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 bus lines numbered $\ell = 1, \ldots, q$, 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 $l = \sum_{\ell=1}^{q} l_\ell$ bus stops constitute the nodes of the transportation network. Each stop $i = 1, \ldots, l$ belongs to *a single bus 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 denote the set of stops which can be reached from stop i within 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 Line edges, transfer edges and trips

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 $j \in S_i$.

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 (ij) belonging to the st-trip, noted $(ij) \in [s,t]$, is unique. Define the edge-trip incidence matrix as

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

A st-trip always starts with the edge $(s, F(s))$, and finishes with $(B(t), t)$. Transfers can occur in-between, but never at the beginning nor at the end of the trip.

2.3 Transportation flows

Let x_{ij} count the number of travelers using edge (ij) 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, $\mathbf{a}$ and $\mathbf{b}$ 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} denote 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 $j \in S_i$. 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 buses at each stop [Boyle, 1998], that is $\mathbf{a}$ and $\mathbf{b}$, which provide the basic raw data of the present study, kindly provided by the Lausanne Transportation Agency (tl), after some preliminary, undocumented corrections (i.e. the components of $\mathbf{a}$ and $\mathbf{b}$ can be non-integer). It may also happen that raw data do not obey the necessary consistency condition $a_\bullet = b_\bullet$, in which case we did rescale the embarking and disembarking counts as

$$\hat{a}_i = \left(1 - \frac{a_\bullet - b_\bullet}{a_\bullet + b_\bullet}\right) a_i \qquad \hat{b}_i = \left(1 + \frac{a_\bullet - b_\bullet}{a_\bullet + b_\bullet}\right) b_i$$

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

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

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 $\mathbf{N} = (n_{st})$. 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], and has been often in transportation modelling [7] [8].

Let $f_{st} = n_{st}/n_{\bullet\bullet}$ be the proportion 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, which will be discussed in section *** .

*** 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 the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{g}$, do not follow the maximum likelihood rationale of the "maximisation step".

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

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

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 rule out forbidden *st*-paths, i.e. obeying $\chi_{\bullet\bullet}^{st} = 0$. Such forbidden *st*-paths consist of *** @G? donner une liste exhaustive de critères d'exclusion ici ? *** The initial prior was chosen as the uniform distribution on the remaining $c < l^2$ admissible *st*-paths, *** @G: que vaut l? que vaut c? ***, that is as

$$g_{ij} = \begin{cases} \frac{1}{c} & \text{if } \chi_{\bullet\bullet}^{st} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

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

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

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

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

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.

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 :

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

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),

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

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

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

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).

*** @G, ici davantage de développements sur l’approche Markovienne, ainsi que le joli schéma ? ***

*** @G, @R, ici la (les) figure de l’exemple “starting from Maladière, Riant-Cour, Dapples”... ? ***

2.6 Algorithm pseudo-code

Algorithm 1 Compute the transportation flow matrix $\mathbf{N} = (n_{st})$ knowing the edge-trip incidence matrix $\boldsymbol{\chi} = (\chi^{st}_{ij})$, 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 hyperparameter θ.

$g_{st} \leftarrow I(\chi^{st}_{\bullet\bullet} > 0)/\sum_{s't'} I(\chi^{s't'}_{\bullet\bullet} > 0)$, $\forall s,t$ ▷ Initialize the prior distribution.
$\alpha_s \leftarrow a_s/a_\bullet$, $\beta_t \leftarrow b_t/b_\bullet$, $\forall s,t$ ▷ Initialize the margin distributions.
$\varepsilon \leftarrow 10^{-40}$ ▷ Fix a small quantity
while $\mathbf{N} = (n_{st})$ has not converge **do** ▷ Main loop.
    $\psi_t \leftarrow 1$, $\forall t$ ▷ Initialize the lagragian multiplier for destinations.
    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.
end while
return $\mathbf{Z}$

3 Case Studies

3.1 Toy Examples

3.2 Real Data

4 Conclusion

Previous submission to complex networks 2022

Statement of the problem

Automatic passenger counters measure the number of passengers entering and leaving buses at each stop [13]. Given this information, can we estimate the complete trajectories of passengers within the entire multi-line network? This communication attempts to propose an estimation of all passenger trajectories in the multi-line network with an algorithm based on iterative proportional fitting (IPF) [14].

The exploited dataset is provided by the Lausanne Transportation Agency (tl) in Switzerland. The dataset includes 42 lines of buses (or subways) and more than 1361 stops, including 497 clusters of stops (Fig. **??**), carrying around 115 million passengers in 2019. Each stop refers to a single directed line, and return lines are considered as distinct. In addition to *line edges*, it is possible to construct pedestrian *transfer edges* (Fig. **??**) to make the graph unilaterally connected by considering, e.g., clusters of stops connected.

Knowing only the network structure and the number of passengers embarking and disembarking at each stop, how can we infer the most probable passenger trajectories in the network? Before examining the general multi-line network problem, we address the estimation of trajectories on a single line.

5 Single line

Consider a one-directional line with n stops indexed regarding the order found in the line. Let $\mathbf{a}_{\text{in}} = (a_s^{\text{in}})$ and $\mathbf{a}_{\text{out}} = (a_t^{\text{out}})$ be two vectors representing, respectively, the passengers entering and leaving the line at each stop. The goal is to estimate the $(n \times n)$ origin-destination matrix $\mathbf{N} = (n_{st})$, where n_{st} represents the number of passengers entering the line at s and leaving at t, subject to constraints $n_{s\bullet} = a_s^{\text{in}}$ and $n_{\bullet t} = a_t^{\text{out}}$. Among many feasible solutions, arguably the most elegant one is the maximum entropy solution, which can be derived from different means: (i) passenger flows can be modelled using a Markovian assumption, which translates by assuming that every passenger has the same probability to continue the trip after having travelled at last one stop ; (ii) an iterative proportional fitting algorithm can be performed, starting with an initial origin-destination affinity matrix $\mathbf{S} = (s_{st})$, defined as the upper triangular $n \times n$ matrix filled with 1, and then iterated to satisfy the margin constraints given by $\mathbf{a}_{\text{in}}$ and $\mathbf{a}_{\text{out}}$. Both approaches give the same solution, but only the latter remains pertinent in the multi-line problem.

6 Multi-lines

In the multi-line problem, a passenger can transfer from a line to another. The problem cannot be tackled with Markov chain modelling anymore, which generate unrealistic random trajectories. Instead, we will assume that passengers follow shortest paths. Starting from an origin-destination matrix $\mathbf{N} = (n_{st})$, where s denotes the stop at which a passenger *enters into the network* (and not simply enters a particular line), and t denotes the stop where the passenger *leaves definitively the network*, this shortest paths assumption allows us to compute the flow matrix on edges $\mathbf{X} = (x_{ij})$. The latter decomposes into the within-line flow and the transfer flow, i.e., $\mathbf{X} = \mathbf{X}_{\text{W}} + \mathbf{X}_{\text{B}}$. Moreover, in the multi-line problem, we also have to distinguish between:

- passengers who enter and leave bus lines at each stop, represented by vectors $\mathbf{a}_{\text{in}}$ and $\mathbf{a}_{\text{out}}$, which are *measured*,

- and passengers who enter and leave the network at each stop i, represented by the *unknown* quantities $n_{\bullet i}$ and $n_{i\bullet}$.

By construction,

$$a_i^{\text{in}} = n_{i\bullet} + x_{i\bullet}^{\text{B}}, \qquad a_i^{\text{out}} = n_{\bullet i} + x_{\bullet i}^{\text{B}}, \tag{19}$$

Using these two constraints, along with the shortest paths assumption and iterative proportional fitting, we propose the following iterative algorithm in order to find $\mathbf{N}$ from measured $\mathbf{a}_{\text{in}}$ and $\mathbf{a}_{\text{out}}$.

Initialisation: $\mathbf{S}^{(0)}$ is filled with 1 excepted for aberrant origin-destination pairs (such as t beeing a previous stop of the same line as s). The margins of $\mathbf{N}$ are fixed as $\mathbf{n}_{\text{in}}^{(0)} = \mathbf{a}_{\text{in}}$ and $\mathbf{n}_{\text{out}}^{(0)} = \mathbf{a}_{\text{out}}$.
Step 1, Iterative proportional fitting: We use IPF to compute $\mathbf{N}^{(r)}$ starting from $\mathbf{S}^{(r)}$, such that margin constraints, defined by $\mathbf{n}_{\text{in}}^{(r)}$ and $\mathbf{n}_{\text{out}}^{(r)}$ are satisfied.
Step 2, Shortest paths flow: Using shortest paths information, we compute $\mathbf{X}_{\text{B}}^{(r)}$ from $\mathbf{N}^{(r)}$.
Step 3, Affinity and margin update: $\mathbf{S}^{(r+1)}$, $\mathbf{n}_{\text{in}}^{(r+1)}$ and $\mathbf{n}_{\text{out}}^{(r+1)}$ are updated in order to respect constraints defined by (1).

Step 1, 2, and 3 are iterated until convergence, giving an admissible solution to the problem.

7 A small example

As an illustration, an estimated solution proposed by the algorithm on a restricted network made of four lines only is depicted on Fig. **??**. A total of $n_{\bullet\bullet} = 16,837,494$ passengers using this network is estimated by the algorithm. The red circle on the bottom left represents the start of the trip s and the size of the circles at stops t represents the estimated number of passengers terminating their trip at t. In this example, the majority of passengers exit the network on the same initial embarkment line. A small fraction of them takes another line.

Table 1 represents the estimated ten most frequented transfer edges. The code of the stop represents the number of the line, the direction and a condensed name of its stop cluster. The third column gives the number (in thousands) of passengers transferring through this edge.

From stop	To stop	Count
S7_A_SF_O	S9_A_SF_O	192k
S9_R_CH_E	S6_A_CH_E	187k
S7_A_SF_O	S8_R_SF_S	135k
S6_R_CH_O	S9_A_CH_O	135k
S8_A_GTE_N	S9_R_GTE_E	103k
S9_R_B-AIR_C	S8_A_B-AIR_N	99k
S9_A_SF_O	S7_A_SF_O	88k
S8_R_B-AIR_D	S6_A_B-AIR_C	87k
S6_R_SF_O	S8_A_SF_O	86k
S9_A_GTE_O	S8_R_GTE_S	84k

Table 1: List of the ten most frequented transfer edges

The current work performs computer-intensive simulations of flow over the entire network (1361 stops), permitting to extract usual network indices (centrality, betweenness...) characterizing both the stops *and* the lines. In parallel, the computational effects of various fine tuning calibration parameters used in the algorithm are investigated.

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)
13. Boyle, D., Mark C, D.: Passenger Counting Technologies and Procedures. National Academy Press, Washington, DC (1998). OCLC: 632725908
14. Bishop, Y.M., Fienberg, S.E., Holland, P.W.: Discrete Multivariate Analysis: Theory and Practice. Springer, New York (2007)

Figures

Figure 4: Sample figure title

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.

Figure 5: Sample figure title

Table 2: 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	..	.	.

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