Modélisation des déplacements à bord de trains pour l'estimation de la charge à bord par zone

Rémi Coulaud, Christine Keribin, Gilles Stoltz

Modélisation des déplacements à bord de trains pour l'estimation de la charge à bord par zone

Rémi Coulaud [1] & Christine Keribin [2] & Gilles Stoltz [2]

[1] *Transilien, SNCF Voyageurs, 10 rue Camille Moke, 93220, Saint-Denis, France*
[2] *Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France*

Résumé. Transilien dispose sur certaines lignes de son réseau de trains équipés de capteurs infra-rouges au niveau des portes, permettant de mesurer le nombre de descentes et montées par porte. Les rames de ces trains sont par ailleurs communicantes, i.e., les voyageurs, une fois montés, peuvent se déplacer dans le train : ainsi, la connaissance même exacte et sans erreur des descentes et montées au cours du trajet ne permet pas d'en déduire la charge à bord du train par zone. Nous considérons des modèles de déplacements simples où la loi du point d'arrivée ne dépend que de la zone de montée, dans la version la plus simple et la moins réaliste, et de cette zone et de la gare, dans une version plus complexe. Dans le premier cas, dit d'échelle globale, une matrice de transition markovienne décrit les lois des mouvements, et elle peut être estimée par moindres carrés ou maximum de vraisemblance. Dans le second cas, dit d'échelle locale, la matrice de transition dépend de la gare de montée ; modulo une hypothèse de modélisation supplémentaire de descentes binomales en la charge à bord, seule subsiste une estimation par maximisation de vraisemblance, mise en œuvre par un algorithme EM.

Mots-clés. Matrice de transition, maximum de vraisemblance, algorithme EM, trains de banlieue

Abstract. Transilien equipped trains of some lines of its Greater Paris network with infra-red sensors measuring the numbers of boarding and alighting passengers per door. These trains are also communicating, i.e., the passengers, after boarding, may move inside the train. Therefore, even a perfect (noiseless) knowledge of boarding and alighting numbers would not be sufficient to fully determine the passenger load by zone of the train. We introduce simple models of passengers' movements, where, in the simplest and less realistic formulation, the law of the arrival point only depends on the boarding zone, and in a more complex version, it also depends on the train station where boarding took place. For the first version, referred to as a global approach, a stochastic transition matrix describes passengers' movements ; it may be estimated by a least-mean-square or a maximum likelihood approach. In the second version, referred to as a local approach, the transition matrix depends on the train station where boarding took place ; an additional convenient modeling assumption (given by a binomial modeling of alighting numbers, in terms of current load) is issued, and maximum likelihood may be carried over with care, thanks to some EM algorithm.

Keywords. Transition matrix, maximum likelihood, EM algorithm, commuter trains

1 Objectif d'estimation de la charge à bord

Les usagers des transports en commun vivent des situations de congestion en particulier aux heures de pointe. Cette congestion est moins acceptée depuis la pandémie de COVID-19. Ainsi, SNCF-Transilien a déployé un service en gare (voir Figure 1) permettant de connaître la charge à bord des trains par zone. Ce service a pu être développé grâce aux nouvelles rames (NAT, Regio2N et bientôt RER NG) qui permettent de mesurer en temps réel le nombre de montées et de descentes depuis chacune des zones. Dans notre cas, une zone est associée à au moins une porte d'une rame. Le renouvellement des rames par la région Île-de-France est massif car depuis 2008 déjà 50% du parc de SNCF-Transilien est renouvelé. La mesure du nombre de montées et de descentes par zone n'est toutefois pas suffisante pour estimer correctement la charge à bord par zone car les rames sont communicantes (voir Figure 1) : la conception même des rames incite les voyageurs à se déplacer à bord. Et on constate dans nos données (voir Figure 2) qu'en effet, un volume significatif de déplacements a lieu, ce qui empêche évidemment de calculer la charge à bord par différences par zone des montées avec les descentes. Il convient de modéliser et d'estimer les déplacements moyens pour corriger ces différences par zone et en déduire une estimation de la charge à bord par zone.

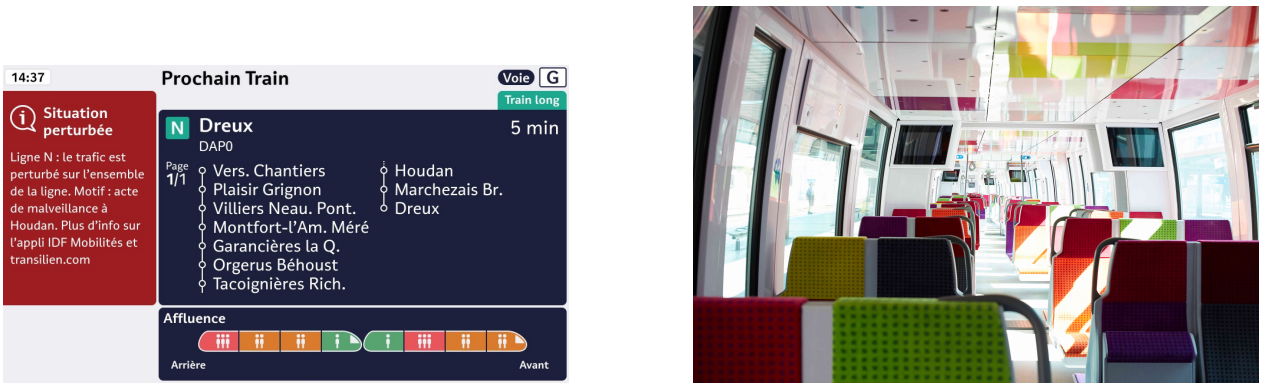

FIGURE 1 – À gauche, affichage déployé dans certaines gares de la charge à bord par zone ; à droite, un exemple de l'intérieur d'une rame communicante.

Revue de la littérature. Pour éviter ce problème, de nombreux opérateurs de transport privilégient une mesure directe de la charge à bord à l'aide de caméras vidéos ou de capteurs de pression. Par exemple, la société Thales (2021) utilise à Londres des caméras de vidéo-surveillance natives pour mesurer le volume de voyageurs. Certains opérateurs à Copenhague (Nielsen et al., 2014), Stockholm (Zhang et al., 2017), Londres (Schmitt, 2017; Rogers, 2019) ou Singapour (Ngauw, 2018) privilégient une mesure semi-directe par la masse à l'essieu en utilisant des capteurs de pression au niveau des essieux.

A défaut, des modélisations des déplacements sont utilisées. Certaines sont théoriques et reposent sur des modèles multi-agents, comme Schöttl et al. (2019) : or, nous nous intéressons ici aux modélisations reposant sur des données. À notre connaissance, de telles modélisations n'ont pas encore été réalisées pour des questions de transports ferroviaires. En revanche, nous pouvons citer des modèles markoviens de déplacements à l'intérieur de bâtiments :

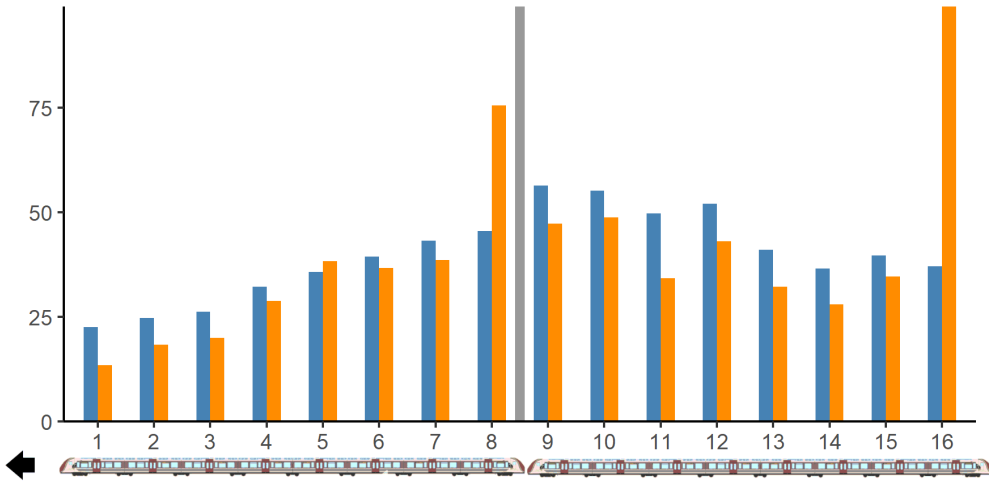

FIGURE 2 – Trajets Paris – banlieue sur la ligne H : diagramme en barres des montées moyennes (en bleu) et des descentes moyennes (en orange) par zones, en nombres de passagers, où les moyennes s'entendent sur l'ensemble des trajets et des gares considérés. La barre grise verticale sépare la rame avant, comportant 8 zones, de la rame arrière, avec elle aussi 8 zones. Les totaux de montées et de descentes moyennes par zones entre les voitures 1 à 8, et 9 à 16, sont égaux : les sommes des hauteurs des barres orange et bleues sont identiques. Dans notre approche, l'estimation des mouvements sera conduite séparément pour chaque rame.

Wang et al. (2011) et Shelat et al. (2020) utilisent à cet effet des données issues de capteurs de CO_2, de température, de son, etc. qui donnent accès directement au volume de personnes présentes dans une pièce, et en déduisent les déplacements. Dans notre cas, nous ne mesurons pas directement le nombre de personnes effectivement présentes dans chaque zone, mais uniquement les flux de voyageurs qui rentrent et sortent par zones.

Source pour ce résumé. Les travaux présentés dans ce résumé sont extraits en bonne partie du chapitre 5 de la thèse de doctorat de Coulaud (2022), auquel nous renvoyons pour plus de détails.

2 Données disponibles et notations

L'ensemble des plus de 3 500 trajets provenant de la ligne H entre janvier et octobre 2021 dans le sens Paris – banlieue est noté $\mathcal{T}$, et un trajet donné sera indexé par $t \in \mathcal{T}$.

Les zones d'une rame seront notées i ou j, et I désignera le nombre total de zones ; $I = 8$ pour chaque rame sur la Figure 2. Nous séparons totalement le problème entre les deux rames, du fait de l'absence de communication possible.

Les montées voyageurs (*boarding* en anglais) à la gare s et dans la zone i seront notées, de manière générique, $b_{s,i}$. Lorsqu'il s'agira des montées du trajet $t \in \mathcal{T}$, on ajoutera un exposant : $b^t_{s,i}$.

Nous employons la notation indicielle point $\bullet$ pour désigner la somme partielle associée à l'indice correspondant. Ainsi, le total des montées dans la zone i sur l'ensemble des gares d'un trajet sera désigné par $b_{\bullet,i}$, et de même pour la somme $b_{s,\bullet}$ des montées par gare sur l'ensemble des zones. Les vecteurs de sommes partielles seront des vecteurs lignes : par exemple, le vecteur ligne $\boldsymbol{b}_{\bullet,1:I} = (b_{\bullet,1}, \ldots, b_{\bullet,I})$ reporte toutes les montées par zones sur l'ensemble d'un trajet.

Des notations analogues s'appliquent pour les descentes (*alighting* en anglais), avec la lettre a.

3 Modélisation des déplacements à une échelle globale

Par échelle globale, nous entendons une modélisation des déplacements commune à toutes les gares du trajet : pour une rame donnée (avant ou arrière), nous supposons que quelle que soit la gare s, en moyenne, une proportion $p_{i,j}$ des voyageurs montés en zone i se déplacent jusqu'en zone j. Cela donne lieu à une matrice de transition $\boldsymbol{P}$ de dimension $I \times I$, où l'on rappelle que $I = 8$.

Estimation par moindres carrés. Par loi des grands nombres sur un trajet donné t, puisqu'il y a de nombreux passagers, et vu l'hypothèse d'homogénéité des déplacements, les descentes totales par zones sont données, en moyenne, par le vecteur ligne $\boldsymbol{b}^t_{\bullet,1:I}\boldsymbol{P}$. Il suffit alors de les confronter au vecteur ligne des descentes observées par zones $\boldsymbol{a}^t_{\bullet,1:I}$, et ce, pour chaque trajet $t \in \mathcal{T}$. Nous choisissons de minimiser des écarts quadratiques et estimons donc $\boldsymbol{P}$ selon

$$\widehat{\boldsymbol{P}}_{\mathrm{MC}} \in \operatorname*{arg\,min}_{\boldsymbol{Q} \text{ stoch.}} \sum_{t \in \mathcal{T}} \left\| \boldsymbol{a}^t_{\bullet,1:I} - \boldsymbol{b}^t_{\bullet,1:I}\, \boldsymbol{Q} \right\|_2^2 ,$$

où le minimum porte sur l'ensemble des matrices stochastiques. Il s'agit d'un problème d'optimisation quadratique sous contraintes d'égalité et d'inégalité linéaires, que nous résolvons numériquement avec la fonction `lsqlincon` du package `pracma` de `R`. (L'ajout de contraintes d'inégalité empêche l'écriture d'une solution analytique du problème.)

Estimation par maximum de vraisemblance. Une alternative consiste à estimer $\boldsymbol{P}$ par maximum de vraisemblance. On suppose que chaque voyageur montant dans un train s'y déplace ensuite indépendamment de tous les autres voyageurs (ce qui est évidemment critiquable, notamment en cas d'affluence) et, lorsqu'il est monté par la zone i, selon la loi de probabilité donnée par la ligne i de $\boldsymbol{P}$. Après ces déplacements, les charges à bord par zones $C_{\bullet,1:I}$ sont donc données comme la somme de I lois multinomiales indépendantes, indexées par i, et de paramètres respectifs $b_{\bullet,i}$ et $(p_{i,1}, \ldots, p_{i,I})$. La densité d'une telle somme n'est pas calculable numériquement (Hong, 2013; Lin et al., 2022). On effectue alors l'approximation de la loi de la somme par une multinomiale particulière :

$$\sum_{i=1}^{I} \text{de lois } \mathcal{M}\Big(b_{\bullet,i},\, (p_{i,j})_{j=1,\ldots,I}\Big) \text{ indépendantes} \;\approx\; \mathcal{M}\left(b_{\bullet,\bullet},\, \left(\sum_i \frac{b_{\bullet,i}}{b_{\bullet,\bullet}} p_{i,j}\right)_{j=1,\ldots,I}\right).$$

Cette approximation repose sur les justifications suivantes : (i) les espérances sont égales ; (ii) dans les deux cas, les covariances sont négatives en dehors de la diagonales ; (iii) cette approximation permet de mener facilement les calculs de maximisation de vraisemblance ; (iv) les résultats numériques d'estimation qui seront obtenus par cette approximation sont bons.

Comme on n'étudie ici qu'une échelle globale, les charges à bord sont égales aux descentes. On a ainsi, pour un trajet donné, la modélisation probabiliste suivante :

$$a_{\bullet,1:I} \quad \text{est la réalisation de} \quad A_{\bullet,1:I} \sim \mathcal{M}\left(b_{\bullet,\bullet},\, \left(\sum_i \frac{b_{\bullet,i}}{b_{\bullet,\bullet}} p_{i,j}\right)_{j=1,\dots,I}\right),$$

où l'on rappelle que les montées b et les descentes a sont observées. En supposant les trajets $t \in \mathcal{T}$ indépendants, on aboutit à l'estimateur suivant :

$$\widehat{\boldsymbol{P}}_{\mathrm{MV}} \in \operatorname*{arg\,min}_{\boldsymbol{Q}\ \text{stoch.}} \sum_{t\in\mathcal{T}} \sum_{j=1}^{I} a^t_{\bullet,j} \ln\left(\sum_{i=1}^{I} \frac{b^t_{\bullet,i}}{b^t_{\bullet,\bullet}}\, p_{i,j}\right) ;$$

nous résolvons ce problème d'optimisation convexe sous contraintes d'inégalité et d'égalité grâce au package `Rsolnp` de `R`.

Comparaison des performances numériques. Nous présenterons les résultats obtenus par les deux méthodes d'estimation, qui se correspondent en très grande partie. Nous indiquerons également l'amélioration obtenue en modélisant les descentes grâce à ce modèle, par comparaison à la situation où les descentes par zones seraient obtenues uniquement en répercutant les montées originelles, sans déplacements—en l'occurrence, une réduction d'erreur de modélisation d'environ 50%.

4 Modélisation des déplacements à une échelle locale

La modélisation homogène précédente néglige l'impact, tout à fait critique, de la position des entrées/sorties du quai sur la répartition des montées et des descentes, et donc des déplacements à bord. Nous allons donc désormais supposer que la matrice de transition $\boldsymbol{P}_s = [p_{s,i,j}]_{i,j}$ dépend de la gare s considérée. À cause de cette inhomogénéité, il sera désormais nécessaire de suivre la charge à bord au fil du trajet, contrairement au cas homogène. Cette charge à bord sera révélée en partie par les descentes, selon l'hypothèse d'émission inhomogène suivante : chaque passager de la zone i descend à la gare s avec probabilité $\alpha_{s,i}$, indépendamment des autres, i.e., les descentes $a_{s,i}$ à la gare s et pour la zone i sont la réalisation d'une variable aléatoire binomiale

$$A_{s,i} \sim \mathrm{Bin}\big(C_{s-1,i}, \alpha_{s,i}\big)\,. \tag{1}$$

Nous notons par ailleurs $Z_{s,i}$ le nombre de passagers montés à la gare s et se trouvant en zone i après déplacement. Avec la même approximation que dans le cas homogène, nous

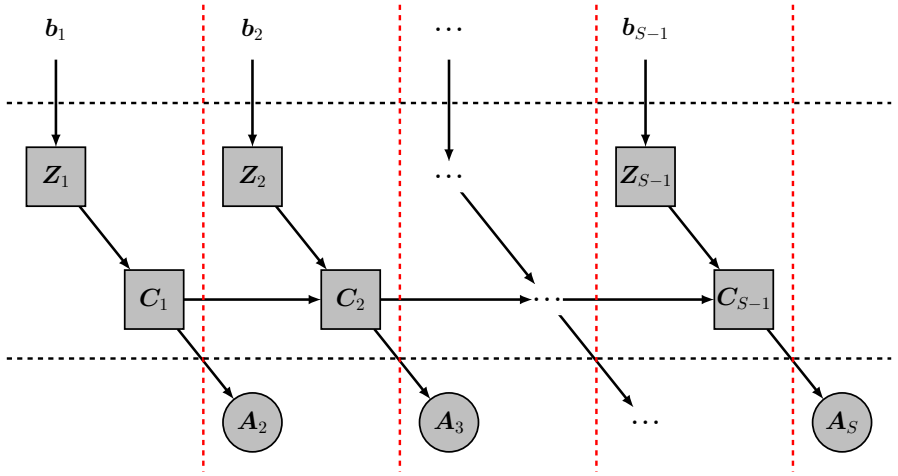

FIGURE 3 – Relation séquentielle entre les variables au fil du trajet : les variables latentes sont indiquées dans des carrés, tandis que les variables observées sont dans des disques ; les traits verticaux rouges séparent les gares.

allons supposer que

$$\boldsymbol{Z}_{s,1:I} \sim \mathcal{M}\left(b_{s,\bullet},\ \left(\sum_i \frac{b_{s,i}}{b_{s,\bullet}} p_{s,i,j}\right)_{j=1,\dots,I}\right). \tag{2}$$

Par conservation de la masse, le vecteur des charges à bord $\boldsymbol{C}_{s,1:I}$ en sortie de gare s est régi par l'équation suivante

$$\boldsymbol{C}_{s,1:I} = \boldsymbol{C}_{s-1,1:I} + \boldsymbol{Z}_{s,1:I} - \boldsymbol{A}_{s,1:I}\,. \tag{3}$$

Cela est illustré à la Figure 3. Les réalisations $\boldsymbol{a}_{s,1:I}$ des descentes $\boldsymbol{A}_{s,1:I}$ et les montées $\boldsymbol{b}_{s,1:I}$ avant déplacements sont observées ; les variables $\boldsymbol{Z}_{s,1:I}$ et $\boldsymbol{C}_{s,1:I}$ sont en revanche latentes. Le paramètre à estimer est $\boldsymbol{\theta} = (\boldsymbol{\alpha}_{2:S-1,1:I}, \boldsymbol{P}_{1:S-1})$, où la dernière gare est notée S. Dans tout ce qui précède, les indices du type $n_1{:}n_2$ désignent des suites d'éléments indexés entre n_1 et n_2.

Vraisemblance. Avec des hypothèses naturelles d'indépendance et de lois conditionnelles, et en tenant compte des effets de bord (aucune descente à la première gare et aucune montée à la dernière gare), la modélisation précédente aboutit à la vraisemblance complète suivante, pour un trajet générique (à un terme $\mathbb{1}_{\{\sum_{s=1}^{S-1}(\boldsymbol{z}_{s,1:I}-\boldsymbol{a}_{s,1:I})=0\}}$ près, qui n'a pas d'importance pour l'estimation du paramètre) :

$$\mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{a}, \boldsymbol{b}) = \prod_{s=2}^{S} \underbrace{\left(\prod_{i=1}^{I} \binom{c_{s-1,i}}{a_{s,i}} (\alpha_{s,i})^{a_{s,i}} (1-\alpha_{s,i})^{(c_{s-1,i}-a_{s,i})}\right)}_{\text{descentes}} \underbrace{\left(b_{s-1,\bullet}! \prod_{i=1}^{I} \frac{(\pi_{s-1,i})^{z_{s-1,i}}}{z_{s-1,i}!}\right)}_{\text{montées}},$$

$$\text{où} \qquad \pi_{s-1,i} = \sum_i \frac{b_{s-1,i}}{b_{s-1,\bullet}} p_{s-1,i,j}\,.$$

Les valeurs $a_{s,i}$ et $b_{s,i}$ sont observées mais celles des $z_{s,i}$ sont latentes, tandis que les valeurs des charges $c_{s,i}$ se déduisent du reste suivant l'équation de récurrence (3). La vraisemblance

complète d'un trajet $t \in \mathcal{T}$ s'écrit donc $\mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{z}^t, \boldsymbol{a}^t, \boldsymbol{b}^t)$. La vraisemblance liée aux observations du trajet t se déduit en sommant la vraisemblance complète sur toutes les valeurs possibles des variables latentes $\boldsymbol{z}^t = \boldsymbol{z}^t_{1:S-1,1:I}$:

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{a}^t, \boldsymbol{b}^t) = \sum_{\boldsymbol{z}^t \in \mathcal{Z}^t} \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{z}^t, \boldsymbol{a}^t, \boldsymbol{b}^t)$$

où $\mathcal{Z}^t$ désigne l'espace des réalisations des variables latentes $\boldsymbol{z}^t$ du trajet t. Cet espace dépend du trajet, car il est déterminé par les observations $\boldsymbol{a}^t$ et $\boldsymbol{b}^t$, qui sont en général différentes d'un trajet à l'autre. Par hypothèse d'indépendance des trajets, la vraisemblance observée sur l'ensemble des trajets est le produit des vraisemblances de chaque trajet :

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{a}^{\mathcal{T}}, \boldsymbol{b}^{\mathcal{T}}) = \prod_{t \in \mathcal{T}} \sum_{\boldsymbol{z}^t \in \mathcal{Z}^t} \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{z}^t, \boldsymbol{a}^t, \boldsymbol{b}^t)\,, \tag{4}$$

où $\boldsymbol{a}^{\mathcal{T}}$ (respectivement, $\boldsymbol{b}^{\mathcal{T}}$) désigne l'ensemble des descentes (respectivement, montées) sur tous les trajets, à toutes les gares et toutes les zones.

Estimation du paramètre. Il s'agit maintenant de maximiser en $\boldsymbol{\theta}$ la vraisemblance donnée par l'équation (4). Nous mettons en œuvre un algorithme EM (Dempster et al., 1977), classiquement utilisé en présence de variables latentes. Cet algorithme itère l'enchaînement de deux étapes, l'étape E et l'étape M. L'étape E calcule l'espérance du logarithme de la vraisemblance complète conditionnellement aux observations et pour une valeur courante du paramètre $\boldsymbol{\theta}^{(c)}$, soit

$$q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(c)}) = \sum_{t \in \mathcal{T}} \mathbb{E}\big(\log \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{Z}^t, \boldsymbol{A}, \boldsymbol{B}) | \boldsymbol{A} = \boldsymbol{a}^t, \boldsymbol{B} = \boldsymbol{b}^t; \boldsymbol{\theta}^{(c)} \big).$$

L'étape M maximise l'expression précédente en $\boldsymbol{\theta}$. La valeur maximisante devient la nouvelle valeur du paramètre en cours $\boldsymbol{\theta}^{(c+1)}$. Cette procédure garantit que $\mathcal{L}(\boldsymbol{\theta}^{(c+1)}; \boldsymbol{a}^{1:T}, \boldsymbol{b}^{1:T}) \geq \mathcal{L}(\boldsymbol{\theta}^{(c)}; \boldsymbol{a}^{1:T}, \boldsymbol{b}^{1:T})$ à chaque itération et la convergence des itérations vers un maximum local de la vraisemblance.

L'expression de $q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(c)})$ se décompose en trois parties et fait intervenir les termes d'espérance conditionnelle de la charge à bord, $\gamma^t_{s,i}(\boldsymbol{\theta}^{(c)}) = \mathbb{E}(C^t_{s,i} | \boldsymbol{A}^t, \boldsymbol{B}^t; \boldsymbol{\theta}^{(c)})$, eux mêmes latents :

$$q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(c)}) = \underbrace{\sum_{t \in \mathcal{T}} \sum_{s=2}^{S} \sum_{i=1}^{I} \Big(a^t_{s,i} \log(\alpha_{s,i}) + (\gamma^t_{s,i}(\boldsymbol{\theta}^{(c)}) - a^t_{s,i}) \log(1 - \alpha_{s,i}) \Big)}_{q_\alpha(\boldsymbol{\alpha}, \boldsymbol{\theta}^{(c)})}$$
$$+ \underbrace{\sum_{t \in \mathcal{T}} \sum_{s=2}^{S} \sum_{i=1}^{I} \gamma^t_{s,i}(\boldsymbol{\theta}^{(c)}) \log(\pi_{s-1,i})}_{q_p(\boldsymbol{P}; \boldsymbol{\theta}^{(c)})} + \sum_{t \in \mathcal{T}} r(\boldsymbol{a}^t, \boldsymbol{b}^t)\,.$$

Le premier terme q_α ne dépend que de α et est relatif aux descentes, le second terme q_p ne dépend que de $\boldsymbol{P}$ et est relatif aux déplacements après les montées ; le dernier terme est

indépendant du paramètre $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{P})$, il est donc inutile pour l'étape de maximisation. La formule récursive (3) permet de calculer facilement les termes $\gamma_{s,i}^t$ à partir de l'espérance conditionnelle des variables latentes $\boldsymbol{Z}$, à savoir les $\mathbb{E}(Z_{s,i}^t | \boldsymbol{A} = \boldsymbol{a}^t, \boldsymbol{B} = \boldsymbol{b}^t; \boldsymbol{\theta}^{(c)})$, pour les gares $s = 1, \ldots, S-1$, les portes $i = 1, \ldots, I$, et les trajets $t = 1, \ldots, T$. Ces derniers termes sont en revanche délicats à calculer à cause de la taille de l'espace latent et nous discuterons différentes solutions pour leur obtention.

L'étape M quant à elle ne pose pas de problème : la maximisation de $q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(c)})$ est faite de façon indépendante sur $\boldsymbol{\alpha}$ et $\boldsymbol{P}$. La maximisation en α intervient sur le terme binomial $q_\alpha(\boldsymbol{\alpha}; \boldsymbol{\theta}^{(c)})$ et donne

$$\arg\max_{\alpha} q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(c)}) = \left(\frac{\sum_{t=1}^{T} \gamma_{s,i}^t(\boldsymbol{\theta}^{(c)})}{\sum_{t=1}^{T} a_{s,i}^t} \right)_{(s,i)}$$

tandis que celle en P procède, comme dans le cas global, d'une optimisation non linéaire sous contrainte de la partie multinomiale q_p, effectuée en utilisant le package `Rsolnp` de `R`.

Références

Coulaud, R. (2022). *Modélisation et prévision des variables d'exploitation ferroviaire et de flux de voyageurs en zone dense*. PhD thesis, Université Paris-Saclay.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–22.

Hong, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis*, 59:41–51.

Lin, Z., Wang, Y., and Hong, Y. (2022). The Poisson multinomial distribution and its applications in voting theory, ecological inference, and machine learning. *arXiv preprint arXiv:2201.04237*.

Ngauw, B. (2018).

Nielsen, B. F., Frølich, L., Nielsen, O. A., and Filges, D. (2014). Estimating passenger numbers in trains using existing weighing capabilities. *Transportmetrica A: Transport Science*, 10(6):502–517.

Rogers, S. (2019).

Schmitt, A. (2017). `https://usa.streetsblog.org/2017/08/03/these-london-trains-have-real-time-displays-to-reduce-crowding`.

Schöttl, J., Seitz, M. J., and Köster, G. (2019). Investigating the randomness of passengers' seating behavior in suburban trains. *Entropy*, 21(6):600.

Shelat, S., Daamen, W., Kaag, B., Duives, D., and Hoogendoorn, S. (2020). A Markov-chain activity-based model for pedestrians in office buildings. *Collective Dynamics*, 5:423–430.

Thales (2021). `https://www.thalesgroup.com/en/group/journalist/press-release/reflecting-passengers-top-public-transport-experience-priorities`.

Wang, C., Yan, D., and Jiang, Y. (2011). A novel approach for building occupancy simulation. *Building simulation*, 4(2):149–167.

Zhang, Y., Jenelius, E., and Kottenhoff, K. (2017). Impact of real-time crowding information: a Stockholm metro pilot study. *Public Transport*, 9(3):483–499.