a_j$. To avoid it, components g_{st} of the prior distribution will be shrinked by a suitable ratio whenever edge flows $(i,j) \in [s,t]$ exhibit overflow. For any edge (i,j), let us compute the *flow ratio* r_{ij} as

$$r_{ij} = \max\left(1, \frac{z_{i\bullet}}{(1-\theta)b_i}, \frac{z_{\bullet j}}{(1-\theta)a_j}\right) \geq 1 \quad , \tag{13}$$

where $r_{ij} > 1$ denotes an overflow through edge (i,j). For a given origin-destination $[s,t]$, define the *orgin-destination flow ratio* $\bar{r}_{st}$ as the largest r_{ij} among edge flows $(i,j) \in [s,t]$, that is as

$$\bar{r}_{st} = \max_{ij} \chi_{ij}^{st} r_{ij} \geq 1 \quad . \tag{14}$$

By construction, $\bar{r}_{st} > 1$ denotes an overflow on some transfer edge between s and t. To adjust the flow, we shall divide the previous flow by this ratio

$$\widetilde{n}_{st} = \frac{n_{st}}{\bar{r}_{st}} \tag{15}$$

and define the new prior distribution as

$$\widetilde{g}_{st} = \frac{\left(\frac{\widetilde{n}_{st}}{\phi_s \psi_t}\right)}{\sum_{s',t'} \left(\frac{\widetilde{n}_{s',t'}}{\phi_{s'} \psi_{t'}}\right)} \quad . \tag{16}$$

where ϕ_s and ψ_t are the values (8) obtained in the previous maximum entropy step.

Algorithm 1 Compute the transportation flow matrix $\mathbf{N} = (n_{st})$ knowing the edge-trip incidence matrix $\boldsymbol{\chi} = (\chi_{ij}^{st})$, the set of transfer edges T, the set of permitted trips P, the embarking flow $\mathbf{a}$, the disambarking flow $\mathbf{b}$, the index of an isolated source node $\tilde{s}$, and the minimum proportion of passengers entering/leaving the network θ.

$g_{st} \leftarrow I([s,t] \in P)/|P|, \forall s,t$ ▷ Initialize the prior distribution
$\alpha_s \leftarrow a_s/a_\bullet, \forall s$ ▷ Initialize the network ingoing distribution
$\beta_t \leftarrow b_t/b_\bullet, \forall t$ ▷ Initialize the network outgoing distribution
$\varepsilon \leftarrow 10^{-40}$ ▷ Fix a small quantity
while $\mathbf{N} = (n_{st})$ has not converge **do** ▷ Main loop
    $\psi_t \leftarrow 1, \forall t$
    while $\boldsymbol{\psi} = (\psi_t)$ has not converge **do** ▷ Iterative fitting loop
        $\phi_s \leftarrow \alpha_s/(\sum_t \psi_t g_{st} + \varepsilon), \forall s$
        $\psi_t \leftarrow \beta_t/(\sum_s \phi_s g_{st} + \varepsilon), \forall t$
    end while
    $n_{st} \leftarrow \frac{a_{\tilde{s}}}{\alpha_{\tilde{s}}} \phi_s \psi_t g_{st}, \forall s,t$ ▷ Compute the transportation flow
    $z_{ij} \leftarrow I((i,j) \in T) \sum_{st} \chi_{ij}^{st} n_{st}, \forall i,j$
    $\alpha_s \leftarrow \frac{\min(\theta a_s, a_s - z_{\bullet s})}{\sum_{s'} \min(\theta a_{s'}, a_{s'} - z_{\bullet s'})}, \forall s$ ▷ Update the network ingoing distribution
    $\beta_t \leftarrow \frac{\min(\theta b_t, b_t - z_{t\bullet})}{\sum_{t'} \min(\theta b_{t'}, b_{t'} - z_{t'\bullet})}, \forall t$ ▷ Update the network outgoing distribution
    $r_{ij} \leftarrow \max\left(1, \frac{z_{i\bullet}}{(1-\theta)b_i}, \frac{z_{\bullet j}}{(1-\theta)a_j}\right), \forall i,j$
    $\bar{r}_{st} \leftarrow \max_{ij} \chi_{ij}^{st} r_{ij}, \forall s,t$
    $\widetilde{g}_{st} \leftarrow \frac{\left(\frac{n_{st}}{\phi_s \psi_t \bar{r}_{st} + \varepsilon}\right)}{\sum_{s',t'} \left(\frac{n_{s't'}}{\phi_{s'} \psi_{t'} \bar{r}_{s't'} + \varepsilon}\right) + \varepsilon}, \forall s,t$ ▷ Update the prior distribution
end while
return $\mathbf{N} = (n_{st})$

2.5 Markov property for a single line

A "network" made of a single line contains no transfers, and flow estimates can be obtained at once by the maximum entropy step only.

Let $i = 1, \ldots, l$ enumerate the bus stops in increasing order, i.e. $F(i) = i+1$. The initial prior is simply $g_{st} = c\, I(s<t)$ and captures solely the unidirectional nature of trips, where $I(.)$ denotes the 0/1 indicator function and $c = \frac{1}{(l-1)(l-2)}$. The margins of the empirical distribution f_{st}, as well as the total flow, are here known :

$$\alpha_s = \frac{a_s}{a_\bullet} \qquad\qquad \beta_t = \frac{b_t}{b_\bullet} \qquad\qquad n_{\bullet\bullet} = a_\bullet = b_\bullet \quad .$$

Following (7) maximum entropy flows are of the form

$$n_{st} = n_{\bullet\bullet}\, c\, I(s<t)\, \phi_s\, \psi_t \tag{17}$$

where (setting $\Psi_s := \sum_{t>s} \psi_t$ and $\Phi_t := \sum_{s<t} c\phi_s$) the constraints (8) equivalently read

$$\phi_s = \frac{\alpha_s}{c \sum_{t>s} \psi_t} = \frac{a_s}{n_{\bullet\bullet}\, c\, \Psi_s} \qquad\qquad \psi_t = \frac{\beta_t}{c \sum_{s<t} \phi_s} = \frac{b_t}{n_{\bullet\bullet}\, c\, \Phi_t} \tag{18}$$

to be solved by iterative fitting.

Interestingly enough, the form (17) for the flows is reminiscent of the *gravity flows* of quantitative Geography [7] [8] [11] [12], where ϕ_s is the *push factor*, ψ_t is the *pull factor*, and $I(s<t)$ the *distance deterrence function*. Yet, instead of being symmetric in s,t and decreasing with the distance $|s-t|$, the distance deterrence function is here asymmetric due to the line orientation, but otherwise constant.

This constancy entails the following Markovian behaviour for flows: let m_{st} be the number of travelers embarking at stop s and still inside the bus at stop $t > s$, and let ρ_{st} the probability that travelers embarking at s will disembark at t. By (17),

$$m_{st} = \sum_{u \ge t} n_{su} = n_{\bullet\bullet}\, c\, \phi_s \sum_{u \ge t} I(s<u)\, \psi_u = n_{\bullet\bullet}\, c\, \phi_s (\psi_t + \Psi_t)$$

The empirical estimate of ρ_{st} is given by the proportion, among the travelers embarking at s and still present at $t > s$, of travelers disembarking at t, that is

$$\rho_{st} = \frac{n_{st}}{m_{st}} = \frac{n_{\bullet\bullet}\, c\, \phi_s\, \psi_t}{n_{\bullet\bullet}\, c\, \phi_s (\psi_t + \Psi_t)} = \frac{\psi_t}{\psi_t + \Psi_t} \le 1$$

which depends on t only: it appears that the disembarkment probability $\rho_t = \frac{\psi_t}{\psi_t + \Psi_t}$ at t is *independent* of the embarkment stop s. Said otherwise, a traveler embarking at any stop s (and thus necessarily in the bus at $F(s) = s+1$) experiences the *same disembarkment probability* at each further stop $t > s$.

This Markov property, enjoyed by maximum-entropic flows, contrasts other possible solutions, such as the "first in, first out" (FIFO) flows (homogenizing the traveled distances among users) or the "last in, first out" (LIFO) flows (tending to generate maximally contrasted traveled distances).

3 Case Studies

Case studies are divided in two sections. In the first section, we test the algorithm on toy examples, which are artificial networks where the transportation flow n_{st} is randomly drawn. These examples enable some kind of validation of the algorithm, as "real" transportation flows are known and can be compared to the solutions given by our method. This setup differs from the second section, which is dedicated to applying the algorithm to the real case of the public transportation network of the city of Lausanne (tl), where embankment and disembarkment flows are measured but transportation flows are unknown. This second case study shows that the algorithm is applicable on large, real datasets and can gives insights about passengers probable routes in the network.

3.1 Error measurements

In all case studies, we obtain a estimation of the transportation flow with the algorithm, noted $\mathbf{N} = (n_{st})$, starting from the real embarkment flow $\mathbf{a}_{\text{ref}}$ and disembarkment flow $\mathbf{b}_{\text{ref}}$. In toy examples, we also have access to the real transportation flow $\mathbf{N}_{\text{ref}}$. There are two types of dissimilarity measures between the data and the solution proposed by the algorithm: (1) if we have access to $\mathbf{N}_{\text{ref}}$, how much $\mathbf{N}$ differs from it, and (2) how well constraints defined by $\mathbf{a}_{\text{ref}}$ and $\mathbf{b}_{\text{ref}}$ are respected. The first dissimilarity is measured through the *mean transportation error*, denoted by $\text{MTE}(\mathbf{N})$, and computed with

$$\text{MTE}(\mathbf{N}) = \sum_{st} \frac{n_{st}^{\text{ref}}}{n_{\bullet\bullet}^{\text{ref}}} \frac{|n_{st} - n_{st}^{\text{ref}}|}{n_{st}^{\text{ref}}} = \frac{\sum_{st} |n_{st} - n_{st}^{\text{ref}}|}{n_{\bullet\bullet}^{\text{ref}}} \tag{19}$$

and the second one with the *mean margin error*, noted $\text{MME}(\mathbf{N})$, obtained with

$$\begin{aligned}\text{MME}(\mathbf{N}) &= \frac{1}{2} \sum_i \frac{a_i^{\text{ref}}}{a_\bullet^{\text{ref}}} \frac{|z_{\bullet i} + n_{i\bullet} - a_i^{\text{ref}}|}{a_i^{\text{ref}}} + \frac{1}{2} \sum_i \frac{b_i^{\text{ref}}}{b_\bullet^{\text{ref}}} \frac{|z_{i\bullet} + n_{\bullet i} - b_i^{\text{ref}}|}{a_i^{\text{ref}}} \\ &= \frac{\sum_i (|z_{\bullet i} + n_{i\bullet} - a_i^{\text{ref}}| + |z_{i\bullet} + n_{\bullet i} - b_i^{\text{ref}}|)}{2 n_{\bullet\bullet}^{\text{ref}}} \end{aligned} \tag{20}$$

where $z_{ij} = I((i,j) \in T) \sum_{st} \chi_{ij}^{st} n_{st}$ is the flow on transfer edges T.

Note that, by construction, the MME should be null when the algorithm converges. However, it can be informative to track down this error along iterations and, in some practical cases where the network is large, the algorithm convergence criterion is reached without having margin constraints perfectly respected.

3.2 Toy Examples

3.2.1 Construction

All constructed toy examples are built following the same approach, which aims at being simple but somewhat realistic. We fix a number of *line tours* $p \geq 2$, each of which is constituted of a forward line and a backward line, for a total of $q = 2p$ lines. Every line has a starting and ending node, which are isolated node, and posses $p-1$ intermediary nodes which allows transfers to the other tour lines, giving a total of $n = 2p(p+1)$ nodes in the network. Examples of these toy networks can be found in Figure 1.

In order to be realistic, permitted st-trips set P is constructed considering all shortest-path between pair of nodes, excluding :

- s and t that are on the same line but with t preceding s in the line order.

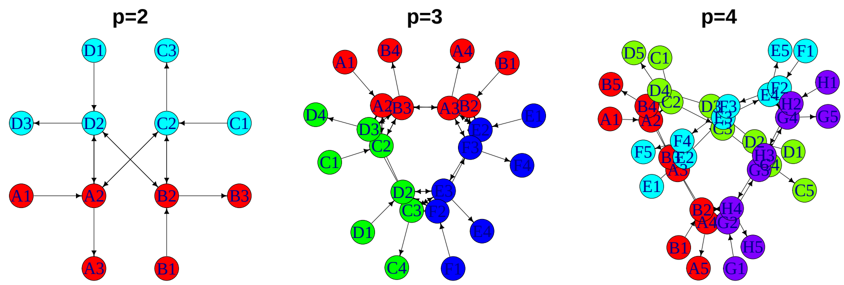

Figure 1: 3 toy examples, with the number of tours $p \in \{2,3,4\}$. Tours are displayed in the same color, line have a unique letter, and each stop a unique combination of a letter and a number. Clusters of nodes represent positions where transfer between tours are possible.

- s and t that are on the same tour but opposite line.
- s and t whose shortest-path starts with a transfer edge, ends with a transfer edge, or possesses two (or more) consecutive transfer edges.

A transportation flow $\mathbf{N}_{\text{ref}} = (n_{st}^{\text{ref}})$ is drawn by setting a fixed number of passengers $n_{\bullet\bullet}^{\text{ref}}$, and each passenger is assigned randomly to a (s,t) pair drawn uniformly among P. From this reference transportation flow $\mathbf{N}_{\text{ref}}$, using the edge-trip incidence matrix $\boldsymbol{\chi}$ and equation (3), we can compute flow on edges $\mathbf{X}_{\text{ref}}$ and, in turn, the number of passengers embarking $\mathbf{a}_{\text{ref}}$ and the number of passengers disembarking $\mathbf{b}_{\text{ref}}$ at each stop.

3.2.2 Algorithm iterations

First, we shows some algorithm iterations on a toy example with $p = 2$, where 100 passengers where drawn uniformly across the $|P| = 20$ possible st-trips. Some iterations of the algorithm with $\theta = 0.1$, along with MTE and MME errors, are shown in Figure 2. On this small example, we can see that the algorithm quickly find an estimation giving small MME error, but still give an MTE of 0.252. This result is due to the fact that only 100 passengers are drawn, giving a large deviation compared to the optimally found solution which maximize the entropy. Interestingly, we see that a better result is found on iteration 6, but margins constraints at this point are not perfectly respected yet.

3.2.3 MTE study

The main goal of toy examples, since we have access to the real transportation flow, is to study how the resulting MTE behave regarding passenger st-trips distribution and hyperparameter θ on different network sizes.

By running the algorithm for line tours $p \in \{2,3,4,5,6,7,8\}$, we observe that MTE depends on different parameters. On the graph on the left in Figure 3, the number of passengers inserted in the network has an influence on MTE. The higher the number of passengers, the smaller the error is because passengers are better distributed throughout the larger network. On the right in Figure 3, the aim is to understand how the proportion $\theta \in (0,1)$ responds to different line tours. Dots represent local minimum MTE for each tour.

3.3 Real Data

after some preliminary, undocumented corrections (i.e. the components of $\mathbf{a}$ and $\mathbf{b}$ can be non-integer). It may also happen that, on some lines ℓ, raw data do not obey the necessary

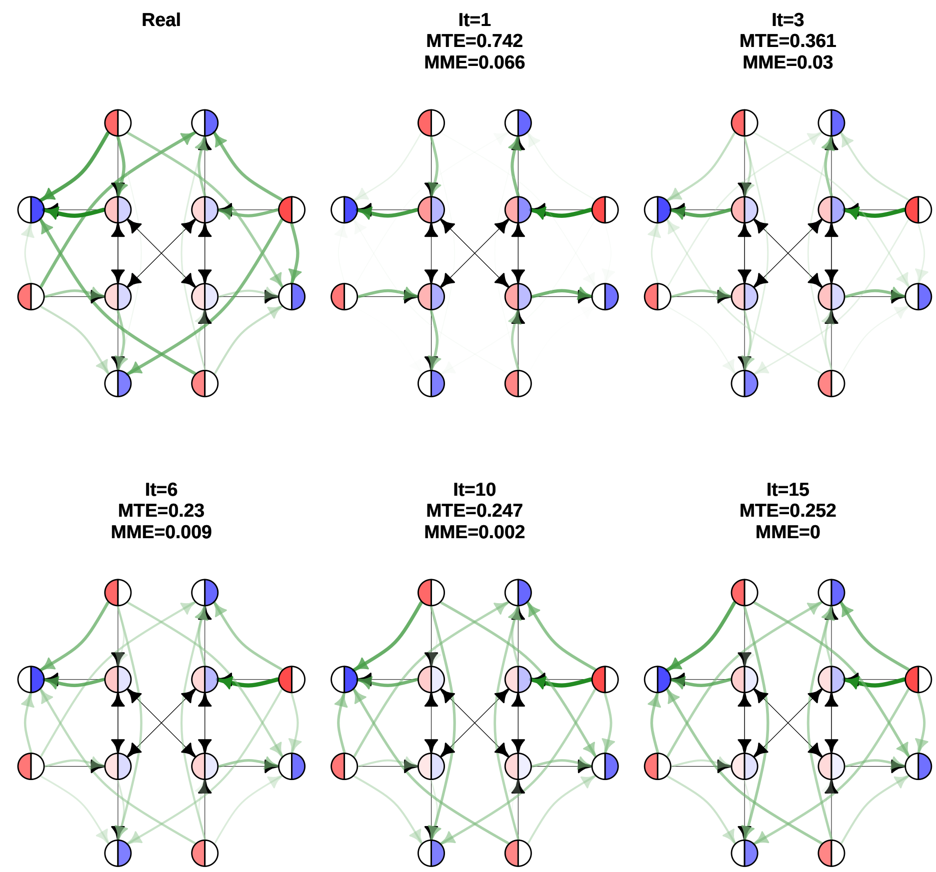

Figure 2: Real transportation flow (green arrows) obtained by randomly drawing 100 passengers on the graph toy example with $p = 2$ line tours, along with iterations $1,3,6,10,15$ of the algorithm with $\theta = 0.1$. MTE and MME errors are computed, and embarkment and disembarkment flows are represented respectively by the red and blue colors on nodes.

consistency condition $a^\ell_\bullet = b^\ell_\bullet$ (where the latter quantities denote the total embarkments and disembarkments on line ℓ), in which case we did rescale the embarking and disembarking line counts as

$$\hat{a}_i = (1 - \frac{a^\ell_\bullet - b^\ell_\bullet}{a^\ell_\bullet + b^\ell_\bullet})\, a_i \qquad\qquad \hat{b}_i = (1 + \frac{a^\ell_\bullet - b^\ell_\bullet}{a^\ell_\bullet + b^\ell_\bullet})\, b_i$$

ensuring $\hat{a}^\ell_\bullet = \hat{b}^\ell_\bullet = 2a^\ell_\bullet b^\ell_\bullet/(a^\ell_\bullet + b^\ell_\bullet)$. However, strongly unbalanced lines such that $|a^\ell_\bullet - b_\bullet|^\ell/a^\ell_\bullet > 0.3$ or $|a^\ell_\bullet - b^\ell_\bullet|/b^\ell_\bullet > 0.3$ (which always turned out to be temporary lines with small counts) were simply disregarded and line ℓ removed from the network.

Also, the geometry of the network permits to derive the edge-trip incidence matrix $\boldsymbol{\chi}$ defined in (1).

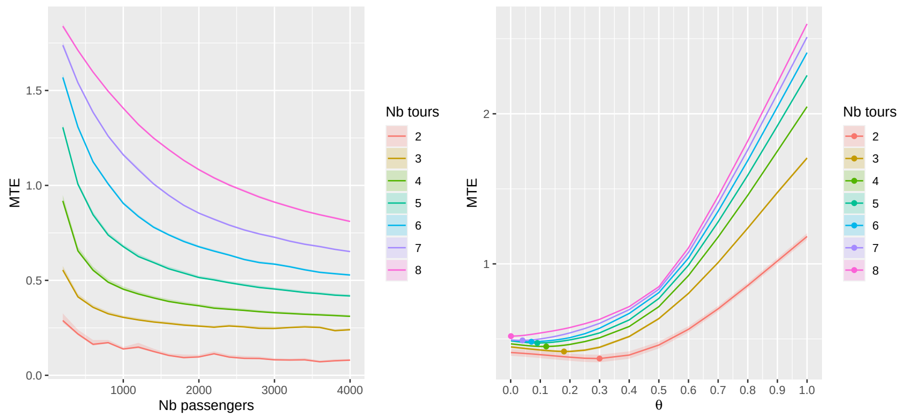

Figure 3: *Mean transportation error* (MTE), according to the number of tours $p \in \{2,3,4,5,6,7,8\}$ and resp. to the number of passengers into the network per line tour and the minimum proportion $\theta = 0.1$ of passengers entering/leaving the network at each stop with a proportional number of passengers to the shortest paths of each line tour.

3.3.1 Dataset construction

The original dataset used in this contribution consists of over 13 million rows of data over more than 1200 stop on 35 bus lines. Each data row represents a stop by a bus to embark and disembark passengers. The dataset used for further analysis is summed by stop over the year 2019.

4 Conclusion

Appendix

Text for this section…

Acknowledgements
Text for this section…

Funding
Text for this section…

Abbreviations
Text for this section…

Availability of data and materials
Text for this section…

Ethics approval and consent to participate
Text for this section…

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Text for this section…

Authors' contributions
Text for this section …

Authors' information
Text for this section…

Author details
[1]Department of Language and Information Sciences, University of Lausanne, Lausanne, Switzerland. [2]Institute of Geography and Sustainability, University of Lausanne, Lausanne, Switzerland.

References

1. Peyré, G., Cuturi, M., *et al.*: Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning **11**(5-6), 355–607 (2019)
2. Bell, M.G., Lida, Y.: Transportation Network Analysis. Wiley, Chichester (1997)
3. Hazelton, M.L.: Estimation of origin–destination matrices from link flows on uncongested networks. Transportation Research Part B: Methodological **34**(7), 549–566 (2000)
4. Ashok, K., Ben-Akiva, M.E.: Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows. Transportation science **36**(2), 184–198 (2002)
5. Cui, A.: Bus passenger origin-destination matrix estimation using automated data collection systems. PhD thesis, Massachusetts Institute of Technology (2006)
6. Jaynes, E.T.: Information theory and statistical mechanics. Physical review **106**(4), 620 (1957)
7. Wilson, A.: A statistical theory of spatial distribution models. Transportation Research **1**(3), 253–269 (1967)
8. Erlander, S., Stewart, N.F.: The Gravity Model in Transportation Analysis: Theory and Extensions vol. 3. VSP, Leiden (1990)
9. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: series B (methodological) **39**(1), 1–22 (1977)
10. Bavaud, F.: Information theory, relative entropy and statistics. In: Sommaruga, G. (ed.) Formal Theories of Information: From Shannon to Semantic Information Theory and General Concepts of Information. LNCS, vol. 5363, pp. 54–78. Springer, Berlin (2009)
11. Bavaud, F.: The quasi-symmetric side of gravity modelling. Environment and Planning A **34**(1), 61–79 (2002)
12. Thomas-Agnan, C., LeSage, J.P.: In: Fischer, M.M., Nijkamp, P. (eds.) Spatial Econometric OD-Flow Models, pp. 2179–2199. Springer, Berlin, Heidelberg (2021)

Figures

Figure 4: Sample figure title

Figure 5: Sample figure title

Table 1: Sample table title. This is where the description of the table should go

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.

Tables

Additional Files

Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.