RESEARCH

Estimation of flow trajectories in a multi-lines network

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

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

Abstract

Blabla

Keywords: sample; article; author

1 Notations and formalism

1.1 Lines, stops and junctions

Consider a transportation network made of bus 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 $l = \sum_{\ell=1}^{q} l_\ell$ bus stops constitute the nodes of the transportation network. Each stop $i = 1, \dots, l$ 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.

1.2 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.

1.3 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, **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} 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 1.1, 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.

1.4 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 **a** and **b**. 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). This issue of incompletely observed data can be tackled by the maximum entropy formalism (*** ref***).

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 1.4.2), i.e. as the distribution closest to g_{st} in the Kullback-Leibler divergence sense under the margin constraints

(2) then (section 1.4.3), 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 1.4.4), the prior will be updated to $\tilde{g}_{st}$ by contracting, 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 *** . This solution method is reminiscent, but quite distinct of the ME algorithm ...***

*1.4.1 ***Maximum entropy solution*

Let $\mathscr{F}$ consist of the empirical distributions $\mathbf{f}$ satisfying R linear constraints of the form $\sum_{st} f_{st}\, o^r_{st} = \bar{o}^r$ for $r = 1, \ldots, R$. Then the minimum Kullback-Leibler divergence

$$\min_{\mathbf{f} \in \mathscr{F}} \sum_{st} f_{st} \log \frac{f_{st}}{g_{st}}$$

is reached for the maximum entropy solution

$$f^0_{st} = \frac{1}{Z(\boldsymbol{\lambda})} g_{st} \exp(-\sum_{r=1}^{R} \lambda_r o^r_{st}) \qquad\qquad Z(\boldsymbol{\lambda}) = \sum_{st} g_{st} \exp(-\sum_{r=1}^{R} \lambda_r o^r_{st}) \tag{6}$$

where the Lagrange multipliers $\boldsymbol{\lambda}$ are determined so as to satisfy the R linear constraints (**see e.g. Bavaud and *** and references therein).

*1.4.2 Maximum entropy solution ***Statement of the problem (version 2, added by GG)*

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). Also suppose a prior guess of row and margin constraints for f_{st}, respectively α_s and β_t, which has to verify

$$\frac{\alpha_i}{\alpha_j} = \frac{a_i}{a_j} \qquad \frac{\beta_i}{\beta_j} = \frac{b_i}{b_j} \qquad \forall i,j \text{ which are not true junctions.} \tag{7}$$

Setting these gives us

$$n_{\bullet\bullet} = \frac{a_i}{\alpha_i} = \frac{b_i}{\beta_i} \qquad \forall i \text{ which are not true junctions.} \tag{8}$$

We are interested in finding

$$\begin{aligned} &\min_{\mathbf{f}\in\mathscr{F}} \sum_{st} f_{st} \log \frac{f_{st}}{g_{st}}, \\ &s.t. \sum_t f_{st} = \alpha_s, \\ &\qquad \sum_s f_{st} = \beta_t. \end{aligned} \tag{9}$$

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,

$$f_{st} = \phi_s \psi_t g_{st} \qquad \text{with } \phi_s := \exp(-1-\lambda_s),\ \psi_t := \exp(-\mu_t). \tag{10}$$

Using constraints in (9), we find

$$\phi_s = \frac{\alpha_s}{\sum_t \psi_t g_{st}}, \qquad \psi_t = \frac{\beta_t}{\sum_s \phi_s g_{st}}, \tag{11}$$

which gives an *iterative fitting algorithm*. Starting with any $\psi_t^{(0)} > 0$, we do

$$\phi_s^{(\iota)} = \frac{\alpha_s}{\sum_t \psi_t^{(\iota)} g_{st}}, \qquad \psi_t^{(\iota+1)} = \frac{\beta_t}{\sum_s \phi_s^{(\iota)} g_{st}}, \tag{12}$$

With equation (8), we can obtain the flow associated to f_{st}

$$n_{st} = n_{\bullet\bullet} f_{st}, \tag{13}$$

and with the edge-trip incidence matrix $\boldsymbol{\chi}$ (1), we obtain the intra-line edge flows $\mathbf{Y} = (y_{ij})$ and the transfer edge flows $\mathbf{Z} = (z_{ij})$.

1.4.3 Margin distributions update

Let us define the hyperparameter $0 \geq \theta > 1$ as the *minimum proportion of passengers (among a_i and b_i) entering/leaving the network at each stop*. Note that we could set a different hyperparmeter for each node, and for embarkations and disembarkations, but without addition information, we will restrain to this simpler case. This hyperparamter means that

$$z_{\bullet s} < (1-\theta)a_s \tag{14}$$

$$z_{t\bullet} < (1-\theta)b_t \tag{15}$$

and with constraints (4) and (5), the update of margins distribution must be

$$\widetilde{\alpha}_s = \frac{\min(\theta a_s, a_s - z_{\bullet s})}{\sum_{s'} \min(\theta a_s, a_{s'} - z_{\bullet s'})} \tag{16}$$

$$\widetilde{\beta}_t = \frac{\min(\theta b_t, b_t - z_{t\bullet})}{\sum_{t'} \min(\theta b_t, b_{t'} - z_{t'\bullet})} \tag{17}$$

1.4.4 Prior distribution update (added by GG)

Transfer edges (i,j) where $z_{i\bullet} > (1-\theta)b_i$ or $z_{\bullet j} > (1-\theta)a_j$ typically denote an overflowing tansfer flow regarding the g_{st} prior distribution. The latter must be adjusted in order to have **Z** comptible with constraints. For any edge (i,j), let us compute the *flow ratio* r_{ij}, with

$$r_{ij} = \max\left(1, \frac{z_{i\bullet}}{(1-\theta)b_i}, \frac{z_{\bullet j}}{(1-\theta)a_j}\right), \tag{18}$$

where $r_{ij} > 1$ typically denotes an overflow through edge (i,j). This flow ratio can be traced back to origin-destination couples s,t in order to compute the *orgin-destination flow ratio* $\bar{r}_{st}$

$$\bar{r}_{st} = \max_{ij} \chi_{ij}^{st} r_{ij}, \tag{19}$$

where again $\bar{r}_{st} > 1$ denotes an overflow between s and t. To ajdust the flow, we can devide the previous flow by this ratio

$$\widetilde{n}_{st} = \frac{n_{st}}{\bar{r}_{st}} \tag{20}$$

And the new prior distribution should follow

$$\widetilde{g}_{st} = \frac{\left(\frac{\widetilde{n}_{st}}{\phi_s \psi_t}\right)}{\sum_{s',t'} \left(\frac{\widetilde{n}_{s',t'}}{\phi_{s'} \psi_{t'}}\right)} \tag{21}$$

2 Single line

Let $i = 1, \ldots, l$ enumerate the bus stops in increasing order, i.e. $F(i) = i+1$. Define $g_{st} = \frac{1}{(l-1)(l-2)} \chi(s<t)$ (where $\chi(.)$ denotes the 0/1 indicator function) as the maximally uniform prior, reflecting only the unidirectional nature of trips. Constraints $n_{r\bullet} = a_r$ (resp. $n_{\bullet r} = b_r$), correspond to $o_{st}^r = \delta_{sr}$ (resp. $o_{st}^r = \delta_{tr}$) in (6), which finally yields, after reparametrization, the maximum entropy flow reads

$$n_{st}^0 = I(s<t)\, c_s\, d_t \qquad \text{where} \quad \sum_{s<t} c_s\, d_t = n_{\bullet\bullet} = a_\bullet = b_\bullet \tag{22}$$

In addition, the (dis-)embarking constraints yield

$$a_i = c_i \sum_{t>i} d_t = c_i D_i \qquad b_j = d_j \sum_{s<j} c_s = d_j C_j \tag{23}$$

where $D_i = \sum_{t>i} d_t$ and $C_j = \sum_{s<j} c_s$. This maximum entropy solution constitutes one possible consistent transportation flow among many others, such as the "first in, first out"

(FIFO) flow. Interestingly enough, (22) is reminiscent of the so-called gravity flows of quantitative Geography (**ref**) $n_{st} = h_{st}\, c_s\, d_t$, but with a purely asymmetric "distance deterrence function" $h_{st} = I(s < t)$. Also, (22) shows that the conditional probability to exit on t, given an entrance on s, is $w_{st} = n^0_{st}/n^0_{s\bullet} = I(s < t)\, d_t/D_s$.

*** ici la (les) figure de l'exemple "starting from Maladière, Riant-Cour, Dapples"... ? ***

*** ici l'équivalence avec l'approche chaîne de Markov de Guillaume ? ***

3 Multiple lines

*** $a_\bullet = b_\bullet$ (where "$\bullet$" denotes summation over the replaced index).

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 [1]. 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) [2].

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.

4 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.

5 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,

$$a_i^{\text{in}} = n_{i\bullet} + x_{i\bullet}^{\text{B}}, \qquad\qquad a_i^{\text{out}} = n_{\bullet i} + x_{\bullet i}^{\text{B}}, \tag{24}$$

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.

6 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.

7 Introduction

7.1 Context

7.2 Statement of the problem

7.3 Related Works

8 Formalism

8.1 Notations

8.2 Problem

8.3 Solution

9 Case Studies

9.1 Toy Examples

9.2 Real Data

10 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]Departement of Language and Information Sciences, University of Lausanne, Lausanne, Switzerland. [2]Institute of Geography and Sustainability, University of Lausanne, Lausanne, Switzerland.

References
1. Boyle, D., Mark C, D.: Passenger Counting Technologies and Procedures. National Academy Press, Washington, DC (1998). OCLC: 632725908
2. 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

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	..	.	.

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.