Introduction au domaine de recherche

Estimateurs statistiques du nombre de voyageurs à bord d'un train de banlieue

Rodrigue Lazarus
École Normale Supérieure Paris

SNCF - Transilien
Datalab' Mass Transit Academy

Table des matières

Chapitre 1

Introduction

J'ai choisi de présenter les travaux effectués lors de mon stade de Master 2 chez SNCF Transilien, au sein d'une équipe de recherche sur des sujets de traitement de données appelée le Datalab'. Cette équipe faisait partie d'un organisme plus grand, nommé la Mass Transit Academy, dont le rôle était d'étudier des problématiques propres au transport de voyageurs en zones denses - problématiques relevant autant de sciences statistiques que sociales. Ce sujet représentait ainsi bien l'étroite interaction entre la recherche et la mise en pratique que l'on retrouve dans ces domaines de mathématiques appliquées, il s'agit également de mon expérience de recherche la plus avancée.

1.1 Contexte

Avec l'ouverture à la concurrence, SNCF Transilien a pour objectif d'améliorer davantage le confort et l'expérience des usagers. Si cela peut être obtenu par des actions matérielles comme la mise en service de trains plus confortables, une approche plus directe et moins coûteuse est d'utiliser les données que l'on possède sur la fréquentation des trains pour offrir aux voyageurs une information fiable sur l'affluence des trains qu'ils s'apprêtent à prendre. Un voyageur correctement averti sur la saturation du train voire du wagon dans lequel il souhaitait monter pourra alors adapter ses horaires de trajet ou son placement sur le quai pour éviter de monter dans une zone bondée. Ceci améliore d'une part son confort à titre personnel mais également réduit les risques d'incidents et de retard liés au temps d'attente pour entrer ou sortir du train en station. Les nouvelles rames déployées par Transilien sur certaines de ces lignes parisiennes amènent une nouvelle manière de compter le nombre de voyageurs par voiture, améliorant drastiquement les données à disposition. Cette méthode n'étant pas parfaite, des travaux mathématiques sont nécessaires pour compenser les erreurs rencontrées et améliorer la fiabilité de l'information. La détermination d'une modélisation permettant l'amélioration de la précision du comptage était ainsi un des sujets du stage, celui qui sera présenté ici. Pour information, deux autres sujets furent traités durant le stage : le premier traitait de la prédiction d'affluence à plusieurs échéances temporelles, le second était une étude visant à estimer l'impact de l'information fournie sur le comportement des voyageurs.

1.2 Travaux précédents sur le sujet

Ce sujet est très spécifique au cas d'usage de la SNCF car il présente deux caractéristiques majeures (qui seront définies en détail dans la prochaine partie) : les estimations du nombre de voyageurs par voiture se font par un comptage des montées et descentes aux portes et les rames sont d'un seul tenant et permettent aux voyageurs d'aller de voiture en voiture durant le trajet

(rames dites de type BOA). Le sujet consiste donc en l'étude de l'impact des déplacements inter-rames sur la fiabilité des modèles de comptage et l'implémentation d'une correction de cet effet. Les premiers travaux à ce sujet ont été faits par Rémi Couland dans [1], au sein de la même équipe. Ces travaux mettaient en valeur une erreur effective induite par ces déplacements et proposaient une méthode de redressement des données. Ces travaux n'avaient été effectués que sur une seule ligne du réseau et proposaient un redressement très global, au détriment de la précision. L'objectif du projet était alors de reprendre la méthode d'étude et de modélisation en l'appliquant sur 9 lignes du réseau avec davantage de données d'étude et évaluer divers niveaux de précision sur l'implémentation d'une solution afin de trouver une modélisation optimale. Enfin, la question d'un nouveau modèle mathématique permettant une approche encore plus précise a été partiellement étudiée mais n'a pas été concluse par manque de temps.

Chapitre 2

Formulation et modélisation du problème

Entrons maintenant davantage dans le détail du sujet. Pour commencer, nous allons formuler explicitement le problème puis le modéliser mathématiquement, avant de proposer une solution de redressement.

2.1 Processus de comptage des voyageurs par voiture

Les données utilisées dans toute l'étude proviennent d'une même source : des capteurs présents à bord des trains.

2.1.1 Compteurs infra-rouges aux portes

Ces capteurs sont disposés au-dessus de chaque porte des nouvelles rames présentes sur le réseau transilien. Il existait au moment de l'étude deux types de rames différentes : les NAT et les Regio2N. Parmi leurs différences, celle qu'il faut retenir est que pour les NAT une porte équivaut à une voiture - ce qui permet de directement attribuer les comptages de la porte à la voiture - alors que pour les Regio2N, une porte dessert plusieurs voitures donc il faut faire une opération de répartition des données de comptages dans les voitures. Ce qu'il faut retenir, c'est que les capteurs sont en mesure, grâce à l'infra-rouge, de compter combien de personnes traversent les portes du train à chaque arrêt, dans le sens des montées et des descentes. À partir des ces données, on peut estimer le nombre de personnes présentes dans une voiture (il s'agit de la somme de toutes les personnes montées dans la voiture depuis le départ du train moins la somme de toutes les descentes).

2.1.2 Principales sources d'erreur

Les capteurs ont une erreur physique de l'ordre de 6% qui est difficilement corrigible mais négligeable car avec les ordres de grandeurs usuels (quelques dizaines de montées/descentes par trajet pour une porte) cela correspond à moins d'une personne d'erreur. Une première source d'erreur, spécifique aux Regio2N, a été mentionnée et consiste en la répartition des comptages aux portes entre les voitures. Celle-ci est assumée en essayant simplement de proposer une répartition simple et cohérente : pour chaque porte, on répartit le comptage entre les voitures adjacentes avec des proportions estimées à partir des capacités théoriques des voitures. Ainsi, une voiture avec 300 places de capacité théorique recevra 2 fois plus de passagers que sa voisine à 150 places. La source d'erreur principale, à l'origine de l'étude, provient des déplacements entre les voitures. En effet, lorsqu'un passager monte en voiture i, il est détecté à la porte correspondante et attribué définitivement à la voiture. S'il décide de se déplacer en voiture j en passant par l'intérieur du train, il n'est détecté par aucun capteur. Ainsi, on peut avoir

potentiellement des dizaines de personnes que l'on croit présentes dans une voiture alors qu'elles sont dans une autre, ce qui fausse complètement les estimations. L'objectif de l'étude est alors naturellement de réussir à estimer le nombre de personnes *réellement* présentes dans chaque voiture.

2.1.3 Formulation du problème

Maintenant que nous avons présenté le problème conceptuellement, nous allons le formuler mathématiquement.

2.2 Modélisation mathématique : concepts généraux

Nous commencerons par définir les variables pertinentes pour modéliser le problème puis nous énoncerons plusieurs hypothèses clés sur ces variables avant de donner le choix de méthodes de résolution adaptées.

2.2.1 Variables pertinentes

Avant de définir les variables, nous donnons quelques conventions de notation :
— Une variable en minuscule représente un scalaire (ex : a_i) ;
— Une variable en majuscule représente ou bien une matrice, ou bien une variable aléatoire composant un vecteur aléatoire (ex : P ou Z_i). Quelques constantes seront notées en majuscules par souci de cohérence de notations, cela sera précisé le cas échéant ;
— Une variable en gras représente un vecteur, aléatoire ou non (ex : $\mathbf{b}$ ou $\mathbf{Z}$)

Pour chaque observation (génération d'un jeu de données par les capteurs), nous définissons un ensemble de variables. Nous ajoutons également des variables agrégées sur plusieurs observations. Nous pouvons séparer les variables en deux catégories : les variables déterminant l'observation (jour, numéro identifiant le train, station où l'arrêt est effectué, nombre de voitures du train ou de la rame) et celles liée aux mesures de comptage (nombre de montées, de descentes, etc...). Toutes ces variables sont données dans le tableau 2.1 :

Variable	Signification	Valeurs possibles
I	Nombre de voitures du train ou de la rame	$\{7, 8, 14, 16\}$
s	Station	$\mathbb{N}^*$
d	Jour	$\{1, ..., 365\}$
k	Numéro du train (et position de la rame)	$\mathbb{N}^*$
$\mathbf{a}_s^{k,d} = (a_{s,i}^{k,d})_{i=1,...,I}$	Comptage des descentes (alightings)	$\mathbb{N}^I$
$\mathbf{b}_s^{k,d} = (b_{s,i}^{k,d})_{i=1,...,I}$	Comptage des montées (boardings)	$\mathbb{N}^I$
$\mathbf{a}^{k,d} = (a_i^{k,d})_{i=1,...,I}$	Comptage des descentes sur tout le trajet	$\mathbb{N}^I$
$\mathbf{b}^{k,d} = (b_i^{k,d})_{i=1,...,I}$	Comptage des montées sur tout le trajet	$\mathbb{N}^I$
$a_s^{k,d}$	Comptage des descentes sur tout le train	$\mathbb{N}$
$b_s^{k,d}$	Comptage des montées sur tout le train	$\mathbb{N}$
$a^{k,d}$	Somme totale des descentes pour un trajet	$\mathbb{N}$
$b^{k,d}$	Somme totale des montées pour un trajet	$\mathbb{N}$
$\mathbf{l}_s^{k,d} = (l_{s,i}^{k,d})_{i=1,...,I}$	Nombre de personnes à bord de chaque voiture (load)	$\mathbb{N}^I$
$l_s^{k,d}$	Nombre de personnes à bord du train entier	$\mathbb{N}$

TABLE 2.1 – Variables du problème

Par souci de clarté, nous enlèverons les notation (k, d) dès lors que nous parlerons d'un trajet générique. Nous noterons en majuscule les valeurs maximales prises par les variables définissant l'observation (s, k et d) pour un contexte donné. Par les définitions, nous pouvons noter les relations suivantes entre les variables :

$$b = \sum_{i=1}^{I} \sum_{s=1}^{S} b_{s,i} = \sum_{s=1}^{S} b_s$$

$$a = \sum_{i=1}^{I} \sum_{s=1}^{S} a_{s,i} = \sum_{s=1}^{S} a_s$$

$$l_s = \sum_{i=1}^{I} l_{s,i}$$

$$l_{s,i} = \sum_{s'=1}^{s} b_{s',i} - as', i$$

La dernière relation traduit le fait que le nombre de personnes à bord d'une voiture correspond au nombre de personnes qui y sont montées depuis la première station moins le nombre de personnes qui en sont descendues.

2.2.2 Hypothèses effectuées

Nous faisons les hypothèses suivantes :
- Les mesures de capteurs sont fiables quant au nombre de personnes traversant les portes ;
- Aucune personne ne monte ou descend du train entre les stations (relations données précédemment) ;
- Une personne montée en voiture j peut s'installer dans n'importe quelle autre voiture i tant que les deux voitures composent la même rame. Ce comportement est aléatoire et modélisé par une probabilité de déplacement $p_{i,j}$. De ce fait, le nombre de personnes montées à la porte i n'est pas forcément le nombre de personnes installées en voiture i ;
- Une personne installée en voiture j descendra du train par la porte j correspondante, sans nouveau déplacement. De ce fait, la mesure $b_{s,j}$ donnant le nombre de personnes descendues par la porte j donne exactement le nombre de personnes *présentes* en voiture j qui descendent à l'arrêt s.

2.2.3 Choix de méthodes

Notons $\bar{\mathbf{b}}_{\mathbf{s}}$ le vecteur donnant le nombre de personnes qui *s'installent* dans chaque voiture à l'arrêt s. $\bar{\mathbf{b}}_s$ est la variable à estimer pour connaître le nombre réel de personnes présentes dans chaque voiture. Puisque cette estimation repose sur un comportement aléatoire, nous choisissons d'utiliser des méthodes statistiques pour résoudre le problème. Plus précisément, nous adopterons deux méthodes : la régression linéaire et la méthode de log-vraisemblance.

2.3 Régression linéaire

La méthode principale utilisée pour la première phase de l'étude est la méthode linéaire.

2.3.1 Formulation de la modélisation

Deux phénomènes aléatoires ont lieu à chaque arrêt : le premier consiste en la répartition des passagers sur le quai et le second en leurs déplacements à l'intérieur du train. Nous posons ainsi deux variables aléatoires qui donnent respectivement le nombre de montées par voiture et le nombre d'installations par voiture pour un arrêt : $\mathbf{T} = (T_i)_{i=1,...,I}$ et $\mathbf{Z} = (Z_i)_{i=1,...,I}$. Le modèle linéaire consiste alors à trouver une matrice P telle que :

$$\mathbf{Z} = P\mathbf{T} + \varepsilon$$

avec $\varepsilon \sim \mathcal{N}(0, \sigma Id)$.
On peut alors estimer les installations en prenant l'espérance : $\bar{\mathbf{b}} = P\mathbf{b}$.
Nous pouvons de plus donner des contraintes sur la matrice P. En effet, la somme des installations doit valoir la somme des montées (puisque toute personne qui monte dans le train s'installe quelque part et réciproquement) et le nombre d'installation par voiture est positif ou nul. On en déduit que les coefficients de P sont positifs ou nuls et vérifient l'équation suivante :

$$\begin{aligned}
&\forall \mathbf{b} \in \mathbb{N}^I : \\
&\qquad \sum_{i=1}^{I} \bar{b}_i = \sum_{i=1} b_i \\
&\Leftrightarrow \sum_{i=1}^{I} \sum_{j=1}^{I} p_i, jb_j = \sum_{i=1} b_i \\
&\Leftrightarrow \sum_{i=1}^{I} b_i \sum_{j=1}^{I} p_{j,i} = \sum_{i=1} b_i \\
&\Leftrightarrow \sum_{i=1}^{I} b_i (1 - \sum_{j=1}^{I} p_{j,i}) = 0
\end{aligned}$$

On en conclut que la somme de chaque ligne de P vaut 1, donc P est stochastique. On retrouve alors bien le fait que le coefficient $p_{i,j}$ donne la probabilité qu'une personne montée en voiture j s'installe en voiture i.

On peut alors estimer la matrice par optimisation. Pour ce faire, on fixe un jeu de données d'entraînement correspondant à des observations $(k, d, s) \in \{1, ..., K\} \times \{1, ..., D\} \times \{1, ..., S\}$ et on utilise l'hypothèse qui stipule que les personnes descendent de la voiture où elles se sont installées. En effet, cela signifie qu'en sommant les montées et descentes sur tout le trajet, on doit avoir une égalité entre la somme des installations et la somme des descentes (i.e : $\bar{\mathbf{b}} = \mathbf{a}$). C'est donc l'écart entre les deux quantités que l'on souhaite minimiser pour trouver la matrice optimale :

$$P^* = \underset{P \text{ stochastique}}{\arg\min} \sum_{(k,d)} \|P\mathbf{b}^{k,d} - \mathbf{a}^{k,d}\|^2$$

Ce problème est en fait facilement résolvable puisqu'il s'agit d'un problème quadratique en P, pouvant être mis sous la forme cannonique :

$$f(\mathbf{p}) = \frac{1}{2} p^T U p + \mathbf{v}^T p$$

où $\mathbf{p}$ est une version "1D" de P et U et $\mathbf{v}$ sont déterminés à partir des variables du modèle (nous n'entrerons pas dans les détails pour ce rapport par soucis de concision).

2.3.2 Choix de scénarios temporels

Une fois le modèle établi, il faut maintenant déterminer quel jeu de données choisir pour son entraînement. Nous indiquons d'emblée que pour cette étude, nous avons souhaité générer une seule matrice par trajet (au contraire d'un modèle dit local qui proposerait par exemple une matrice par station) - puisque l'estimation se fait en sommant les données sur toutes les stations. De plus, par souci d'homogénéité sur le nombre de voitures I, nous comptons un train ayant plusieurs rames comme plusieurs trains faisant le même trajet au même moment (distingués par la position de la rame dans le train réel). Tout se passe ainsi comme s'il n'y avait que des trains possédant une seule rame à l'étude. Avant de s'intéresser à la plage horaire choisie pour une estimation, nous regardons déjà le type de trajets étudiés.

Nous en retenons trois :

— **Scénario par ligne** : Scénario de base, pour chaque ligne du réseau nous prenons tous les trajets ayant eu lieu sur une durée fixée pour estimer une matrice qui s'appliquera à tous les trajets ayant lieu jusqu'à la prochaine estimation. Dans notre cas d'usage, cela fait 9 matrices à estimer ;

— **Scénario par ligne et sens de circulation** : Dans le contexte des trains de banlieue, le sens de circulation (Paris vers banlieue ou l'inverse) peut avoir une incidence sur la manière dont les personnes s'installent dans un train. L'exemple le plus fréquent est celui d'une gare terminus comme la gare du Nord avec la sortie au bout du quai : les passagers seront a priori plus susceptibles de se déplacer vers l'avant du train durant le trajet pour gagner du temps. À l'inverse, les trains au départ de Gare du Nord peuvent avoir beaucoup de passagers qui montent dans la première voiture accessible et traversent le train de l'intérieur pour s'installer. Dans ce scénario, on estime une matrice pour tous les trains correspondant à une ligne et un sens de circulation donnés, cela représente 18 matrices à estimer ;

— **Scénario par ligne, sens de circulation et position de la rame** : Pour un train possédant plusieurs rames, la position de la rame (avant ou arrière du train) peut influencer les déplacements voyageurs puisque les accès aux quais ne seront pas placés de la même manière par rapport à la rame. On estime ainsi une matrice pour toutes les rames d'une ligne donnée, voyageant dans un sens donné et placée à une position donnée. La majorité des lignes du réseau ont des trains à 1 ou 2 rames mais certaines peuvent aller jusqu'à 3, ce qui donne au total 38 matrices à estimer.

Pour ce qui est des plages temporaires, nous testerons 3 scénarios :

— **Scénario 4 mois fixes** : nous prenons quatre mois de données pour estimer la matrice et testons les résultats sur le mois suivant ;

— **Scénario 1 semaine fixe** : nous prenons une semaine de données pour estimer la matrice et testons les résultats sur le mois suivant ;

— **Scénario 1 semaine roulante** : nous testons toujours sur un mois entier mais en procédant de la manière suivante : entraînement de la matrice sur la semaine $n-1$ pour le test sur la semaine n. Cela correspond à une estimation hebdomadaire des matrices faite avec la semaine précédente.

2.3.3 Résultats numériques

Pour estimer l'erreur commise lors d'un test sur N données, on définit la métrique MAE :

$$MAE(K, D, S, N) = \frac{1}{N} \sum_{(k,d,s)} |\mathbf{a}_s^{k,d} - P\mathbf{b}_s^{k,d}|$$

Nous résumons dans le tableau 2.2 les résultats obtenus en faisant la moyenne des erreurs calculées pour chaque matrice estimée, nous donnons aussi l'erreur obtenue sans déplacement ($P = Id$) :

Scénario temporel	Scénario physique	MAE
4 mois fixes	Lignes	7
	Lignes, sens	7
	Lignes, sens, positions	7
1 semaine fixe	Lignes	7
	Lignes, sens	7
	Lignes, sens, positions	7
1 semaine fixe	Lignes	7
	Lignes, sens	7
	Lignes, sens, positions	7
1 semaine roulante	Lignes	7
	Lignes, sens	7
	Lignes, sens, positions	7
Sans déplacement		10

TABLE 2.2 – Résultats généraux

Ce premier tableau montre la pertinence du redressement puisque l'erreur est réduite, mais il est difficile de déterminer le scénario le plus avantageux en regardant toutes les lignes confondues. Nous allons donc observer les résultats par lignes en les groupant par similarité du matériel roulant dessus, cela est résumé dans les tables 2.3, 2.4 et 2.5 :

Ligne	Sans déplacement	Lignes	Lignes, sens	Lignes, sens, position
E	9	8	7	7
H	12	7	7	6
K	8	6	6	5
P	15	10	9	9
J	15	7	7	7
L	13	8	7	7
D	3	3	3	3
N	7	6	6	6
R	9	9	9	9

TABLE 2.3 – MAE par ligne pour le scénario "4 mois fixes"

Ligne	Sans déplacement	Lignes	Lignes, sens	Lignes, sens, position
E	9	8	8	7
H	12	7	7	7
K	8	6	6	6
P	15	10	10	10
J	15	7	7	7
L	13	8	7	7
D	3	3	3	3
N	7	5	5	5
R	9	8	7	9

TABLE 2.4 – MAE par ligne pour le scénario "1 semaine fixe"

Ligne	Sans déplacement	Lignes	Lignes, sens	Lignes, sens, position
E	9	8	7	7
H	12	7	7	7
K	8	6	6	5
P	15	10	10	9
J	15	7	7	7
L	13	8	7	7
D	3	3	3	3
N	7	5	5	4
R	9	8	7	9

TABLE 2.5 – MAE par ligne pour le scénario "1 semaine roulante"

En regardant la colonne "sans déplacement", on peut d'emblée remarquer que certaines lignes n'ont pas vraiment de phénomène de déplacements voyageurs et que l'estimation des matrices est moins pertinente, au contraire d'autres lignes pour lesquelles une différence est notable. On distingue les deux premiers groupes du troisième, cela n'est pas surprenant puisque qu'ils correspondent aux rames de type NAT là où le dernier groupe correspond aux Regio2N, qui ont une architecture très particulière et facilite moins les déplacements. La seule différence entre les deux premiers groupes réside dans le nombre de voitures par rames (respectivement 8 et 7).

Sur ces observations, il a été décidé de n'appliquer les matrices de déplacement qu'aux lignes composées de rames NAT et d'étudier une autre manière d'estimer les données pour les Regio2N. Les scénarios retenus sont "1 semaine roulante" et "Lignes et sens" car ils présentent le meilleur compromis entre coût de calcul et performance.

Nous montrons en figure 2.1 à quoi ressemble une matrice estimée, ici pour les lignes H et J :

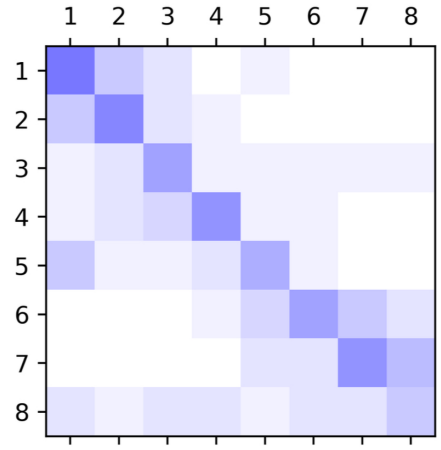

FIGURE 5 – Matrice de la ligne H

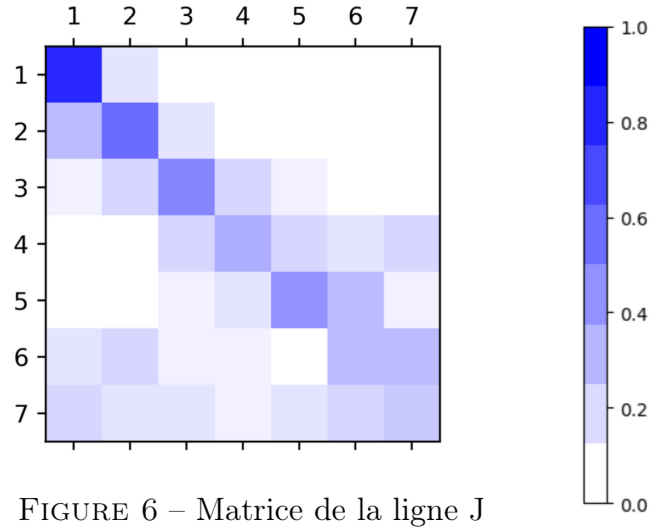

FIGURE 6 – Matrice de la ligne J

FIGURE 2.1 – Matrices obtenues pour les lignes H et J

On peut notamment constater que la plupart des passagers restent dans leur voiture de montée (diagonale forte) et que certaines voitures ont tendance à beaucoup distribuer vers les autres, notamment celles aux extrémités (phénomène Gare du Nord déjà mentionné).

2.4 Modélisation par vraisemblance

La seconde approche, proposée par Coulaud et al dans [1], est celle du calcul d'une vraisemblance et sa maximisation.

2.4.1 Expression d'une vraisemblance

Afin d'exprimer une vraisemblance, il faut faire de nouvelles hypothèses sur les variables aléatoires, en supposant notamment des comportements multinomiaux. Cette méthode étant surtout utile pour les travaux futurs mais n'ayant pas été utilisée en pratique pour cette étude, nous ne donnons pas les détail des hypothèses effectuées. Nous pouvons donner une formule close pour la log-vraisemblance :

$$l(P) = C + \sum_{n=1}^{N} \sum_{i=1}^{I} a_i^{(n)} ln(\sum_{j=1}^{I} \frac{b_j^{(n)}}{b^{(n)}} p_{j,i})$$

où C est une constante en P, n désigne une observation et $b^{(n)}$ le nombre total de montées pour cette observation. On peut ensuite trouver P en maximisant la vraisemblance par un algorithme d'optimisation convexe sous contrainte. Quelques essais ont montré que les résultats obtenus par cette méthode étaient similaires à ceux obtenus par la régression linéaire, d'où le choix de ne pas l'implémenter davantage.

Chapitre 3

Prochaines réflexions

Bien que ce modèle présente des résultats concluants, il présente quand même plusieurs problèmes. Il y a tout d'abord la question des Regio2N et du fait qu'une porte donne accès à plusieurs voitures ce qui complique l'association "montées à la porte" et "installations en voiture". Une solution proposée est de se concentrer uniquement sur ce qu'il se passe aux portes, en ignorant les voitures sans portes, estimer les matrices alors, effectuer le déplacement par ces matrices et seulement à la fin appliquer une répartition des voyageurs dans les voitures adjacentes aux portes.
Le second souci est la globalité du modèle : quoi qu'il arrive, on raisonne sur un trajet entier et pas à l'échelle de la station. Or, les géographies et niveaux d'affluence très divers des gares ont très probablement un impact sur le comportement des usagers. Se pose donc la question de la modélisation par stations.

3.1 Modélisation par station

Modéliser par station, cela signifie estimer une matrice P_s pour chaque station s d'une ligne et ensuite appliquer cette matrice aux observations obtenues dans cette station. Tous les scénarios énoncés prédecemment peuvent être appliqués pour affiner l'estimation.

3.1.1 Nouvelles difficutés

Plusieurs difficultés se présentent avec cette approche. Il y a d'une part les difficultés matérielles : estimer une matrice par station multiplie par 10 le nombre de matrices à estimer, ce qui peut demander une capacité de calcul consistante, surtout sur un scénario déjà assez précis. De plus, pour estimer la charge à bord il faut accéder aux données sur toutes les stations antérieures à celle étudiée, donc cela implique de faire appel à chaque matrice estimée jusque là et cela rajoute de la complexité au modèle. D'autre part, il y a une difficulté théorique : pour le modèle global, on regardait l'ensemble des descentes sur tout le trajet en décrétant qu'il fallait faire correspondre les montées à ces valeurs. Or, à l'échelle d'une station, il n'y a aucune raison que le nombre de descentes d'une voiture corresponde au nombre de personnes s'y installant - puisque les personnes qui s'installent dans la voiture à la station ne vont évidemment pas en descendre immédiatemment. De ce fait, il faut trouver une nouvelle variable objectif pour pouvoir faire des estimations.

3.1.2 Approche considérée

L'approche considérée alors, proposée par Coulaud et al dans [1], consiste en l'ajout de chaînes de Markov cachées et de variables latentes. Dans ce cadre, seule l'estimation par vrai-

semblance peut fonctionner, à condition de réussir à trouver une forme close de cette fonction. Quelques réflexions avaient été menées à ce sujet, sans avoir pu avancer davantage dessus.

References

[1] Rémi Coulaud. “Modélisation et prévision des variables d’exploitation ferroviaire et de flux de voyageurs en zone dense - Ecole doctorale de mathématiques Hadamard (2022)