Modélisation et prévision des variables d'exploitation ferroviaire et de flux de voyageurs en zone dense

Rémi Coulaud

université
PARIS-SACLAY

Modélisation et prévision des variables d'exploitation ferroviaire et de flux de voyageurs en zone dense

Modeling and forecasting of railway operation variables and passenger flows for dense traffic areas

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 574, Mathématique Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Mathématiques, Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche du Laboratoire de mathématiques d'Orsay **(Université Paris-Saclay, CNRS)** sous la direction de **Gilles STOLTZ**, Directeur de recherche, la co-direction de **Christine KERIBIN**, Maîtresse de conférences, le co-encadrement de **Pierre MESSULAM**, Docteur Ingénieur.

Thèse soutenue à Paris-Saclay, le 30 novembre 2022, par

Rémi COULAUD

Composition du jury

Membres du jury avec voix délibérative

Élisabeth Gassiat Professeure, Université Paris-Saclay	Présidente
Francesco Corman Professeur, ETH Zürich	Rapporteur & Examinateur
Yohann De Castro Professeur, Institut Camille Jordan	Rapporteur & Examinateur
Étienne Côme Chargé de recherche, Université Gustave Eiffel	Examinateur

Thèse de doctorat

NNT : 2022UPASM031

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre : Modélisation et prévision des variables d'exploitation ferroviaire et de flux de voyageurs en zone dense

Mots clés : Transport ferroviaire urbain, Temps de stationnement, Déplacements des voyageurs à bord, Prévision à court terme, Modèles d'apprentissage automatique, Graphes et réseaux

Résumé : Grâce à ses rames connectées, Transilien mesure en temps réel le nombre de montées et de descentes par porte du train. Nous contribuons à une meilleure synchronisation en phase opérationnelle des flux de trains et de voyageurs à l'aide de ces données uniques. Nous évaluons plusieurs modèles d'apprentissage statistique afin d'estimer les temps de stationnement en fonction des variables d'exploitation ferroviaire et des flux de voyageurs. Ces modèles permettent d'isoler des situations critiques où les flux de voyageurs impactent les temps de stationnement. Nous prévoyons chacune des variables, à l'horizon d'un arrêt, à partir de modèles autorégressifs bidirectionnels exploitant leur passé proche. Ces modèles se simplifient grâce aux motifs issus de la grille horaire. Nous estimons enfin des taux d'occupation par zone des rames traversantes, afin d'informer les voyageurs sur le confort à bord et proposons deux modèles de déplacement des voyageurs à bord.

Title : Modeling and forecasting of railway operation variables and passenger flows for dense traffic areas

Keywords : Urban rail transit, Dwell time, On-board passenger's movements, Short-term forecasting, Machine learning models, Graphs and networks

Abstract : Thanks to its new connected trains, Transilien is now able to measure the number of passengers boardings and alightings per train door in real time. Our research uses this unique dataset to contribute to a better synchronization of train and passenger flows during the operational stage of railway operations. We first evaluate several statistical learning models to estimate dwell time as a function of railway operation and passenger flows variables. These models allow us to isolate critical situations where passenger flows significantly impact dwell time. Our research forecasts each of the railway operation and passenger flows variables, one stop ahead, from bidirectional autoregressive models exploiting their recent past. We then simplify those models' using patterns derived from timetables. Finally, we estimate the occupancy rate by zone of the open gangways rolling stocks in order to inform passengers about the comfort on board and build two models of on-board passenger movements.

Table des matières

Remerciements

Cette thèse a été l'occasion de me mettre sur les rails de la science des transports. J'ai apprécié mêler apprentissage statistique et science des transports durant 3 années de travail, de sueur et des larmes. C'est une véritable joie de partager ce manuscrit qui est une synthèse de ces deux mondes. Merci à tou.te.s de m'avoir aidé à tracer mon sillon. L'apprentissage statistique en science des transports reste encore marginal. Nous ne sommes qu'au début du voyage !

La thèse CIFRE est une chance pour les laboratoires de recherche comme pour les entreprises. Elle permet aux entreprises de prendre le temps de se poser des questions fondamentales. Elle permet aux laboratoires d'identifier de nouvelles thématiques avec des applications directes comme le comptage des voyageurs pour SNCF Transilien.

Ces remerciements sont l'occasion d'exprimer ma profonde gratitude à toutes celles et ceux qui m'ont donné de leur temps sous toutes ses formes : échange de vive voix, présentations, cours, échange à distance, etc. Je remercie ceux qui m'ont guidé parfois malgré eux dans ce travail extrêmement riche. Tout oubli dans cette litanie des remerciements serait malheureux.

Merci Gilles et Christine pour votre encadrement de thèse que je souhaite à tout.e doctorant.e. Vous avez été à mes côtés à chaque instant et grâce à votre présence mon travail de thèse a avancé envers et contre tout durant ces trois années. Vous m'avez permis de passer d'étudiant à chercheur au prix de nombreuses heures de travail. Lorsque vous m'avez vu vaciller, vous m'avez soutenu. La recherche, c'est de la joie mais c'est aussi beaucoup de déceptions : des idées qui ne fonctionnent pas, un grain de sable qui change tout, une erreur dans un script... Merci à tous les deux. Vous êtes complémentaires et cohérents, Gilles ton exigence dans la rédaction, notamment en anglais, fait de toi un exemple à suivre. Christine, ta facilité à traduire une formulation mathématique en expression numérique sur R m'a toujours bluffé. Grâce à toi Gilles, my english has improved slightly since the beginning, et grâce à toi Christine, je m'entraine à faire des phrases courtes. Merci beaucoup à tous les deux. J'ai pris conscience à la fin de ces trois années de l'importance d'être extrêmement rigoureux et intègre lorsqu'on fait de la recherche, en mathématique comme en transport. Nous sommes les garants de la qualité des futures recherches. Désolé si, à vos yeux, je n'étais parfois pas suffisamment rigoureux, je le serai pour les années à venir. Merci à tous les deux de m'avoir ouvert ce vaste terrain de recherche qu'est celui de l'interaction entre statistique et science des transports. Dans ce domaine, je pense que les statistiques ont 10-20 ans de retard mais ce n'est pas grave car nous sommes là. Merci à tous les deux de m'avoir fait confiance dans la phase finale de rédaction du manuscrit que j'aurais dû amorcer 3 mois plus tôt mais qui s'est, au prix de sueur et de larmes, globalement bien passée.

Merci Christine de m'avoir sélectionné pour intégrer le master que tu dirigeais d'une main de maître. Tu es une statisticienne idéale ayant à la fois une grande rigueur dans la formalisation mathématique et une agilité naturelle à exprimer numériquement tes idées. Durant ces années de thèse, tu as été une véritable source d'inspiration pour

les sujets les plus compliqués allant du co-clustering des gares et des jours jusqu'à la modélisation des mouvements des passagers à l'aide d'un modèle de Markov caché.

Merci Gilles, on s'est dit « oui » alors que j'étais devant la mare d'Orsay, dans laquelle j'ai appris plus tard qu'il y avait des tortues. Tu m'as alors donné une citation qui ressemblait étrangement à celle de Churchill « Je n'ai à offrir que du sang, du labeur, des larmes et de la sueur ». Je m'engageais alors pour une pérégrination intellectuelle de 3 ans et demi. J'ai appris beaucoup de choses : (1) la recherche scientifique demande de la patience, énormément de rigueur et d'abnégation ; (2) le choc des images, il m'a fallu plusieurs années avant de me rendre compte que l'important est de rendre visuelles nos idées ; (3) malgré les doutes il faut garder la foi. Merci Gilles d'avoir été un encadrant de thèse hors pair.

Quelques expressions qui resteront liées à ma thèse : « Amen » qui me rappelle à mon absence de catéchisme, « Soit » lorsqu'il n'y avait rien à ajouter, « C'est pas faux » qui inquiète grandement Christine. Nous avons partagé une grande intimité tout au long de ces semaines. Merci pour votre confiance. J'espère pouvoir travailler avec vous dans le futur sur des sujets aussi importants que la prévision d'affluence ou de retard dans les transports en commun, le clustering de gares ou la modélisation des mouvements des passagers à bord des trains.

Merci à Pierre d'avoir été cherché un laboratoire de mathématiques, en particulier une équipe de statisticiens/probabilistes, non des plus appliqués, pour attaquer un problème en science des transports qui en ouvre plein d'autres. Nos échanges brefs mais intenses au siège de Transilien ou à l'agence Tram-Train m'ont toujours permis d'avancer. Ta vision pour Transilien était riche et portait un ensemble de sujets stratégiques. Merci Pierre. Marc, merci. Tu m'as offert d'être confortable dans mes travaux de recherche. J'ai pu avancer sans me soucier des contingences matérielles : recrutement de stagiaire, congés, etc. J'ai beaucoup appris au Lab' avec toi notamment en termes de management d'équipe. En bref, ton enseignement serait le suivant : l'important est de faire confiance à ses équipes.

Je remercie chaleureusement Yohann et Francesco d'avoir accepté d'être rapporteur dans mon jury de thèse. Leurs retours sont précieux. En espérant que dans 10 ans, nous arrivions à construire une excellence française mêlant science des transports et apprentissage statistique. Je remercie chaleureusement Élisabeth d'être présidente de mon jury ainsi qu'Étienne d'avoir accepté d'être examinateur. Étienne tu m'as transmis le goût de la visualisation de données de mobilités. Je serai ravi de collaborer avec toi sur des sujets en lien avec les données de SNCF-Transilien.

Merci Mathilde pour ton aide de tous les instants lors de cette thèse. Le Chapitre 5 est très largement inspiré de travaux que nous avons mené ensemble. Notre collaboration a été fructueuse aussi bien intellectuellement qu'industriellement. J'ai dû ménager du temps pour ma thèse mais vraiment cette première expérience d'encadrement m'a donné envie de continuer d'encadrer des stages. Merci donc aux suivants, à Joshua, Laura, Martine et Marine pour votre entrain et vos idées toujours intéressantes. J'ai beaucoup appris à votre contact. J'espère vous avoir un peu transmis de mon goût pour la recherche : prendre un sujet, le creuser, le

décortiquer, le conceptualiser et analyser les résultats. J'espère pouvoir suivre vos travaux dans les années à venir. Merci beaucoup Valentine, tu as été l'étoile de cette dernière année de thèse. Tu as su prendre en main des sujets compliqués qui demandaient une bonne capacité d'adaptation. Merci pour ta compétence et ta bonne humeur. Je te souhaite beaucoup de réussite dans le futur.

La présidence de l'association des doctorants de la SNCF fut l'occasion de prendre conscience de la richesse de la SNCF. Merci à la DTIPG d'apporter aux doctorants de nouvelles compétences notamment grâce au programme « Ma thèse en 180s ». Merci à Laurent d'être venu au LMO faire travailler des doctorants dans le cadre d'une semaine en entreprise sur les données de ma thèse, quelques mois avant le début du projet. Laurent et Tom, nous avons mené plusieurs projets qui n'ont pas toujours eu les retombés espérées mais que nous avons menés haut les cœurs, en collaboration avec le plateau IA du groupe SNCF. Merci à Lisa et à tout le Lab' pour son accueil, sa tolérance et ses coups de mains.

Merci à tous les membres de Transilien pour votre confiance lorsque nous nous attaquions à des sujets avec le DataLab'. Vous m'avez toujours donné de précieux conseils aussi bien concernant la richesse des données de Transilien que sa stratégie pour les prochaines années. Je suis impatient de relever avec vous les prochains défis qui attendent Transilien.

Merci Zoi, pour l'opportunité que tu m'as donnée d'enseigner à l'École des Ponts, un projet qui me tenait à cœur autour du calcul de marge au niveau des temps de stationnement théoriques. Merci Capucine-Marin, tu as été un ancrage mouvant mais un ancrage tout de même pour me guider lors de mes débuts à la SNCF et dans l'univers de la recherche en science des transports.

Thank you to Taku, Georgio, Gerrit, Andrew, ... for our discussions on many different transportation subjects. Thank you very much Oded for your very interesting lecture on transport planning. We initiate an interesting collaboration then on platform passengers' strategies. Thank you very much Carl-William for your visit at Lab' MTA of SNCF-Transilien. We are starting a fruitful collaboration with Swedish network using APC data to compute dwell time margins. Thank you Erik for your proposal to review an interesting paper on short-term forecasting of crowding. Merci Jeremy, Noëlie, Fatma, Latifa pour vos visions sur un des trois sujets de cette thèse. Merci Maguelonne et Xavier de m'avoir montré que l'on peut faire des statistiques à la SNCF. Le projet Reluance était précurseur !

Je tiens à remercier toute ma famille ainsi que ma belle-famille pour leur soutien lors des confinements successifs. Merci à tou.te.s les relec.teurs.trices des chapitres. Toutes erreurs restantes sont de ma propre responsabilité.

Merci à toi Louise. Tu as toujours été présente pour m'accompagner dans cette aventure. Merci pour ta patience et ton soutien. Tu es ma plus fidèle lectrice. Tu as su m'écouter et me rassurer quand le train déraillait. Et si nous avons vécu ces trois années avec joie, c'est grâce à toi. Merci.

Notations

Cette Table résume les notations communes aux différents chapitres. L'indexation des variables diffère d'un chapitre à l'autre, par exemple pour une variable générique x, l'indexation sera : $x_{k,s,d}$ dans le Chapitre 3 ; $x^d_{k,s}$ dans le Chapitre 4 et $x^{k,d}_{s,i}$ dans le Chapitre 5. En effet, les indices définissent un changement de modèle tandis que les exposants caractérisent les répétitions d'un même phénomène stochastique. Cette logique n'est toutefois pas respectée pour le Chapitre 3 où les répétitions sont en indice car l'exposant est réservé au caractère théorique (theo) ou observée (obs) des variables d'exploitation ferroviaire.

k	Numéro de train
s	Indice de gare
d	Indice du jour
i	Indice de zone
I	Nombre de zones
S	Nombre de gares
K	Nombre de trains
D	Nombre de jours
$\mathcal{N}^d$	Ensemble des arrêts (k, s) pour un jour d
a^{theo}	Heure d'arrivée théorique
d^{theo}	Heure de départ théorique
y^{theo}	Temps de stationnement théorique $(d^{\text{theo}} - a^{\text{theo}})$
a^{obs}	Heure d'arrivée observée
d^{obs}	Heure de départ observée
y^{obs}	Temps de stationnement observé $(a^{\text{obs}} - a^{\text{theo}})$
Δa	Écart à l'heure d'arrivée ou retard à l'arrivée $(a^{\text{obs}} - a^{\text{theo}})$
z	Régime de ponctualité (z = 1 si le train est en avance, z = 2 si le train est à l'heure et z = 3 si le train est en retard)
b	Nombre de montées à l'échelle du train
a	Nombre de descentes à l'échelle du train
ℓ	Charge à bord à l'échelle du train
c	Taux d'occupation à l'échelle du train (ℓ/capacité)
m	Affluence à la porte critique
b_i	Nombre de montées en zone i
a_i	Nombre de descentes en zone i
ℓ_i	Charge à bord en zone i
w_i	Déplacements jusqu'en zone i
$x_{1:(k-1),s}$	Voisinage en gare $\left(x_{1,s}, \ldots, x_{(k-1),s}\right)$, Chapitre 4
$x_{k,1:(s-1)}$	Voisinage en train $\left(x_{k,1}, \ldots, x_{k,(s-1)}\right)$, Chapitre 4
$\boldsymbol{x}^k_s$	Vecteur par zones $\left(x^k_{s,1}, \ldots, x^k_{s,i}, \ldots, x^k_{s,I}\right)$, Chapitre 5
$x^k_{\bullet,i}$	Somme par trajet en zone i $\left(\sum_{s=1}^{S} x^k_{s,i}\right)$, Chapitre 5
$x^k_{\bullet,\bullet}$	Somme par trajet à l'échelle du train $\left(\sum_{i=1}^{I}\sum_{s=1}^{S} x^k_{s,i}\right)$, Chapitre 5

1 Introduction générale

L'objectif de cette thèse est de développer des modèles d'estimation et de prévision des variables d'exploitation ferroviaire et de flux de voyageurs en *Mass Transit*. Le Mass Transit est l'exploitation des trains de banlieue en zone dense. L'introduction débute par une rapide présentation des enjeux et du jeu de données. Elle est structurée autour des contributions des chapitres de la thèse. Elle se termine par un guide de lecture ainsi que la liste des publications associées.

Contents

1.1 Contexte et notations

Cette partie est l'occasion de présenter succinctement les éléments communs aux différents chapitres : le Mass Transit ouvert qui motive le sujet de la synchronisation des flux de voyageurs et des flux de trains, ainsi que le jeu de données qui est essentiel au développement des modèles statistiques. L'ensemble des notations qui rendent familiers des objets mal définis au début de la thèse sont décrit dans la Table des notations.

1.1.1 Enjeux du Mass Transit ouvert

Le Mass Transit ouvert est l'exploitation des trains de banlieue dans un contexte où la synchronisation des flux de voyageurs et des flux de trains est nécessaire pour maximiser la capacité [1] du réseau. En Mass Transit ouvert, l'exploitation des trains se rapproche de celle de l'exploitation des métros car les circulations sont très rapprochées et les voyageurs arrivent de façon continue dans les gares. Les enjeux du Mass Transit sont à relever à toutes les étapes de la planification ferroviaire : de la commande des rames jusqu'à la gestion des circulations en passant par le dimensionnement des temps de stationnement. Les opérateurs suivent en temps réel la progression des trains sur le réseau pour gérer au mieux les circulations. Transilien, l'opérateur de trains de banlieue d'Île-de-France, suit de surcroit les flux de voyageurs sur son réseau. La connaissance précise et en temps réel de ces flux de voyageurs est nécessaire pour gérer les circulations en fonction des volumes de voyageurs transportés.

Les flux de voyageurs se concentrant aux heures de pointe, leur représentation sur une journée forme une courbe bi-modale les jours de semaine (cf. Figure 2.3). Les autorités organisatrices de transport dimensionnent le réseau en fonction du flux de voyageurs maximal en pointe. Pour cette raison, l'opérateur exploite généralement un réseau proche de la saturation aux heures de pointe. Il doit chercher à toujours mieux comprendre et anticiper l'interaction entre les flux de voyageurs et les flux de trains afin d'augmenter la capacité globale du réseau ainsi que le confort des voyageurs.

1.1.2 Jeu de données de la thèse

Le jeu de données de la thèse comprend l'ensemble classique [RO] des données d'exploitation ferroviaire par train k : heures d'arrivée a^{theo} et de départ d^{theo} théoriques, heures d'arrivée a^{obs} et de départ d^{obs} observées. Ces quantités permettent de calculer le temps de stationnement observé ou théorique ainsi que le retard à l'arrivée ou au départ. Les heures d'arrivée et de départ observées sont

1. La capacité d'un réseau étant le nombre maximal de trains pouvant circuler sur une portion d'infrastructure à un temps donné [Hansen and Pachl, 2014].

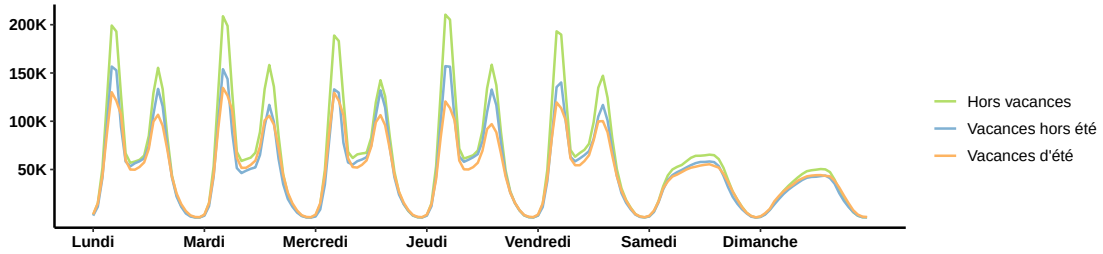


FIGURE 1.1 – Fréquentation hebdomadaire par heure sur le réseau Transilien (moyennée sur l'année 2019 hors mois de décembre pour cause de grève). La fréquentation est mesurée en nombre d'entrées dans toutes les gares de Transilien exceptées celles partagées avec la RATP.

très précises car elles proviennent de la mesure des tours de roue par la rame et non de balises au sol. La disponibilité de l'information précise et en temps réel des flux de voyageurs [PF] rend notre jeu de données original. Les flux de voyageurs (le nombre b de montées, le nombre a de descentes, etc.) sont mesurés directement par le train à chaque arrêt à l'aide de capteurs infra-rouges au niveau de chaque porte du train. La combinaison des deux jeux de données [RO + PF] est le terreau fertile de ces travaux de thèse. Les jeux de données sont utilisés selon différents périmètres spatiotemporels précisés dans chacun des chapitres.

1.2 Modélisation des temps de stationnement (Chapitre 3)

Au commencement du projet de thèse, la commande de Transilien était :

modéliser le temps d'échange pour mieux planifier et mieux gérer en opérationnel le temps de stationnement.

Le premier constat des débuts de travaux de thèse est que le temps d'échange est mal défini et mal mesuré à Transilien, voir en ce sens l'Annexe D. La problématique a donc évolué de la modélisation du temps d'échange vers celle du temps de stationnement (y^{obs}). Les temps de stationnement sont non seulement bien définis (écart entre le dernier et premier tour de roue) mais également bien mesurés, ce qui en fait un sujet idéal pour la thèse.

Toutefois, dans le cas général, il n'est pas facile de déduire un temps d'échange du temps de stationnement en particulier parce que celui-ci dépend du statut du train par rapport à son heure d'arrivée théorique (à l'avance, à l'heure ou en retard). Par exemple sur la Figure 1.3, les temps de stationnement des trains en avance (en vert) croissent d'autant que le train est en avance. La situation se simplifie dans le cas particulier où il n'y a pas de marge (en vert) sur la Figure 1.2. .

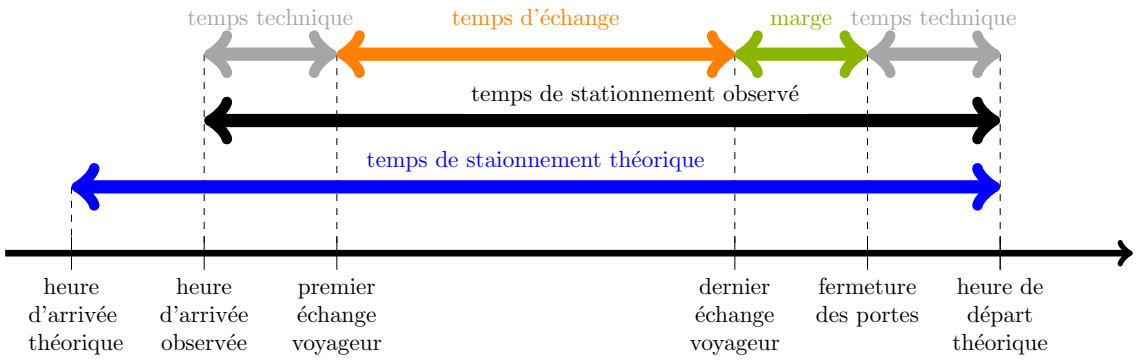


FIGURE 1.2 – Décomposition du temps de stationnement.

Ainsi, l'enjeu du Chapitre 3 est d'isoler les situations où les temps de stationnement sont contraints par les flux de voyageurs *i.e.* sans marge. L'utilisation de modèles d'apprentissage statistique classiques permet d'identifier ces situations à partir de quatre jeux de variables. Le premier [PF] composé uniquement des flux de voyageurs à l'échelle du train (nombre b de montées, nombre a de descentes et taux d'occupation c). Le second [RO] composé uniquement des variables d'exploitation ferroviaire (notamment l'écart à l'heure d'arrivée théorique Δa et le temps de stationnement théorique y^{theo}). Le troisième [RO + PF] est la combinaison des deux précédents. Le quatrième [RO + PF + M] est augmenté de l'affluence à la porte critique (m) [2].

1.2.1 Synthèse des contributions

Dans ce chapitre, l'objectif est de modéliser le temps de stationnement y^{obs} comme une fonction aléatoire d'un jeu de variables X défini ci-dessous :

$$y^{\text{obs}} = f(X) + \varepsilon, \tag{1.1}$$

où f est la fonction de régression et ε est un bruit aléatoire qui ne dépend pas de X. L'enjeu est de comparer six fonctions de régression différentes et les quatre jeux de variables X que nous avons présentés à la section précédente. La Table 1.1 récapitule les performances obtenues pour ces vingt-quatre configurations, en particulier, nous mesurons le gain à utiliser des méthodes non linéaires. Nous présentons brièvement ici les six familles de modèles d'apprentissage, trois issues d'une modélisation linéaire, trois autres issues d'une modélisation non linéaire. Les modèles de régression linéaire classiques expriment la fonction de régression comme une fonction linéaire des variables. S'il y a des variables qui ne sont pas quantitatives (par exemple, le régime de ponctualité z), elles sont transformées en autant de variables indicatrices qu'il y a de niveaux à cette variable. On introduit des interactions entre variables (*i.e.* leurs produits), ce qui permet d'élargir la complexité du modèle. Ainsi, la méthode reste linéaire, mais dans un espace de

2. L'affluence à la porte critique est la valeur maximale des voyageurs échangés (nombre de montées plus descentes) par porte.

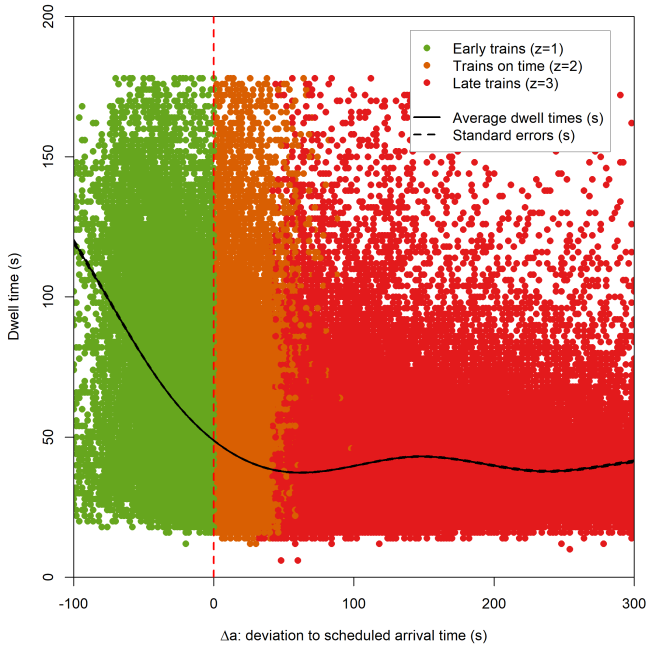


FIGURE 1.3 – Temps de stationnement en fonction de l'écart à l'heure d'arrivée $\Delta a = a^{\text{obs}} - a^{\text{theo}}$. Chaque point est un arrêt d'un train colorié en fonction de son régime de ponctualité, en avance (en vert), à l'heure (en orange) ou en retard (en rouge).

co-variables plus complexe. Par exemple, l'équation 1.2 formalise un modèle de régression linéaire avec effets multiplicatifs de l'écart à l'heure d'arrivée théorique Δa selon le régime de ponctualité z :

$$
\begin{aligned}
& f(s, w_k,\, t_{k,d},\, y^{\text{theo}}_{k,s,d},\, \Delta a_{k,s,d},\, z_{k,s,d},\, b_{k,s,d},\, a_{k,s,d},\, c_{k,s,d},\, m_{k,s,d}) \qquad (1.2)\\
&= \left.\begin{aligned} & \beta^0 + \beta^{\text{way}} \mathbb{1}_{[w_k=1]} + \sum_{s'=1}^{S-1} \beta^{\text{station}}_{s'} \mathbb{1}_{[s=s']} \end{aligned}\right\} \text{dans tous les cas}\\
&\quad \left.\begin{aligned} & + \sum_{z\in\{1,2,3\}} \mathbb{1}_{[z_{k,s,d}=z]}\, \beta^{(\Delta a)}_z\, \Delta a_{k,s,d} \end{aligned}\right\} \begin{array}{l}\text{variables RO,}\\ \text{effet multiplicatif } \Delta a \text{ et } z\end{array}\\
&\quad \left.\begin{aligned} & + \beta^{(y)}\, y^{\text{theo}}_{k,s,d} + \beta^{\text{type}} \mathbb{1}_{[t_{k,d}=\text{double}]} + \beta^{\text{early}} \mathbb{1}_{[z_{k,s,d}=1]} + \beta^{\text{late}} \mathbb{1}_{[z_{k,s,d}=3]} \end{aligned}\right\} \text{RO}\\
&\quad \left.\begin{aligned} & + \beta^{(a)}\, a_{k,s,d} + \beta^{(b)}\, b_{k,s,d} + \beta^{(c)}\, c_{k,s,d} \qquad + \beta^{(m)}\, m_{k,s,d} \end{aligned}\right\} \text{PF+M}
\end{aligned}
$$

où w_k représente le sens de circulation et $t_{k,d}$ la composition du train (simple ou double rames). Le but du modèle de régression linéaire avec effets multiplicatifs selon le triplet (sens, gare et régime de ponctualité) est d'étendre les interactions à d'autres variables. Ces transformations sont implicitement effectuées par les autres méthodes d'apprentissage automatique que nous avons testés que ce soit pour les modèles d'ensemble à base d'arbres (forêts aléatoire et gradient boosting) ou les réseaux de

TABLE 1.1 – Performances en termes de MAE des différents modèles d'apprentissage statistique en ligne avec les différents jeux de variables en colonne. Les colonnes indiquent le jeu de variable utilisé : comprenant seulement les flux de voyageurs [PF], comprenant seulement les variables d'exploitation ferroviaire [RO], la combinaison des deux [PF + RO] ou la combinaison des deux augmentée de l'affluence à la porte critique [RO + PF + M].

Modèles	MAE			
	PF	RO	RO PF	RO PF+M
1. Régression linéaire avec effets additifs	13.7	10.5	10.2	10.1
2. Régression linéaire avec effets multiplicatifs Δa selon z	13.7	9.1	8.9	8.8
3. Régression linéaire avec effets multiplicatifs selon le triplet (sens, gare et régime de ponctualité)	13.3	8.8	8.3	8.3
4. Forêt aléatoire	13.7	8.4	8.1	8.0
5. Gradient boosting avec arbres de régression	12.9	8.5	8.0	7.9
6. Réseau de neurones	12.7	8.4	8.0	8.0

neurones. L'évaluation des modèles et des jeux de variables du Chapitre 3 est faite classiquement en divisant le jeu de données en un jeu de données d'entrainement et un jeu de données test. Les modèles d'apprentissage statistique nécessitent de fixer un ensemble d'hyperparamètres (*tunning parameters* en anglais) qui sont optimisés sur le jeu de données d'entrainement par validation croisée.

#1 Évaluation des modèles : les modèles d'apprentissage automatique sont équivalents aux modèles de régression linéaire avec interactions

Le modèle linéaire répond à un aspect explicatif : comprendre la corrélation de la variable réponse, le temps de stationnement, avec les différentes variables explicatives. Des procédures non linaires issues de l'apprentissage automatique, mettent l'accent sur l'atteinte d'un objectif de bonne estimation, au risque de perdre des clés de compréhension du phénomène sous-jacent.

Dans la Table 1.1, le modèle de régression linéaire avec effet additif est largement moins performant que les modèles d'apprentissage statistique à partir du moment où le jeu de variables [RO] est utilisé. Cette contre-performance s'explique par la relation non linéaire, en forme de coude, entre la durée des temps de stationnement et l'écart à l'heure d'arrivée théorique, voir Figure 1.3. La relation en forme de

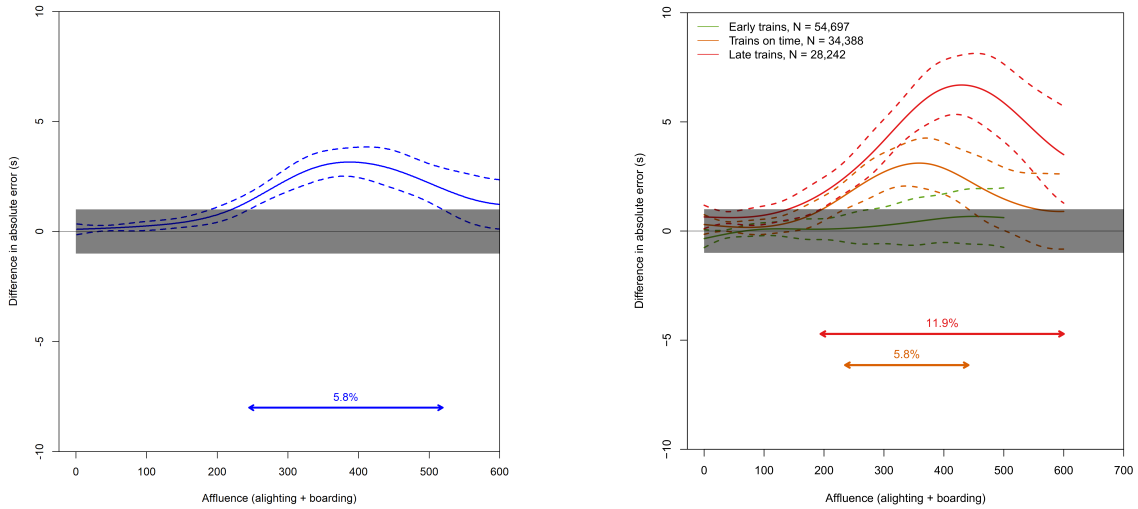


FIGURE 1.4 – Différence des erreurs moyennes en valeurs absolues du modèle de forêt aléatoire ne comprenant que les variables [RO] (RF–RO) et celui comprenant toutes les variables (RF–All). La différence des erreurs MAE pour la modélisation des temps de stationnement (en ordonnée) en fonction du volume de voyageurs (en abscisse). Les résultats sont globaux (à gauche), suivant les régimes de ponctualité (à droite). Chaque courbe représente la moyenne, conditionnelle à l'affluence, de la différence des MAE de RF–RO et de RF-All calculée à partir d'un modèle additif généralisé. Les flèches horizontales indiquent la plage d'observation où une amélioration moyenne significative de RF–All sur RF–RO est observée ; les pourcentages indiquent la part d'observations concernées.

coude, en noir sur la Figure 1.3, est une moyenne conditionnelle des temps de stationnement en fonction de l'écart à l'heure d'arrivée théorique estimée grâce à un modèle additif généralisé. La prise en compte d'interaction entre l'écart à l'heure d'arrivée théorique et le régime de ponctualité réduit considérablement l'écart de performances entre les modèles de régression linéaire et ceux d'apprentissage automatique. L'ajout de variables d'interactions multiplicatives en fonction de la gare, du sens et du régime de ponctualité fait que le modèle de régression linéaire a des performances similaires aux modèles d'apprentissage automatique. Le nombre de paramètres des modèles de régression linéaire reste bien inférieur, malgré l'ajout successif de variables pour améliorer les performances d'estimation, à ceux des modèles d'apprentissage automatique où le nombre de paramètres est énorme que ce soit pour les forêts aléatoires ou les réseaux de neurones. L'évaluation des performances des modèles de régression linéaire avec effets multiplicatifs précise certaines non linéarités entre les temps de stationnement et les variables explicatives. Enfin, pour calculer des résultats locaux dans le Chapitre 3, nous nous concentrons sur les modèles de forêts aléatoires, notés RF, directement comparables au modèle de Kecman and Goverde [2015].

#2 Évaluation des variables : [PF] $<$ [RO] $\leq$ [RO + PF] $\approx$ [RO + PF + M]

Le jeu de données permet de confirmer l'intuition de Hansen et al. [2010] et Kecman and Goverde [2015] : les variables les plus importantes dans la modélisation des temps de stationnement des trains de banlieue sont bien les variables d'exploitation ferroviaire, notamment Δa. Dans la Table 1.1, quel que soit le modèle sélectionné, l'erreur moyenne absolue globale décroit de plusieurs secondes entre le jeu de variables ne comprenant que des variables de flux voyageurs [PF] et celui ne comprenant que des variables d'exploitation ferroviaire [RO]. L'ajout des variables de flux de voyageurs à celles des variables d'exploitation ferroviaire a un faible impact global, l'erreur moyenne absolue ne décroissant que de 0.5 s. L'ajout de l'affluence à la porte critique ne fait décroitre l'erreur moyenne absolue que de quelques dixièmes de seconde.

#3 [RO] $\ll$ [RO + PF + M] dans les situations critiques

Les variables de flux de voyageurs semblent ne pas avoir un impact important, au global, sur l'estimation de l'ensemble des temps de stationnement, cf. Table 1.1. Pour autant, dans certaines situations dites critiques, elles sont importantes.

Les deux graphes de la Figure 1.4, construits sur le même principe, permettent de caractériser ces situations critiques. Ils représentent la différence de l'erreur moyenne conditionnelle en valeur absolue (MAE) entre le modèle de forêt aléatoire RF–RO ne contenant que les variables [RO] et le modèle de forêt aléatoire RF–All contenant toutes les variables [RO+PF+M] (et en particulier les variables de flux de voyageurs). Le graphe de droite prend en compte l'information de ponctualité, tandis que celui de gauche est global. La valeur moyenne représentée par une courbe en fonction du volume de voyageurs (nombre de montées plus nombre de descentes) est estimée à l'aide d'un modèle additif généralisé. Le graphe de gauche de la Figure 1.4 montre qu'il est intéressant d'ajouter des variables de flux de voyageurs quand ceux-ci sont importants. La Figure 1.4 de droite permet de confirmer l'intuition de Pedersen et al. [2018] et de Medeossi and Nash [2020] selon laquelle les trains en retard sont des trains avec des temps de stationnement contraints par les flux de voyageurs. D'une part, le gain de performance de RF–All sur RF–RO est visible à partir d'une affluence plus faible pour les trains en retard que pour les trains à l'heure. D'autre part, RF–All n'est jamais meilleur que RF–RO pour les trains en avance, tandis que le nombre d'observations concernées par une meilleure performance de RF–All sur RF–RO croit avec l'augmentation de l'écart à l'heure d'arrivée. Autrement dit, plus les trains arrivent après leur heure d'arrivée théorique, plus les variables de flux de voyageurs impactent les temps de stationnement, aussi bien en nombre d'observations qu'en intensité. En résumé, l'analyse locale des performances montre que l'intégration des flux de voyageurs dans la modélisation des temps de stationnement a un intérêt pour les situations critiques, les situations de forte affluence ou de trains en retard (non contraints par leur heure de départ théorique).

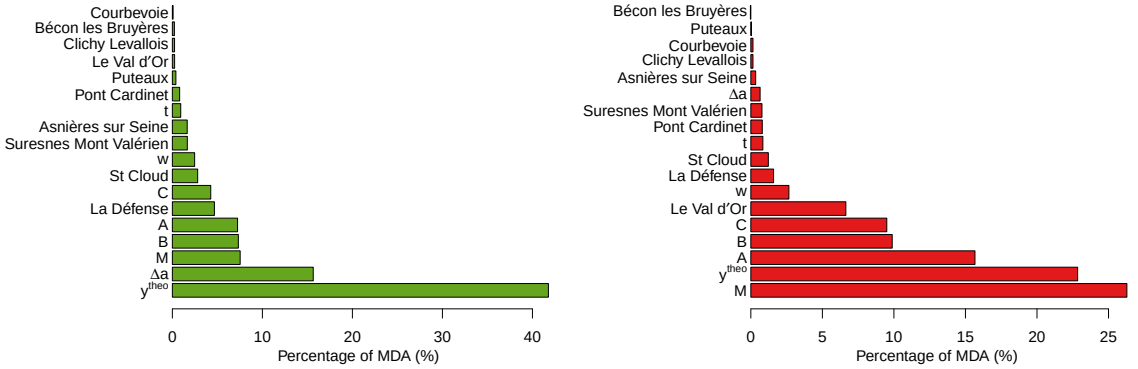


FIGURE 1.5 – Les variables les plus importantes suivant l'indice MDA normalisé pour une forêt aléatoire entrainée avec RO + PF + M sur une partition en fonction du régime de ponctualité du jeu de données. Les trains en avance (à droite), les trains en retard (à gauche).

#4 Importance des variables en fonction du régime de ponctualité

L'importance des variables est calculée lors de l'estimation des paramètres d'une forêt aléatoire. Celle-ci se traduit dans le MDA (*mean decrease in accuracy*) qui est calculé en permutant de façon aléatoire les valeurs prises par chacune des variables. Si une variable est importante dans la construction d'un arbre, une permutation aléatoire de ses valeurs sur les différentes observations conduira à une dégradation significative des performances du modèle. Ce point est détaillé dans la Section 3.4.4.

La Figure 1.5 représente l'importance des variables calculée pour deux modèles de forêts aléatoires. La première forêt aléatoire (à gauche, en vert) est entrainée uniquement sur des observations de trains en avance : les variables d'exploitation ferroviaire, notamment l'écart à l'heure d'arrivée, ressortent comme les variables les plus importantes, ce qui confirme les résultats de la Table 1.1. La seconde forêt aléatoire (à droite, en rouge) est entrainée uniquement sur des observations de trains en retard : les variables de flux de voyageurs sont les plus importantes, notamment l'affluence à la porte critique.

Les quatre grands résultats présentés ici avec les données de la ligne L entre janvier 2018 et septembre 2019 ont été reproduits dans l'Annexe 3.B avec les données de la ligne H sur la même période. Ces résultats ouvrent deux perspectives, d'une part, pour le calcul des marges au niveau des temps de stationnement à partir des trains en retard, d'autre part, pour l'utilisation du modèle en opérationnel.

1.2.2 Calcul des marges des temps de stationnement en phase tactique (Annexe D)

Ce travail permet de calculer *a posteriori* la marge effective des temps de stationnement théoriques, représentés sur la Figure 1.7 de gauche pour cinq trains de la ligne H en 2019. Les résultats du Chapitre 3 sur la modélisation des temps de

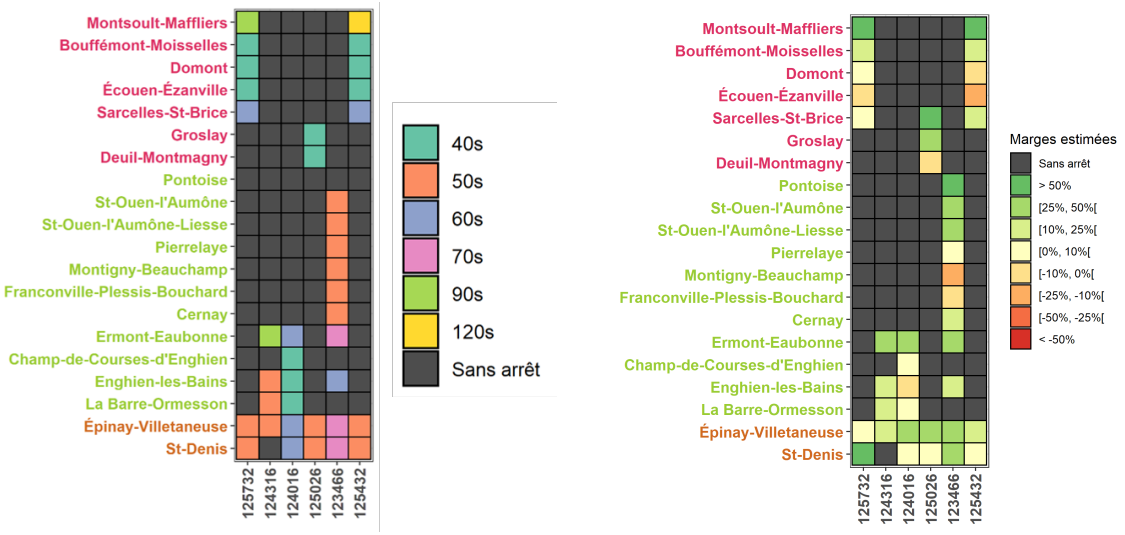


FIGURE 1.6

FIGURE 1.7 – Exemple de calcul des marges pour 5 trains de la ligne H allant vers Paris en heures de pointe du matin de janvier à septembre 2019 pour les jours ouvrés. À gauche, les temps de stationnement théoriques. À droite, les marges estimées par la méthode des trains en retard.

stationnement confirment que les temps de stationnement des trains en retard sont proches d'un temps de stationnement sans marge. Ce modèle des temps de stationnement sans marge permet d'estimer *a posteriori*, quelque soit le régime de ponctualité, un temps de stationnement sans marge en fonction des flux de voyageurs. L'idée est de calculer la différence relative entre les temps de stationnement théoriques et les temps de stationnement sans marge estimés *a posteriori*. Cette différence est égale à la marge effective qui est représentée sur la Figure 1.7 de droite. Les marges effectives sont entre 10 et 50 % du temps de stationnement théorique. Une analyse plus précise développée dans l'Annexe D a permis de montrer que la marge effective représente entre 40 et 50 % des temps de stationnement théoriques pour la ligne R et N. Les perspectives de ce travail d'analyse *a posteriori* des marges sont nombreuses pour Transilien. Elles permettraient, d'une part, d'optimiser en phase tactique les temps de stationnement théoriques en calculant le temps de stationnement sans marge puis en ajoutant la marge souhaitée, d'autre part, ce travail permet d'identifier les temps de stationnement non respectés à cause des volumes de voyageurs trop importants.

1.2.3 Temps de stationnement en phase opérationnelle

La prévision des temps de stationnement du ou des prochains arrêts est un pré-requis pour les prendre en compte lors de la gestion opérationnelle des circulations. Les temps de stationnement prédits permettraient d'indiquer aux conducteurs le meilleur moment pour enclencher la séquence de fermeture des

portes. Ils pourraient également permettre aux centres opérationnels d'anticiper les retards liés à un trop grand volume de voyageurs. Cependant, la modélisation des temps de stationnement du Chapitre 3 repose sur un ensemble de variables explicatives inconnues aux prochains arrêts. L'objectif du Chapitre 4 est de prédire ces variables à l'arrêt suivant pour les utiliser ensuite dans le modèle du Chapitre 3. Les principales variables explicatives à prédire sont : le nombre b de montées, le nombre a de descentes, la charge à bord ℓ et l'écart à l'heure d'arrivée théorique Δa. La méthode de prévision proposée est suffisamment générique pour être également testée sur la prévision des temps de stationnement y^{obs}.

1.3 Modèles de prévision à court terme (Chapitre 4)

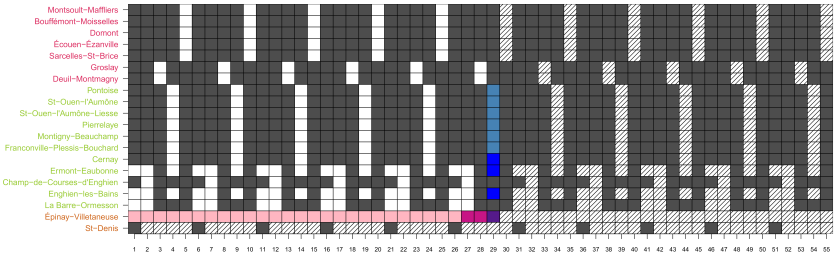
La prévision à court terme (à l'horizon d'une gare dans le Chapitre 4) des variables de flux de voyageurs et d'exploitation ferroviaire dessine une voie vers la prévision des temps de stationnement eu égard aux modèles construits dans le Chapitre 3. La prévision des retards [3] et de l'affluence est nécessaire, indépendamment des temps de stationnement, pour alimenter les canaux d'information des voyageurs comme les écrans à quai. Ces deux utilisations montrent l'importance de la prévision à court terme en science des transports et à Transilien. Le Chapitre 4 propose un modèle de prévision à court terme dont les principes de base sont applicables à toutes les variables : le nombre b de montées, le nombre a de descentes, la charge à bord ℓ, le retard à l'arrivée Δa ou le temps de stationnement y^{obs}.

Le Chapitre 4 est l'occasion de résoudre un problème mathématiquement intéressant car il nécessite d'identifier des structures de dépendances pertinentes pour la prévision à court terme. Les modèles de prévision à court terme, par exemple les modèles auto-régressifs pour les séries temporelles, reposent généralement sur l'exploitation de valeurs passées observées dans un passé proche de la réalisation à prédire. Cependant, sur un réseau de transport, la notion de proximité peut être définie dans deux directions, l'une spatiale, celle des gares précédentes pour un train donné k et l'autre temporelle, celle des trains précédents à une gare donnée s. Ces deux directions se rejoignent à l'arrêt (k, s), défini comme l'association du train k à la gare s. Ces liens spatio-temporels, permis notamment par la desserte de plusieurs gares par un même train, définissent une structure de dépendance peu étudiée dans la science des transports. Nous proposons dans ce chapitre plusieurs modèles auto-régressifs bi-directionnels en gare et en train.

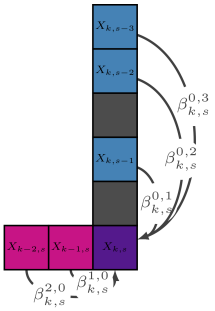
1.3.1 Synthèse des contributions

Les principales contributions du Chapitre 4 s'articulent autour de la généralisation d'un modèle de prévision à court terme dont les performances sont équivalentes à l'état de l'art mais dont la complexité est moindre.

3. Appelé « écart à l'heure d'arrivée théorique »dans le Chapitre 3.



(a) Représentation de la structure de graphe gares-trains issue de la grille horaire pour la pointe du matin. Les cases gris foncés indiquent qu'il n'y a pas d'arrêt. Les informations disponibles des gares et des trains précédents sont respectivement en bleu et en rose. Les bleu ou rose foncé indiquent les informations les plus récentes. Les cases hachurées représentent des informations futures ou non utilisées pour la prévision.



(b) Exemple de régression linéaire en forme de L pour un arrêt (k, s).

FIGURE 1.8 – Structure de graphe gares-trains avec voisinage et modèle de régression linéaire en forme de L.

#1 Un même type de modèle pour plusieurs variables

Le Chapitre 4 propose une modélisation unique pour prédire les flux de voyageurs (le nombre b de montées, le nombre a de descentes et la charge à bord ℓ) et les variables d'exploitation ferroviaire (le retard à l'arrivée Δa et le temps de stationnement observé y^{obs}).

La Figure 1.8a donne un exemple de structure de graphe gares-trains issu de la grille horaire où l'ordre des gares desservies par chaque train est parfaitement défini. La notion de graphe gares-trains est inspirée de la notion de graphe espace-temps (GET) où l'espace se résume aux gares et le temps à l'ordre de passage des trains. La grille horaire permet de définir un voisinage en forme de L (en forme de L inversé pour être exact) pour une variable générique x à un arrêt (k, s). Ce voisinage est composé d'un voisinage en gare (en rose sur la Figure 1.8a), noté $x_{1:k,s}$, et d'un voisinage en train (en bleu clair sur la Figure 1.8a), noté $x_{k,1:s}$. En Mass Transit, il est cependant rare que l'ensemble des trains circule suivant l'ordre prévu par la grille horaire. Pour réduire localement le risque de déviation, les profondeurs des voisinages en gare et en trains sont respectivement restreintes à P et Q. Par ailleurs, l'ordre observé et prévu de chaque arrêt × jour sont systématiquement comparés, s'ils diffèrent, l'arrêt × jour est supprimé. Les arrêts sont observés chaque jour de la semaine du lundi au vendredi, si pour un jour et un arrêt donnés, le voisinage en gares ou en trains n'est pas cohérent avec celui de la grille horaire, nous supprimons l'observation. De fait, ceci arrive rarement, en moyenne 20 % du temps (voir la Section 4.2.2 pour plus de détails), ainsi la suppression de ces observations n'empêche pas d'estimer les paramètres du modèle. Le modèle le plus général de ce chapitre définit un ensemble de régressions linéaires locales en chaque arrêt dont une illustration est donnée sur la Figure 1.8b à droite avec $P = 2$ et $Q = 3$. De façon générique, le modèle linéaire associé à un arrêt (k, s), pour chaque variable d'intérêt $x_{k,s}$, conditionnellement au

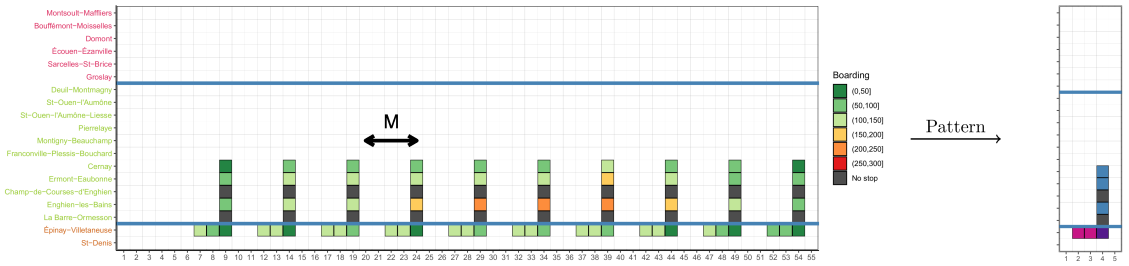


FIGURE 1.9 – Illustration de la répétition d'un voisinage en forme de L sur le graphique gares-trains de l'heure de pointe du matin.

voisinage en forme de L de profondeur P et Q est formulé :

$$x_{k,s} = \beta^{0,0}_{k,s} + \sum_{p=1}^{P} \beta^{p,0}_{k,s} x_{k-p,s} + \sum_{q=1}^{Q} \beta^{0,q}_{k,s} x_{k,s-q} + \varepsilon_{k,s},$$

où $\beta^{0,0}_{k,s}$ est l'ordonnée à l'origine pour chaque arrêt, $\beta^{1:P,0}_{k,s}$ sont les coefficients d'auto-régression avec les trains précédents à la même gare, $\beta^{0,1:Q}_{k,s}$ aux gares précédentes pour le même train et $\varepsilon_{k,s}$ est un bruit indépendant et identiquement distribué. Nous appelons ce modèle, inspiré de Corman and Kecman [2018], *modèle non stationnaire*, puisqu'il ne se factorise pas suivant les gares ou les trains. Un tel modèle est générique pour l'ensemble des variables, et même si cela peut paraitre évident, ses paramètres seront estimés différemment pour chaque variable d'intérêt.

#2 L'introduction de motifs, un compromis entre parcimonie et performance

La seconde idée du Chapitre 4 est l'introduction de la notion de motif, permettant la définition d'un modèle beaucoup plus parcimonieux sans toutefois perdre en performance.

La Figure 1.9 représente une même répétition d'un voisinage en forme de L tous les 5 trains sur la grille horaire. Un motif est donc la répétition régulière d'un même ensemble de dessertes. Les motifs qui sont issus d'horaires cycliques sont là pour aider les voyageurs à mémoriser plus facilement les horaires. Cette répétition à l'intérieur d'une même plage horaire vient s'ajouter aux répétitions entre les jours. Soit M la taille d'un motif. Chaque indice de train k correspond à un indice de course dans le motif, obtenu en calculant la valeur k modulo M, notée $k[M]$. Le nombre d'arrêts (k, s) à modéliser, et donc le nombre de paramètres, diminue d'autant qu'il y a de répétitions du motif contenant k. L'arrêt $(k[M], s)$ est l'arrêt projeté issu de (k, s). Les motifs permettent de définir des modèles linéaires projetés sur un motif dont l'équation pour un arrêt est :

$$x_{k,s} = \beta^{0,0}_{k[M],s} + \sum_{p=1}^{P} \beta^{p,0}_{k[M],s} x_{k-p,s} + \sum_{q=1}^{Q} \beta^{0,q}_{k[M],s} x_{k,s-q} + \varepsilon_{k,s}.$$

Le modèle de prévision basé sur la notion de motif est appelé *modèle stationnaire*. Le modèle stationnaire permet d'explorer des voisinages plus grands avec moins de paramètres pour les variables d'exploitation ferroviaire. L'utilisation des motifs dans le modèle stationnaire impose qu'en heures de pointe les valeurs moyennes et les relations de dépendance entre les arrêts soient identiques quelque soit l'heure. Or, il apparait que le volume de voyageurs varie fortement entre le centre et les extrémités de l'intervalle de temps défini par les heures de pointe. L'idée du *modèle semi-stationnaire* est de permettre que les coefficients associés aux ordonnées à l'origine varient d'un arrêt à l'autre, tout en conservant la stabilité des relations de dépendance entre les arrêts. L'équation pour un arrêt (k, s) du modèle semi-stationnaire où seul $\beta^{0,0}_{k,s}$ peut varier en fonction de l'arrêt est :

$$x_{k,s} = \beta^{0,0}_{k,s} + \sum_{p=1}^{P} \beta^{p,0}_{k[M],s} x_{k-p,s} + \sum_{q=1}^{Q} \beta^{0,q}_{k[M],s} x_{k,s-q} + \varepsilon_{k,s}.$$

La Table 1.2 synthétise (en ligne) les performances des différents modèles de prévision à court terme. Ces modèles sont caractérisés (en colonne) par une profondeur de voisinage en forme de L symétrique, en gare (P) et en train (Q) ainsi que par un nombre de paramètres. Les performances en MAE sont données pour les variables d'exploitation ferroviaire (le temps de stationnement y^{obs} et le retard à l'arrivée Δa) puis pour les variables de flux de voyageurs (le nombre b de montées, le nombre a de descentes et la charge à bord ℓ). Les performances sont calculées à l'horizon d'une gare. Le jeu de données comprend plus de 34 000 arrêts des trains circulant en heures de pointe du matin entre janvier et juillet 2019 sur la ligne H. Il est découpé en un jeu de données d'entrainement (70 %) et de test (30 %). Dans la Table 1.2, les performances du modèle stationnaire sont similaires à celles du modèle non stationnaire pour les variables d'exploitation ferroviaire avec huit fois moins de paramètres. Le modèle semi-stationnaire permet, pour les variables de flux de voyageurs, d'obtenir des performances quasiment équivalentes pour deux fois moins de paramètres. Le nombre de paramètres est un enjeu pour Transilien qui fait rouler plus de 6 200 trains par jour.

#3 Sélection d'un voisinage pour chaque arrêt

Nous proposons dans le Chapitre 4 une stratégie de sélection automatique du meilleur voisinage pour chaque arrêt. Cette sélection évite la comparaison fastidieuse et incomplète des différentes valeurs de P et Q menée dans la Table 1.2. L'enjeu scientifique et industriel pour Transilien est de mieux comprendre pour chaque arrêt (k, s) les phénomènes de propagation des flux de voyageurs ou des retards sur la grille horaire. La sélection automatique de voisinage permet de faire ressortir certains arrêts isolés ou à l'inverse très dépendants des autres arrêts. Cette stratégie permet également de réduire au maximum la complexité des modèles.

Nous avons proposé une représentation graphique, que nous avons appelée *graphe des vents*, dont une instance est représentée sur la Figure 1.10. Elle permet de visualiser sous forme d'une flèche, pour chaque arrêt, la profondeur de voisinages en gares et

TABLE 1.2 – Caractéristiques et performances des modèles de prévision à court terme sur une grille horaire. Pour chaque modèle, le nombre de paramètres (colonne 3) ainsi que le MAE globale des cinq variables à prédire (colonne 4-8) : le temps de stationnement (y^{obs}), le retard à l'arrivée (Δa), le nombre a de descentes et le nombre b de montées, la charge à bord ℓ. Le modèle de référence est en bleu. Les modèles sélectionnés sont en vert.

Modèles			Exploitation ferroviaire		Flux de voyageurs		
Nom	L-forme	Nombre de paramètres	y^{obs} [s]	Δa [s]	a [voy]	b [voy]	ℓ [voy]
Non-stationnaire	P = Q = 0	337	9.7	35.8	10	21	69
	P = Q = 1	956	9.5	16.1	9	18	20
Semi-stationnaire	P = Q = 1	417	9.3	18.6	10	19	23
	P = Q = 2	455	9.2	18.1	9	19	23
	P = Q = 3	482	9.2	18.1	9	18	23
Stationnaire	P = Q = 1	80	9.3	16.2	10	21	27
	P = Q = 2	118	9.2	15.8	8	20	27
	P = Q = 3	145	9.2	15.9	8	20	27

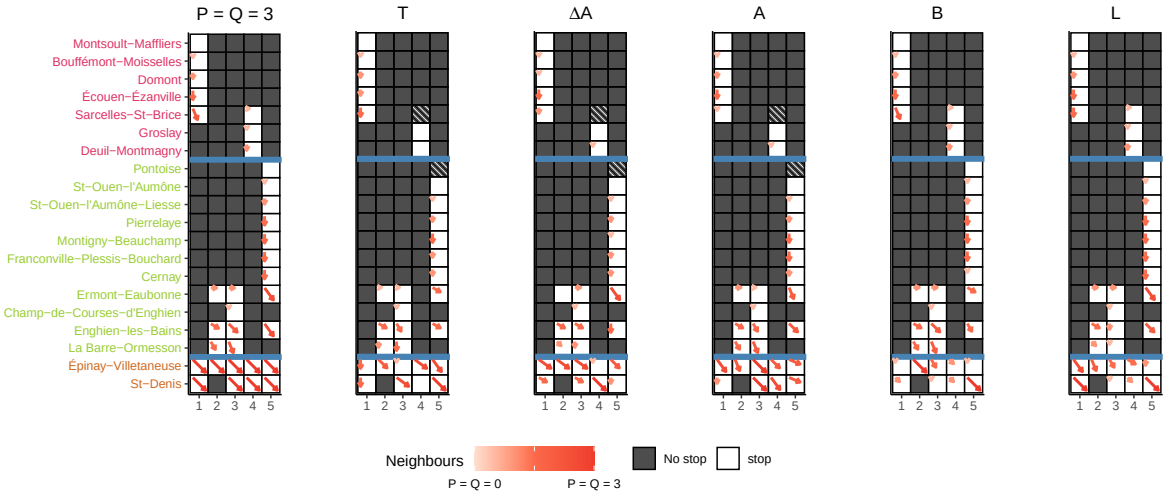


FIGURE 1.10 – Représentation par des flèches du voisinage optimal pour chaque arrêt. Le premier motif à gauche, $P = Q = 3$, est la borne supérieure *i.e.* le plus grand voisinage possible par arrêt. Le voisinage maximal pour un arrêt est atteint s'il est entouré de suffisamment d'arrêts à l'intérieur d'un motif. Chaque motif représente une variable d'intérêt. Pour chaque arrêt, la taille et l'intensité de la couleur des flèches représentent le nombre de voisins sélectionnés. L'orientation des flèches indique la direction privilégiée : un voisinage en gare (flèche horizontale) ou un voisinage en train (flèche verticale). Les cases hachurées sont les gares origines.

en trains sélectionnée par validation croisée avec le critère de MAE, à partir d'une recherche exhaustive de voisinages contiguës de profondeur au plus $P = Q = 3$. La diversité des voisinages optimaux par arrêt pour une même variable confirme qu'il est pertinent de ne pas utiliser systématiquement le voisinage maximal. L'intensité moyenne des voisinages, quel que soit l'arrêt, est plus faible pour le nombre b de montées que pour le retard à l'arrivée Δa ou le nombre a de descentes.

1.3.2 Perspectives

Le Chapitre 4 pose les bases de la prévision à court terme sur une grille horaire. Les perspectives de recherche de ce sujet d'actualité sont nombreuses et s'orientent dans trois directions :

- **Un approfondissement de l'évaluation des performances des modèles** : en élargissant l'horizon de prévision aux gares $s + 2$, $s + 3$, $s + 4$, ... ; en prédisant aussi dans des situations faiblement perturbées ; en étendant les modèles à d'autres lignes.
- **La conception de nouveaux modèles exploitant les voisinages** : en imaginant des modèles non linéaires aux arrêts ; en expérimentant un modèle global de type réseau Bayésien ; en proposant un modèle plus adaptatif pour faire face à des grands événements comme les Jeux Olympiques de 2024.
- **L'élargissement des connaissances** : en injectant les prévisions dans le modèle de temps de stationnement ; en proposant un état de l'art de la prévision des flux de voyageurs à l'échelle du train.

1.4 Modélisation probabiliste des déplacements (Chapitre 5)

Le Chapitre 3 et l'Annexe D confirment que le volume de voyageurs échangé à la porte critique conditionne le temps d'échange minimal. Parallèlement, l'épidémie de COVID-19 a accentué le besoin des voyageurs d'anticiper leur niveau de confort [4] pendant leur voyage. Pour les aider à mieux l'anticiper avant de monter dans le train, nous avons développé le service Hector, détaillé en Section 2.4.3. Hector permet aux voyageurs de la ligne H de consulter *via* un site web l'affluence par zone [5] en temps réel des trois prochains trains, voir la maquette à gauche de la Figure 1.11.

Un des enjeux d'Hector est d'estimer l'affluence à bord par zone à partir du nombre de montées $\boldsymbol{b}_s^{k,d} = \left(\boldsymbol{b}_{s,1}^{k,d}, \ldots, \boldsymbol{b}_{s,I}^{k,d}\right)$ et de descentes $\boldsymbol{a}_s^{k,d} = \left(\boldsymbol{a}_{s,1}^{k,d}, \ldots, \boldsymbol{a}_{s,I}^{k,d}\right)$ par zone, alors que les rames sont communicantes [6], voir la photo à droite de la Figure 1.11.

4. Dans cette thèse, le confort est lié au niveau d'affluence *i.e.* au taux d'occupation.

5. Une zone est une partie de train caractérisée par un nombre de portes ou un nombre de voitures.

6. De façon équivalente traversantes ou BOA (le métro BOA est une rame traversante

FIGURE 1.11 – À gauche, une maquette du site web Hector. À droite, une photo de l'intérieur d'une rame NAT communicante.

Une rame communicante permet aux voyageurs de se déplacer librement entre les différentes zones. Le graphe du bas de la Figure 1.12 donne un exemple de l'ampleur de ces déplacements en comparant le volume de descentes et de montées pour chaque zone à l'échelle d'un trajet. Il est clair qu'il y a en moyenne beaucoup plus de montées que de descentes en zone 16 ou 8.

La problématique scientifique du Chapitre 5 est :

comment estimer les déplacements des voyageurs à bord des rames communicantes à partir du nombre de montées et de descentes par porte ?

Cette problématique est nouvelle en science des transports car habituellement les services d'information des voyageurs d'affluence en temps réel, *real time crowding information* en anglais, exploitent des données de capteurs de masse au niveau des essieux. En 2022, peu d'opérateurs ont des données CAVE/APC accessibles en temps réel et à la maille de la zone. Le Chapitre 5 a pour origine une modélisation naïve des déplacements, présentée dans l'Annexe B, utilisée pour déployer rapidement le site web Hector. Hector est un succès scientifique et industriel car il permettra à Transilien de déployer ce service sur les écrans à quai de la ligne H et de la ligne N d'ici la fin de l'année 2022. C'est une des contributions industrielles majeure de cette thèse. Parallèlement, l'algorithme d'Hector a mûri pour faire émerger un problème mathématique particulièrement intéressant, présenté dans le Chapitre 5 et dont nous donnons ici les principales contributions.

1.4.1 Synthèse des contributions

La modélisation des déplacements des voyageurs dans le Chapitre 5 est abordée à deux échelles, celle du trajet et celle de la gare. La modélisation des déplacements à l'échelle du trajet suppose que les déplacements à l'intérieur du train sont identiques

expérimentale du métro parisien des années 80', l'aspect traversant lui donne l'apparence intérieure d'un long serpent).

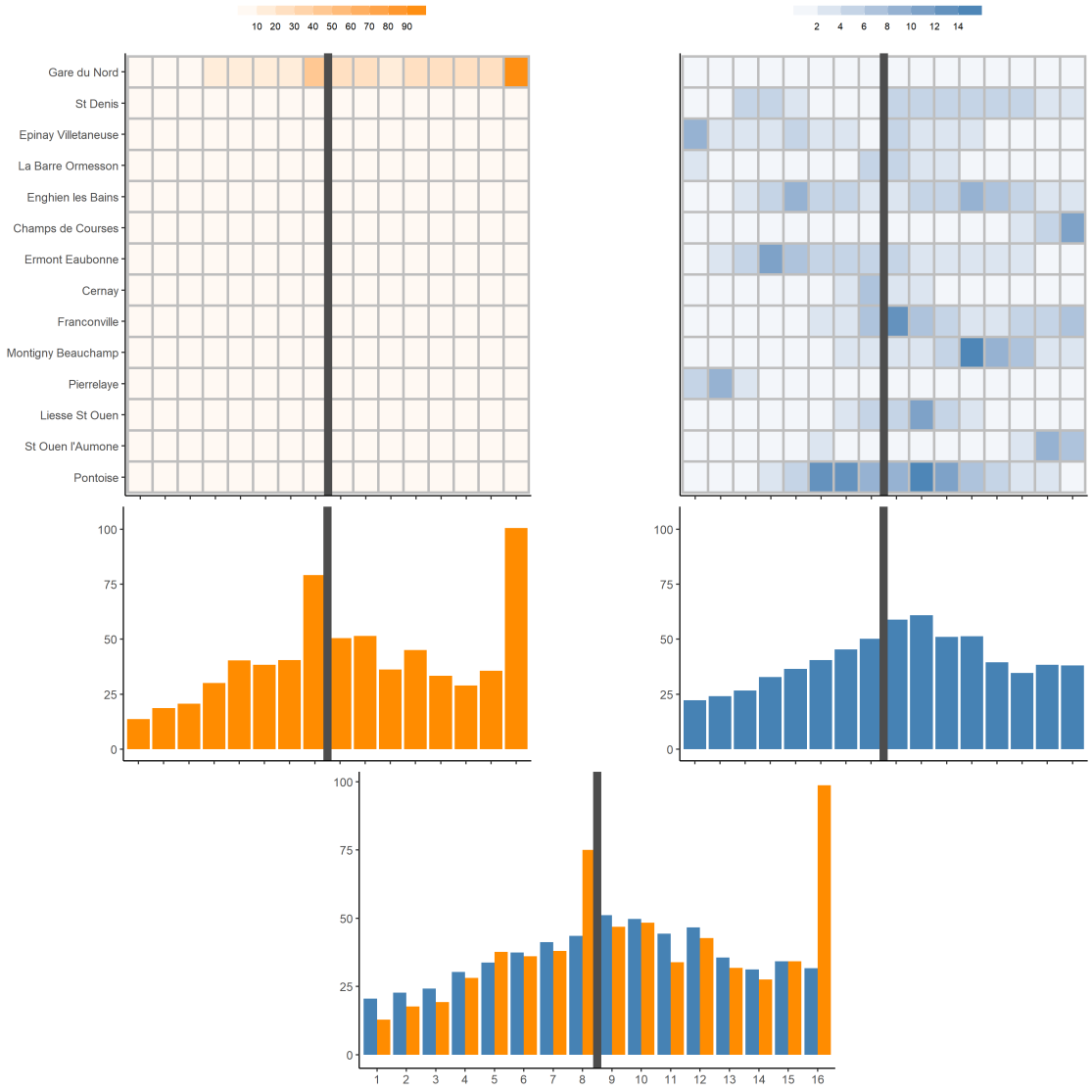


FIGURE 1.12 – Illustration de la somme du vecteur des montées $\boldsymbol{b}_s$ (en orange) et des descentes $\boldsymbol{a}_s$ (en bleu) par gare à l'échelle du trajet. En haut à gauche, les montées représentées par une carte de chaleur, où chaque entrée est la moyenne des montées par gare pour chaque zone, que nous sommons à l'échelle du trajet sur le diagramme en barres juste en-dessous. En haut à droite, les descentes représentées par une carte de chaleur puis sommées en-dessous. Les barres grises verticales séparent la rame avant, numérotée de 1 à 8, et la rame arrière, numérotée de 9 à 16. Le diagramme en barres en bas représente la moyenne des montées et des descentes sommées à l'échelle du trajet pour toutes les zones $(\boldsymbol{b}_\bullet, \boldsymbol{a}_\bullet)$.

pour toutes les gares du trajet. La modélisation à l'échelle de la gare, à l'inverse, permet d'estimer des probabilités de déplacement différentes à chaque gare du trajet. La modélisation des déplacements à l'échelle du trajet consiste à modéliser la somme suivant toutes les gares du trajet du nombre $A_{\bullet,i}$ [7] de descentes par zone en fonction du nombre $b_{\bullet,i}$ de montées. Nous faisons l'hypothèse que les voyageurs descendent depuis la zone où ils se sont déplacés. Cette hypothèse est valable à l'échelle du trajet comme à l'échelle de la gare. Cependant, à l'échelle de la gare, l'ensemble des voyageurs présents dans une zone ne descendent pas, contrairement à l'échelle du trajet, où tous les voyageurs sont descendus à la fin de trajet. Ainsi, à l'échelle du trajet, la distribution des descentes est égale, par hypothèse, à la distribution des déplacements.

La modélisation du nombre de descentes $\boldsymbol{A}_\bullet$ à l'échelle du trajet consiste à faire l'hypothèse que chaque voyageur monté depuis une zone i a une probabilité de descendre (et donc de se déplacer vers les autres zones) issue d'une loi multinomiale $\boldsymbol{A}_{\bullet,i} \sim \mathcal{M}(b_{\bullet,i}, p_{i,1}, \ldots, p_{i,j}, \ldots, p_{i,I})$ où $p_{i,j}$ est la probabilité de se déplacer de la zone i vers la zone j conditionnellement au nombre $b_{\bullet,i}$ de montées en zone i. $\boldsymbol{A}_{\bullet,i}$ est un vecteur aléatoire de taille I, avec I le nombre de zones. La loi jointe du nombre de descentes aux différentes zones est donc une somme de lois multinomiales indépendantes mais non identiquement distribuées :

$$\boldsymbol{A}_\bullet = \sum_{j=1}^{I} \boldsymbol{A}_{\bullet,j}, \tag{1.3}$$

dont nous ne connaissons pas la loi exacte. Nous proposons dans le Chapitre 5 deux stratégies pour contourner ce problème. La première consiste à modéliser l'espérance de la somme de l'équation 1.3, ce qui revient à définir un modèle de régression linéaire multivarié. La seconde consiste à proposer une approximation de la loi de $\boldsymbol{A}_\bullet$ puis de développer une approche par maximum de vraisemblance. Nous comparons les deux modèles à l'échelle du trajet sur un unique jeu de données réelles. La dernière contribution du Chapitre 5 est l'extension du modèle précédent à l'échelle de la gare.

#1 Un modèle linéaire multivarié sous contraintes à l'échelle du trajet

Le modèle en espérance à l'échelle du trajet consiste à calculer l'espérance du vecteur aléatoire défini dans l'équation (1.3) conditionnellement au nombre de montées :

$$\mathbb{E}[\boldsymbol{A}_\bullet|\boldsymbol{b}_\bullet] = \sum_{i=1}^{I} \mathbb{E}\big[\boldsymbol{A}_{\bullet,j}|\boldsymbol{b}_\bullet\big] = \sum_{i=1}^{I} b_{\bullet,i}\boldsymbol{p}_i.$$

7. Les majuscules sont réservées aux variables aléatoires.

où $\boldsymbol{p}_i = (p_{i,1}, \ldots, p_{i,j}, \ldots, p_{i,I})$. Les probabilités de déplacement sont définies par la matrice stochastique $\boldsymbol{P}$ décrite dans l'équation 1.4 :

$$\boldsymbol{P} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,I} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,I} \\ \vdots & \vdots & \ddots & \vdots \\ p_{I,1} & p_{I,2} & \cdots & p_{I,I} \end{pmatrix}. \tag{1.4}$$

On rappelle qu'une matrice stochastique vérifie que la somme des probabilités en ligne vaut un et que chaque entrée $p_{i,j}$ appartient à l'intervalle $[0, 1]$. L'estimation des descentes $\boldsymbol{A}_\bullet$ peut être vue comme un problème de régression d'une variable à expliquer multivariée $\boldsymbol{A}_\bullet$ (le vecteur du nombre de descentes par zone à l'échelle du trajet) en fonction de variables explicatives : le nombre $\boldsymbol{b}_\bullet$ de montées. Les paramètres du modèle sont les probabilités de déplacement $\boldsymbol{P}$. Ce problème est un problème des moindres carrés classique :

$$\boldsymbol{A}_\bullet = \boldsymbol{b}_\bullet \boldsymbol{P} + \text{ bruit.} \tag{1.5}$$

Les observations, pour chaque trajet k, un jour donné d, permettent de proposer un modèle. Ainsi, en ajoutant les contraintes sur les paramètres, on obtient la matrice estimée $\widehat{P}$ de P qui est solution du problème d'optimisation sous contraintes suivant :

$$\begin{aligned} \underset{\boldsymbol{P}}{\text{argmin}} \quad & \sum_{(k,d)\in\mathcal{N}} \left\| \boldsymbol{a}_\bullet^{k,d} - \boldsymbol{b}_\bullet^{k,d} \boldsymbol{P} \right\|_2^2 \\ \text{s.c} \quad & \boldsymbol{P} \text{ est stochastique.} \end{aligned} \tag{1.6}$$

Le problème (1.6) est un problème simple d'optimisation quadratique sous contraintes d'égalités et d'inégalités linéaires. Ce modèle formulé par régression linéaire multivarié a l'avantage d'être simple mais il ne s'étend pas facilement à l'échelle de la gare.

#2 Une modélisation probabiliste des déplacements à l'échelle du trajet

Dans le Chapitre 5, nous abordons le problème de la modélisation des déplacements à l'échelle du trajet de façon probabiliste. Le vecteur du nombre de descentes à l'échelle du trajet $\boldsymbol{A}_\bullet$ est équivalent par hypothèse au nombre de déplacements jusqu'à chaque zone. Il s'agit de poser un modèle probabiliste sur le vecteur aléatoire du nombre de descentes, conditionnellement au nombre de montées par zone. Nous avons exprimé dans l'équation (1.3), le vecteur des descentes $\boldsymbol{A}_\bullet$ comme la somme de lois multinomiales indépendantes mais non identiquement distribuées, où chacune des lois impliquées modélise les déplacements des voyageurs montés par une des zones. La loi de cette somme n'étant pas connue, nous en proposons une approximation 1 :

Approximation 1 *La loi de $\boldsymbol{A}_\bullet$ est approchée par la loi multinomiale de même espérance : $\mathcal{M}(b_{\bullet,\bullet}, \pi_{\bullet,1}, \ldots, \pi_{\bullet,I})$ avec $\pi_{\bullet,j} = \sum_{i=1}^{I} r_{\bullet,i} p_{i,j}$ où $r_{\bullet,i} = b_{\bullet,i}/b_{\bullet,\bullet}$.*

Nous discutons la qualité de cette approximation dans l'Annexe 5.A. L'approximation 1 permet d'exprimer la log-vraisemblance associée à une observation (k, d) ne dépendant que des observations et des paramètres :

$$\ell(\boldsymbol{a}_{\bullet}^{k,d}; \boldsymbol{P}, \boldsymbol{b}_{\bullet}^{k,d}) = \sum_{j=1}^{I} a_{\bullet,j}^{k,d} \log \left(\sum_{i=1}^{I} r_{\bullet,i}^{k,d} p_{i,j} \right).$$

La log-vraisemblance sous contraintes d'inégalités et d'égalités linéaires du problème d'optimisation convexe (1.7) est optimisée numériquement à l'aide d'une librairie standard :

$$\begin{aligned} \underset{\boldsymbol{P}}{\text{argmax}} \quad & \sum_{(k,d)\in\mathcal{N}} \sum_{j=1}^{I} a_{\bullet,j}^{k,d} \log \left(\sum_{i=1}^{I} r_{\bullet,i}^{k,d} p_{i,j} \right) \\ \text{s.c} \quad & \boldsymbol{P} \text{ est stochastique.} \end{aligned} \tag{1.7}$$

L'estimation du nombre de descentes est $\widehat{\boldsymbol{a}}_{\bullet} = \boldsymbol{b}_{\bullet}\widehat{P}$.

#3 Deux modèles cohérents à l'échelle du trajet

La Table 1.3 compare la qualité d'estimation du vecteur du nombre de descentes à l'échelle du trajet par rapport au vecteur du nombre de descentes estimées par zone. Les erreurs MAE des deux modèles de régression et probabiliste (en ligne) ont été comparés pour la rame avant et la rame arrière (en colonne) à une stratégie d'estimation sans déplacement. Le jeu de données considéré comprend plus de 3 500 trajets de Gare du Nord vers Pontoise sur la ligne H entre janvier et septembre 2021. Les performances en MAE ont été évaluées en découpant le jeu de données en un jeu de données d'entrainement (75 %) et de test (25 %). L'erreur d'estimation calculée sur le jeu de données test est la suivante :

$$\text{MAE} = \frac{1}{N_{\mathcal{T}\text{est}}} \sum_{(k,d)\in\mathcal{T}\text{est}} \left\| \boldsymbol{a}_{\bullet}^{k,d} - \widehat{\boldsymbol{a}}_{\bullet}^{k,d} \right\|_1,$$

où $\widehat{\boldsymbol{a}}_{\bullet}^{k,d} = \boldsymbol{b}_{\bullet}^{k,d}\widehat{\boldsymbol{P}}$ est l'estimation du nombre de descentes conditionnellement aux montées. Les deux modèles ont des performances similaires pour la rame avant et arrière. Il apparait cependant dans le Chapitre 5 que les matrices de passage estimées $\widehat{\boldsymbol{P}}$ différent légèrement entre les deux modèles. Les modèles de régression et probabiliste commettent une erreur moyenne d'estimation du nombre de descentes deux fois moins importante qu'un modèle sans déplacement.

La modélisation des déplacements à l'échelle du trajet permet d'estimer correctement le nombre de descentes. Cependant, une telle modélisation fait l'hypothèse que les déplacements sont identiques d'une gare à l'autre. Cette hypothèse est discutable pour deux raisons : la première parce que la géographie des quais influence le positionnement à quai des voyageurs et donc leurs déplacements à bord ; la seconde est qu'une telle modélisation revient à perdre de l'information en sommant les montées et les descentes à l'échelle du trajet.

TABLE 1.3 – Performances d'estimation du vecteur du nombre de descentes à l'aide du vecteur du nombre de montées à l'échelle du trajet. La différence est calculée selon le critère de MAE en fonction du modèle de déplacement des montées utilisé pour les rames avant et arrière. En ligne, les différents modèles estimés : sans déplacement (matrice identité, $\boldsymbol{P} = I_I$), régression ($\widehat{\boldsymbol{P}}_{\text{MLS}}$) ou probabiliste ($\widehat{\boldsymbol{P}}_{\text{MLE}}$).

	Avant	Arrière
Modèles	MAE [voy]	MAE [voy]
Sans déplacement	10.9	17.5
$\widehat{\boldsymbol{P}}_{\text{MLS}}$	6	8.5
$\widehat{\boldsymbol{P}}_{\text{MLE}}$	6	8.5

#4 Un modèle probabiliste avec variables cachées à l'échelle de la gare

Le modèle probabiliste à l'échelle de la gare a été formalisé et une stratégie d'estimation a été identifiée. Cependant, nous n'avons pas pu le tester sur des données réelles dans le cadre de cette thèse ; cette confrontation à la réalité est un projet de l'immédiat après-thèse. Pour concevoir le modèle de déplacements à l'échelle de la gare, nous ne pouvons plus utiliser l'hypothèse d'identification des déplacements au nombre de descentes. En effet, à chaque arrêt, seule une partie des voyageurs descendent, entrainant une accumulation des déplacements non révélés (cachés) au fil du trajet. Certes, le nombre de descentes $\boldsymbol{A}_s$ observées à une gare ne révèle pas complètement les déplacements mais il donne des informations sur le nombre de voyageurs présents dans chaque zone. Ainsi, pour relier le nombre de descentes et le nombre de déplacements par zone, nous définissons une loi d'émission des descentes qui dépend de la charge à bord cachée $\boldsymbol{L}_{s-1}$. Pour rappel, la charge à bord à l'arrivée de la gare s' en zone i est égale à la somme cumulée le long du trajet de la différence des déplacements jusqu'à cette zone et du nombre de descentes :

$$L_{(s-1),i} = \sum_{g=1}^{s-1} W_{g,i} - A_{g,i}.$$

Nous avons posé une hypothèse importante sur la loi du nombre de descentes, à savoir que le nombre de descentes en zone i à la gare s conditionnellement au passé ne dépend que de la charge à bord en sortie de gare $s-1$. Le nombre de descentes en zone i à la gare s conditionnellement à la charge à bord $\ell^{k,d}_{s-1,i}$ en sortie de gare $s-1$ suit une loi binomiale :

$$A^{k,d}_{s,i} \sim \mathcal{B}(\ell^{k,d}_{s-1,i}, \alpha_{s,i}), \quad s = 2, \ldots, S-1.$$

La modélisation probabiliste des déplacements à l'échelle de la gare nécessite la modélisation conjointe de la loi du nombre de descentes (loi d'émission) et celle du nombre de déplacements cachés à chaque gare. La loi des déplacements à l'échelle de la gare est issue de la même approximation que celle à l'échelle de la gare. Les déplacements conditionnellement aux montées à la gare suivent la loi multinomiale :

$$\boldsymbol{W}^{k,d}_s \sim \mathcal{M}(b^{k,d}_{s,\bullet}, \pi^{k,d}_{s,1}, \ldots, \pi^{k,d}_{s,I}), \quad s = 1, \ldots, S-1,$$

où $\pi_{s,j}^{k,d} = \sum_{i=1}^{I} r_{s,i}^{k,d} p_{s,i,j}$ avec $r_{s,i}^{k,d} = b_{s,i}^{k,d} / b_{s,\bullet}^{k,d}$. Un ensemble d'hypothèses supplémentaires justifiées dans le Chapitre 5 permettent d'exprimer la loi jointe des déplacements et des descentes à partir de celles du nombre de déplacements, de loi multinomiale, et du nombre de descentes, de loi binomiale :

$$\begin{aligned}
&\mathbb{P}\Big(\boldsymbol{a}_{2:S}, \boldsymbol{w}_{1:(S-1)}; \boldsymbol{b}_{1:(S-1)}, \boldsymbol{\theta}\Big) \\
&= \prod_{s=2}^{S} \underbrace{\left(\prod_{i=1}^{I} \binom{\ell_{s-1,i}}{a_{s,i}} (\alpha_{s,i})^{a_{s,i}} (1-\alpha_{s,i})^{(\ell_{s-1,i}-a_{s,i})} \right)}_{\mathbb{P}(\boldsymbol{a}_s | \boldsymbol{\ell}_{s-1}; \boldsymbol{\theta})} \underbrace{\left(\prod_{i=1}^{I} \frac{(b_{s-1,\bullet}!)}{(w_{s-1,i}!)} (\pi_{s-1,i})^{w_{s-1,i}} \right)}_{\mathbb{P}(\boldsymbol{w}_{s-1}; \boldsymbol{b}_{s-1}, \boldsymbol{\theta})},
\end{aligned} \tag{1.8}$$

où $\boldsymbol{\theta} = (\boldsymbol{P}_{1:(S-1)}, \boldsymbol{\alpha}_{2:S})$ est l'ensemble des paramètres des modèles probabilistes. La vraisemblance observée associée à la loi jointe de l'équation (1.8) n'est pas calculable à cause du nombre exponentiel d'états cachés ℓ et w. Une solution classiquement utilisée pour optimiser la vraisemblance dans le cas de variables cachées est l'algorithme EM. Un travail de l'immédiat après-thèse est d'écrire et d'implémenter l'algorithme EM pour le tester sur des données réelles.

1.4.2 Perspectives

L'objectif à très court terme est d'implémenter et de tester le modèle proposé à l'échelle de la gare. À court terme, il s'agit également d'implémenter la stratégie de déplacements des voyageurs afin de fiabiliser le service industriel d'information des voyageurs à la zone qui sera déployé à la fin de l'année 2022. L'approximation 1 a été appuyée par simulation dans l'Annexe 5.A, toutefois le calcul de bornes théoriques pour valider cette approximation serait intéressant. Le modèle avec variables cachées permet d'estimer des probabilités de déplacements différentes pour chaque gare. Un objectif à plus long terme serait de pouvoir les comparer et les classer automatiquement.

1.5 Structure de la thèse et publications associées

Cette thèse s'articule autour de quatre grands blocs représentés sur la Figure 1.13 qui traitent de la modélisation et la prévision des variables d'exploitation ferroviaire et des flux de voyageurs en zone dense.

Ancrage industriel Le Chapitre 2 pose les bases industrielles et scientifiques nécessaires à la compréhension des enjeux de la thèse. Il est écrit pour des lectrices et des lecteurs, qui comme nous il y a trois ans, sont étrangers au monde des transports en commun. Il présente les étapes clés de la planification ferroviaire ainsi que les connaissances nécessaires à la compréhension de la demande. Le Chapitre 2 est également l'occasion de présenter les contributions industrielles ainsi que d'inscrire

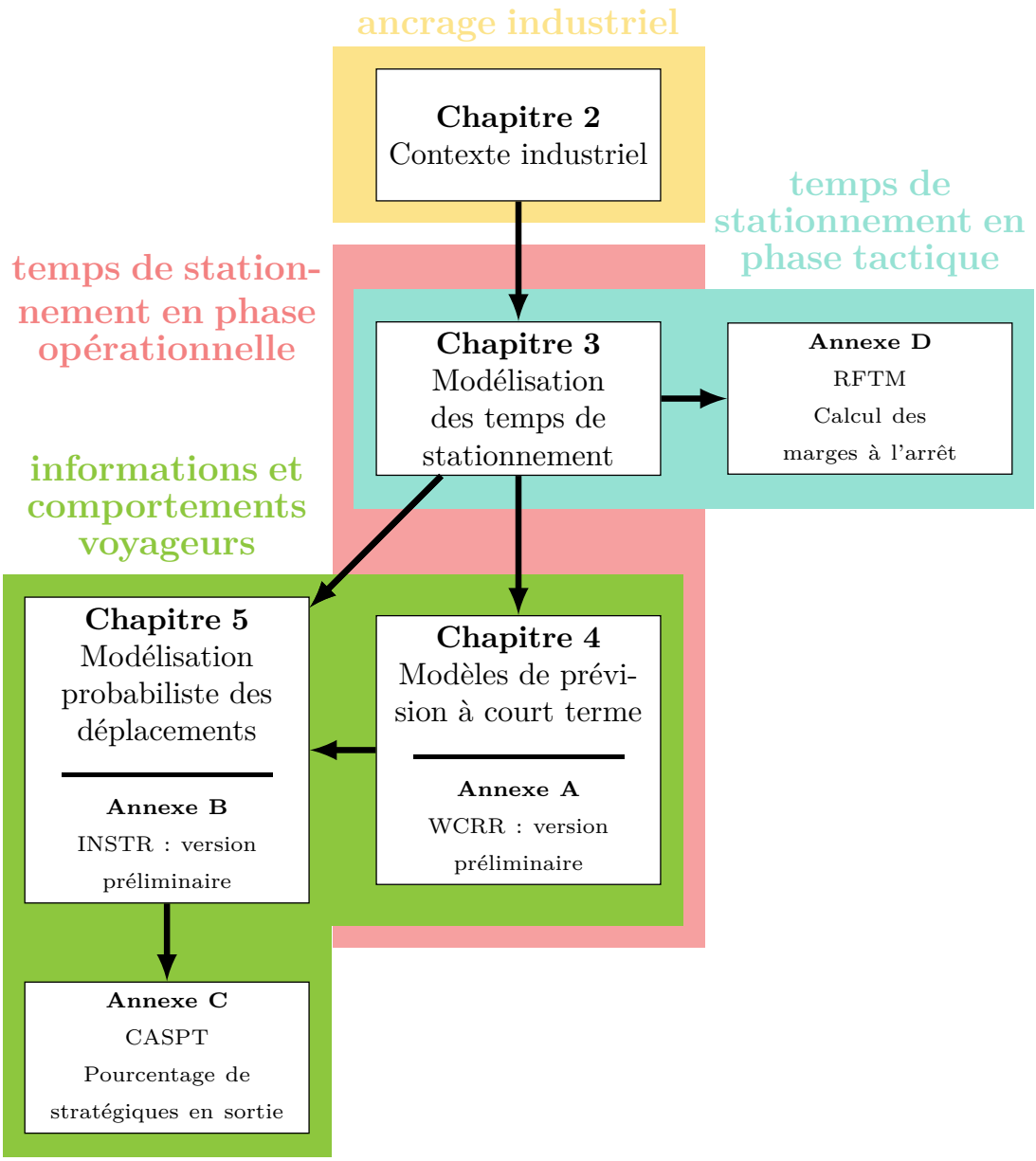


FIGURE 1.13 – Plan schématique des chapitres de la thèse.

la thèse dans un champ de recherche et une stratégie d'entreprise qui restent encore largement à construire.

Temps de stationnement en phase opérationnelle Les Chapitres 3–4 et l'Annexe A sont complémentaires. Le Chapitre 3 propose un modèle statistique des temps de stationnement qui permet d'identifier les principaux déterminants du temps de stationnement : importance des heures de départ théoriques et des flux de voyageurs pour les trains en retard. Pour rendre ce modèle opérationnel, l'Annexe A et surtout le Chapitre 4 sont l'occasion de développer une stratégie de prévision à court terme (à l'horizon d'une gare) des variables explicatives du modèle du Chapitre 3. Ces deux chapitres traitent de la problématique des temps de stationnement en opérationnel ainsi que de la question plus générale de la prévision à court terme sur un réseau de trains de banlieue.

Informations et comportements voyageurs Les Chapitres 4 et 5 ainsi que les Annexes B et C traitent de la qualité de l'information des voyageurs ainsi que de la modélisation de leurs comportements. Le Chapitre 4 propose une stratégie de prévision à court terme des flux de voyageurs et des retards pouvant alimenter les canaux d'information des voyageurs. La modélisation des déplacements, qui permet de calculer une affluence fiable par zone, a été amorcée dans l'Annexe B et a largement été développée dans le Chapitre 5. Le travail en Annexe C est une ouverture pour mieux comprendre les comportements des voyageurs à l'interface quai-train. L'ambition de ces travaux est d'embrasser la question de la compréhension de l'interaction des voyageurs avec le train, au moment de monter ou de descendre du train, ainsi qu'au cours de leur voyage.

Temps de stationnement en phase tactique Ce bloc comprend un résumé de conférence D et le Chapitre 3. Il traite de la problématique, esquissée dans cette thèse, de la prise en compte des flux de voyageurs dans la conception de grille horaire, et en particulier pour dimensionner les temps de stationnement théoriques. C'est un sujet clé pour permettre aux opérateurs comme Transilien de gagner en rigueur et en efficacité.

Publications associées aux travaux de doctorat

Articles de journaux soumis ou en cours de rédaction

Le Chapitre 3 est une version de l'article :

Coulaud, Rémi, Keribin, Christine, et Stoltz, Gilles. Modeling dwell time in a data-rich railway environment: with operations and passenger flows data. Re-soumis à *Transportation Research Part C* (TRC) après corrections. Preprint accessible ici `hal.archives-ouvertes.fr/hal-03651835/`, 2022

Le Chapitre 5 donnera lieu à la rédaction d'un article.

Conférences internationales

L'article suivant est en Annexe C :

Coulaud, Rémi, Mazon, Valentine, Sanchis, Laura, et Cats, Oded. Share of strategic alighting passengers combining automatic passenger counting and OpenStreeMap. In *Conference on Advanced Systems in Public Transport (CASPT)*, 2022

Les deux articles suivants sont respectivement des versions préliminaires des Chapitre 4 et Chapitre 5. Ils sont respectivement en Annexe A et en Annexe B.

Coulaud, Rémi, Keribin, Christine, et Stoltz, Gilles. One-station-ahead forecasting of dwell time, arrival delay and passenger flows on trains equipped with automatic passenger counting (apc) device. In *13th World Congress on Rail Research (WCRR)*, 2022

Coulaud, Rémi et Vimont, Mathilde. How to use APC data to model passenger movement on-board? An application to Paris suburban train network. In *8th International Symposium On Transport Network Reliability (INSTR)*, 2021

Conférence nationale

L'article suivant est en Annexe D :

Coulaud, Rémi et Grangé, Martine. Modélisation de l'impact des flux voyageurs sur les temps d'échange pour la simulation des marges d'exploitation : une application à la ligne N de transilien. In *4èmes Rencontres Francophones Transport Mobilité (RFTM)*, 2022

2

Contexte industriel

Ce chapitre pose les bases industrielles et scientifiques nécessaires à la compréhension des enjeux de la thèse. Il est écrit pour les lectrices et les lecteurs qui, comme nous il y a trois ans, sont étrangers au monde des transports en commun. Il présente les étapes clés de la planification ferroviaire ainsi que les connaissances nécessaires à la compréhension de la demande. Ce chapitre est également l'occasion de présenter les contributions industrielles ainsi que d'inscrire la thèse dans un champ de recherche et une stratégie d'entreprise qui restent encore largement à construire.

Contents

2.1 Introduction

Les transports en commun [1] sont l'épine dorsale des métropoles. Ils permettent le déplacement d'un grand volume de personnes grâce à des horaires et des trajets fixes, tout en ne consommant que peu d'énergie et d'espace. Les transports en commun font ainsi partie de la solution pour accélérer la transition énergétique [Farandou, 2022]. La fréquentation des transports en commun en Île-de-France a connu une forte croissance entre 2001 et 2018 : le nombre de déplacements journaliers est passé de 7 à 9,4 millions selon la dernière enquête globale transport [OMNIL, 2018].

L'épidémie de COVID-19 a cassé cette dynamique mais au vu des perspectives de croissance démographique de l'Île-de-France ainsi que l'ouverture du Grand Paris Express, le besoin de transporter plus de voyageurs sur le réseau historique reste d'actualité. La bonne adéquation entre l'offre et la demande est une nécessité pour que l'augmentation du nombre de voyageurs ne soit pas synonyme de dégradation du service. L'inadéquation de l'offre à la demande se traduit souvent par de la congestion. Cette congestion peut causer des retards, par exemple à Transilien, un chiffre est partagé selon lequel 1 pour cent de croissance de la fréquentation d'une ligne entraine une hausse de 0,5 pour cent des retards. L'objectif de cette introduction est de poser, dans un premier temps, le contexte et les enjeux des transports en commun en Île-de-France pour ensuite, nous focaliser sur le périmètre de cette thèse, le réseau de trains de banlieue opéré par Transilien.

2.1.1 Fonctionnement des transports en commun en Île-de-France

Une des richesses de l'Île-de-France est son réseau de transports en commun que nous esquissons dans cette partie. Il s'agit également de présenter le rôle et les missions de l'autorité organisatrice des transports Île-de-France Mobilités (IdFM). Nous concluons cette partie en donnant un aperçu des enjeux associés à la mise en concurrence des transports en commun d'Île-de-France.

L'Île-de-France : un maillage de transports en commun dense

En Île-de-France, 43 millions de déplacements sont réalisés chaque jour par des voyageurs se déplaçant en moyenne 1h30 et 18 km selon la dernière Enquête Globale Transport OMNIL [2018]. Ces déplacements se font pour 22 % en transports en commun. Parmi les déplacements en transports en commun, 70 % sont sur un réseau ferré, le reste sur les réseaux de tramway ou de bus. Les déplacements sur un réseau ferré sont réalisés pour un peu plus de la moitié sur un réseau de type métro et le reste (3,4 millions) sur le réseau de trains de banlieue. Pour absorber ce volume de déplacements, l'Île-de-France est sillonnée par un réseau dense comprenant 13 lignes

1. Public transport ou mass transit en anglais

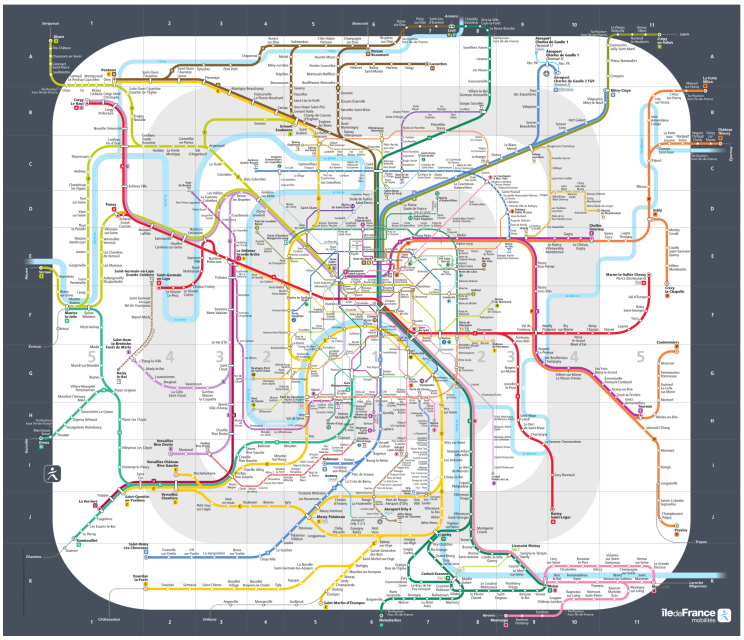


FIGURE 2.1 – Plan du réseau de transports en commun d'Île-de-France comprenant le réseau Transilien.

de trains régionaux dont 5 lignes de réseau express régional (RER), 14 lignes de métro, 9 lignes de tramway et plus de 1 500 lignes de bus représentées en partie sur la Figure 2.1.

Deux grandes entreprises de transports en commun se partagent le réseau ferré d'Île-de-France. La société nationale des chemins de fer français (SNCF), au travers de son activité Transilien, exploite le réseau de trains de banlieue d'Île-de-France composé des lignes nommées par des lettres sur la Figure 2.1. Les gares au sud de Gare du Nord sur la ligne B et les gares à l'est de la Défense sur la ligne A sont exploitées par la RATP. La régie autonome des transports parisiens (RATP) exploite les lignes de métro. La RATP et la SNCF exploitent chacune une partie du réseau de tramway. Les bus sont exploités par Keolis, Transdev, RATP Dev ou d'autres opérateurs indépendants qui sont regroupés dans l'Organisation professionnelle des transports d'Île-de-France (Optile). La densité et le nombre de voyageurs par an, font de Transilien le deuxième opérateur de trains de banlieue au monde [Laurent et al., 2018]. Les trains de banlieue jouent un rôle structurant dans la desserte des territoires d'Île-de-France, qu'ils soient faiblement ou densément peuplé.

Le fonctionnement d'Île-de-France Mobilités (IdFM).

Les transports en commun, et en particulier le réseau de trains de banlieue, nécessitent des investissements colossaux. En Île-de-France, ces investissements sont assurés par l'autorité organisatrice des transports (AOT) Île-de-France Mobilité (IdFM) dont les trois grandes missions sont :

1. imaginer le réseau de demain : les nouvelles lignes, les extensions de lignes,

etc. ;

2. organiser le réseau existant : renouveler les rames ou les bus, financer les travaux de régénération du réseau, etc. ;
3. financer l'exploitation et définir les trames de desserte : commander aux opérateurs une offre de transport qu'ils devront exécuter.

Pour mener à bien ces missions, les revenus d'IdFM proviennent pour 25 % de la vente des titres de transport, 50 % des taxes aux entreprises et 25 % des subventions publiques. Les deux premières missions d'IdFM concernent ce que nous appelons par la suite la phase stratégique de la planification ferroviaire, voir Figure 2.4. Les étapes de cette phase sont considérées comme des données d'entrée de cette thèse. La troisième mission de planification et de financement des transports en commun d'IdFM est quant à elle au cœur de nos enjeux. IdFM définie tous les 4-9 ans au travers d'un appel d'offre ou d'un contrat [IdFM, 2020], une offre de référence, aussi appelée manchette pour les trains de banlieue. Dans cette offre de référence, IdFM spécifie les horaires des trains pour une année. Les trains de banlieue ont des horaires fixes qui se répètent suivant différents types de jours : les jours ouvrés de bases (JOB) du lundi au vendredi [2], les samedis, les dimanches et jours fériés. Il y a des adaptations pour les vacances d'été (six semaines) et d'hiver (une semaine). Pour les trains de banlieue l'offre d'IdFM spécifie en plus le numéro du train, le code de mission, le nombre et type de rames avec leur capacité totale (places assises + places debout [3]), voir le contrat IdFM [2020].

Appels d'offre et mise en concurrence.

La mise en concurrence des transports en commun consiste en l'ouverture par un appel d'offre de la gestion d'une ou plusieurs lignes de transport. Il est à prévoir que le rôle d'IdFM dans la planification et/ou l'exploitation des transports en commun se renforcera dans les prochaines années. La mise en concurrence qui s'échelonne de 2021 à 2040 concerne tous les modes de transports en commun (trains de banlieue, métros, tramways, bus). Les premières lignes à être concernées par l'ouverture à la concurrence sont les lignes de bus et de tramways, suivies des lignes de trains de banlieue, pour finir par les lignes de métro et RER.

L'ouverture à la concurrence amène les opérateurs historiques à se questionner sur leurs modèles d'exploitation et de services. Les opérateurs historiques ont besoin de se différencier des concurrents pour montrer leur plus-value au regard des standards d'exploitation internationaux. Transilien cherche à renforcer son expertise dans l'exploitation et la planification des trains en *Mass Transit* pour se distinguer de ses concurrents. Par ailleurs, les nombreux projets de RER dans d'autres métropoles françaises (Strasbourg, Lyon, Bordeaux, Toulouse, etc.) sont une opportunité pour Transilien et le groupe SNCF, dont Keolis fait partie, de valoriser leur expertise dans l'exploitation ferrée en zone dense. Cette thèse qui

2. Pour certains départements à Transilien, les JOB désignent exclusivement les mardis et jeudis.
3. Une densité théorique de 4 personnes par mètre carré

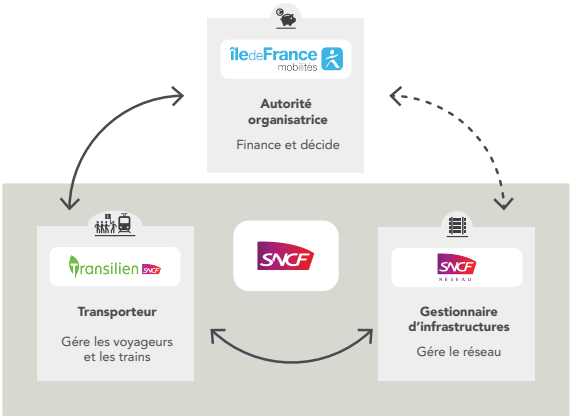


FIGURE 2.2 – Schéma illustrant les différents acteurs de l'exploitation des trains de banlieue en Île-de-France.

porte sur l'adéquation de l'offre et de la demande en zone dense, développe des idées particulièrement importantes pour Transilien dans ce contexte.

2.1.2 Transilien et le Mass Transit ouvert

Transilien opère des trains de banlieue qui circulent sur le réseau ferré national (RFN) géré par SNCF-Réseau. SNCF-Réseau est le gestionnaire d'infrastructure [4] du réseau ferré en France. La construction des horaires en Île-de-France est le résultat d'une étroite collaboration entre l'autorité organisatrice (IdFM), l'opérateur (Transilien) et le gestionnaire d'infrastructure (SNCF-Réseau), voir figure 2.2. Ainsi, il est écrit dans le dernier contrat 2020-2023 IdFM [2020] : *« [Transilien] sera force de proposition tant en ce qui concerne l'évolution de l'offre, l'amélioration de la qualité du service, la tarification, que la modernisation des réseaux, leur interopérabilité, et la coordination avec les autres opérateurs »*. Ainsi, la modification des horaires en amont ne peut se faire qu'en concertation avec l'autorité organisatrice et le gestionnaire d'infrastructure. L'offre de référence étant contractuelle, tout écart à cette offre entraine des pénalités pour l'opérateur.

Milieu ouvert

Une des contraintes fortes qui distingue Transilien des autres modes de transports urbains est le fait que les trains circulent sur le réseau ferré national (RFN). Le RFN est un milieu ouvert au sens où il est partagé avec d'autres opérateurs ferroviaires en particulier TER, TGV et SNCF-Fret. Les métros ou les tramways circulent eux en milieu fermé avec des voies dédiées ce qui leur donne une grande latitude pour la planification et la gestion opérationnelle des circulations [5]. Par

4. L'infrastructure est l'ensemble des installations fixes telle que : la voie ferrée, les caténaires, le système de signalisation, les gares, etc.

5. Les circulations définissent les déplacements des trains sur le réseau ferré.

exemple, il est relativement simple de modifier un système de signalisation pour augmenter la capacité [6] d'une ligne quand la même opération sur le RFN se révèle très compliquée compte tenu des différents opérateurs impactés. De même, la gestion opérationnelle des circulations pour un réseau de métro se fait de façon efficace car l'opérateur et le gestionnaire d'infrastructure sont une unique entité dirigée vers un objectif commun.

Pour Transilien, la circulation en milieu ouvert nécessite une coopération constante avec le gestionnaire d'infrastructure SNCF-Réseau. Ce dernier peut avoir des vues divergentes avec Transilien :

- En amont, au moment de la commande des sillons [7] où il faut arbitrer entre les sillons commandés et ceux alloués pour les travaux ou les autres opérateurs ;
- En temps réel, au moment de la gestion opérationnelle des circulations, où le centre opérationnel de gestion des circulations (COGC), la tour de contrôle du réseau ferré, doit arbitrer entre les différents opérateurs et ne met pas toujours Transilien en priorité.

En milieu ouvert, les horaires théoriques sont très contraignants. Le partage des voies nécessite l'adaptation à un système de signalisation générique, appelé block automatique lumineux (BAL), dont nous donnons les grands principes dans l'encadré ci-dessous. Le cantonnement impose aux opérateurs d'espacer les trains d'au moins deux cantons. Or sur certaines portions du réseau, un espacement de deux cantons ($\approx$ 2 min) est déjà trop important pour répondre à la demande. La contrainte d'un milieu ouvert vient aussi de la nécessité de s'adapter à tous les matériels roulants, par exemple la longueur des cantons et les limites de vitesse sont fixées en fonction du matériel avec la distance de freinage la plus importante. Le réseau et les performances des trains imposant une capacité maximale, il serait intéressant de réfléchir aux solutions pour réduire cet espacement minimal. Ici, nous nous intéressons aux marges de manœuvre disponibles lors des arrêts en gare pour embarquer plus de voyageurs ou raccourcir les temps de stationnement. Notre objectif est de mieux prendre en compte les flux de voyageurs dans la planification et la gestion opérationnelle des circulations.

6. La capacité est le nombre maximum de trains qu'il est possible de faire circuler sur une portion de voie et une plage de temps donnés selon Hansen and Pachl [2014].

7. Un sillon est la capacité d'infrastructure dans l'espace-temps (réseau) requise pour faire circuler un train donné entre deux points sur un réseau ferré pendant une période de temps donnée.

Le poids, la vitesse et le contact fer-fer (roue-rail) accroissent les distances de freinage par rapport à la route. Ces distances de freinage empêchent un conducteur de conduire à vue. Le système de cantonnement ou block automatique lumineux (BAL) permet de remédier à ce problème en utilisant la voie et un système de feux tricolores pour indiquer la présence au conducteur d'un train dans les prochains kilomètres.
Le cantonnement fixe fonctionne à partir de cantons (block). Un canton est délimité par un feu tricolore qui indique l'état du canton suivant. Un feu rouge indique que le canton est occupé : un nouveau train ne doit pas entrer dans ce canton. Un feu orange indique que le canton est libre mais que le suivant est occupé. Un train peut entrer dans un canton avec un feu orange mais il doit freiner jusqu'à atteindre 30 km/h pour pouvoir conduire à vue. Un feu vert indique que la voie est libre et que le train peut entrer sans restriction. La taille d'un canton est dimensionnée en fonction de la longueur nécessaire à l'arrêt du train. Elle dépend de la vitesse limite, de la géométrie des voies et, de façon négligeable, du poids du train. Sur la Figure ci-dessus, le train entrant dans le canton 1 circule à voie libre tandis que le train dans le canton 3 est au bloc, il doit donc s'arrêter en attendant que la voie se libère.

Flux de voyageurs denses

Transilien transporte environ 3,4 millions de voyageurs avec plus de 6 200 trains (ce qui représente 70 % du trafic national) chaque jour, le tout sur à peine 10 % du réseau ferré national. Les flux de voyageurs de Transilien vont majoritairement vers Paris en heures de pointe du matin (6h-10h), et vers la banlieue en heures de pointe du soir (16h-20h). En effet, les voyageurs sont majoritairement des actifs (76 %) qui font des trajets domicile-travail. Or la majorité des emplois en Île-de-France se concentrent dans Paris ou en proche banlieue (68 % des emplois se situent sur moins de 6 % de la surface de la région Île-de-France). Les flux de voyageurs sont donc concentrés dans le temps et dans l'espace. En conséquence, le réseau est saturé par endroit avec une densité maximale de trains, un toutes les trois minutes, et des densités de voyageurs dépassant le *taux d'occupation*[8] butoir de 80 % des places assises. Le seuil de 80 % des places assises peut paraître faible, cependant les temps de trajet des voyageurs utilisant les trains de banlieue sont des longs, ce qui incite l'opérateur à éviter les situations où les voyageurs sont debout. Au même moment sur certaines branches de Transilien, la fréquence de trains est d'un par heure et les taux d'occupation sont autour de 30 %. Transilien est un réseau à deux vitesses : des circulations très denses

8. Le taux d'occupation est égal à la la charge à bord divisée par la capacité.

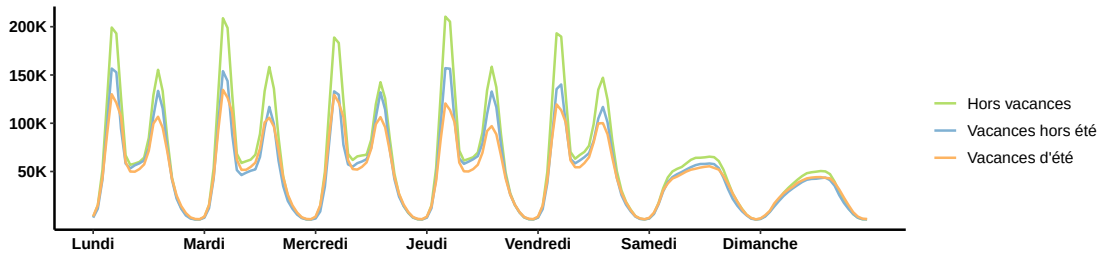


FIGURE 2.3 – Fréquentation hebdomadaire sur le réseau Transilien (moyennée sur l'année 2019 - hors mois de décembre) en nombre d'entrées sur le réseau Transilien pendant et hors vacances.

pour les gares proches de Paris en heures de pointe ; des circulations diffuses pour les gares éloignées de Paris en heures creuses. L'exploitation en zone dense est un enjeu pour Transilien car (1) un retard sur un train se propage immédiatement aux autres trains (2) les arrêts sont autant de séquences à gérer parfaitement pour ne pas rallonger le temps de trajet car ils représentent entre 15 et 30 % du temps de trajet total (3) il y a souvent plusieurs branches qui convergent vers une même branche ce qui impose un respect strict des horaires (4) Transilien partage par endroit ses voies avec d'autres opérateurs ferroviaires dont les trains grandes lignes et le transport de marchandises.

Définition du Mass Transit

En zone dense, la demande s'exprime sous la forme d'un flux de voyageurs : ceux-ci arrivent en gare de façon spontanée. À l'inverse, en zone peu dense, les voyageurs arrivent en fonction d'horaires fixes. Une métaphore empruntée à Pierre Messulam est qu'en zone dense les voyageurs se déversent sur les quais comme l'eau dans une baignoire ; le départ des trains est comme l'ouverture du bouchon d'une baignoire permettant d'évacuer régulièrement les voyageurs. À Transilien, la courbe de fréquentation par heure d'une journée en semaine, peut être graphiquement représentée par un M dont les sommets se situent aux heures de pointe comme sur la Figure 2.3. Le M perd une jambe le week-end pour ne plus former qu'une montagne, avec un sommet vers 17h. Cette connaissance globale de la demande guide la construction de l'offre de référence en la décomposant en heures de pointe et en heures creuses.

La zone peu dense ainsi que le partage du réseau avec d'autres opérateurs imposent une exploitation à l'horaire, c'est-à-dire de circuler suivant des horaires connus par les voyageurs. À l'inverse, l'exploitation à la fréquence, typique des réseaux de métros ou de tramways, consiste à garantir un espacement entre les différents véhicules. À Transilien, pour répondre à la demande, le nombre de trains est à certains endroits si important, que même si les trains circulent à l'horaire, ils donnent l'impression de circuler à la fréquence. Pour preuve, Transilien affiche sur les écrans à quai le temps d'attente avant le prochain train. Ces trains ne forment alors qu'un flux. La rencontre des flux de voyageurs incontrôlables, et des flux de trains, très sensibles

aux aléas, est au cœur de ce qu'on nomme le Mass Transit. Nous en donnons une définition formelle ci-dessous.

Définition 1 *Le Mass Transit*[9] *est un réseau de transports en commun où la synchronisation des flux de voyageurs et des flux de trains est une nécessité pour maximiser sa capacité et minimiser sa congestion. Le Mass Transit ouvert est un réseau dont l'infrastructure est partagée entre le réseau Mass Transit et d'autres circulations extérieures.*

Transilien est un opérateur en Mass Transit ouvert qui doit gérer des trains avec les contraintes du métro (nombreux arrêts et des flux de voyageurs incontrôlables) et les contraintes du ferroviaire (circulation sur le réseau ferré national avec respect des horaires stricts).

Pour développer cette expertise à la SNCF, Transilien a créé la Mass Transit Academy qui promeut la culture du Mass Transit. Le Lab' Mass Transit Academy est la partie dédiée à l'innovation de cette entité. Un DataLab' s'est créé au cours de cette thèse pour valoriser les données de comptage et de circulation des trains. Il a pour mission de développer le savoir-faire de Transilien et d'améliorer le parcours des voyageurs en Mass Transit. Le DataLab' est une entité souple qui porte des sujets de recherche et développement pragmatiques et fondamentaux. Un des enjeux de cette thèse est de donner à Transilien, au monde académique et aux autres acteurs de la mobilité, les clés de compréhension et d'anticipation de l'impact des flux de voyageurs sur la régularité des flux de trains en Mass Transit. Cette thèse ne traite pas des problématiques d'optimisation de la grille horaire mais s'intéresse à l'estimation et la prévision des temps de stationnement et des flux de voyageurs.

2.1.3 Plan du chapitre

Dans la Section 2.2, nous donnons les principaux éléments de compréhension de la structure de l'offre ferroviaire en nous focalisant sur Transilien. Nous présentons dans la Section 2.3 l'évolution de la mesure de la demande grâce aux comptages automatiques permettant des usages en temps réel. Nous présentons nos principales contributions dans la Section 2.4 sur l'estimation et la prévision aux services d'une meilleure compréhension et anticipation de l'interaction des flux de voyageurs et des flux de trains *i.e.* de l'offre et de la demande.

2.2 Offre ferroviaire

La construction de l'offre ferroviaire est un compromis visant à optimiser la capacité d'emport (en nombre de places offertes par période de temps) face à une

9. La traduction littérale serait transport en commun

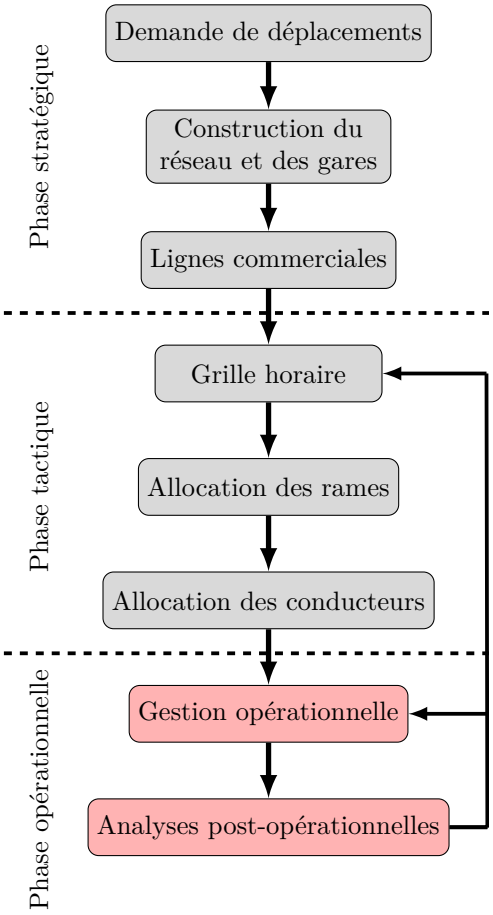


FIGURE 2.4 – Étapes de la planification ferroviaire, adaptation libre en français de Goverde [2005]. En rouge, les étapes importantes pour cette thèse.

demande qui est l'agglomération de décisions individuelles. Ce compromis tient compte des contraintes économiques (coûts et ressources) et des contraintes physiques du mode ferroviaire (espacement des trains). L'objectif de cette partie est de donner les éléments de base de la planification et de l'exploitation des trains de banlieue.

Nous présentons dans la Section 2.2.1, les différentes étapes de la planification ferroviaire d'un point de vue théorique. Nous insistons dans la Section 2.2.2 sur la conception des grilles horaires et leur évaluation à Transilien. Nous montrons dans la Section 2.2.3 le lien entre la gestion opérationnelle des circulations et l'arrivée des trains automatiques pour Transilien.

2.2.1 Étapes de la planification ferroviaire

L'arrêt à quai est le lieu de la rencontre entre les voyageurs et le système de transport. Cette rencontre est l'aboutissement du long processus de la planification ferroviaire, représenté sur la Figure 2.4. Goverde [2005] et Caprara et al. [2007] divisent celle-ci en trois phases exécutées successivement. La première phase stratégique (à long terme *i.e.* supérieur à 5 ans), aussi appelée *Line Planing Problem*, consiste à estimer la demande stratégique, construire le réseau de transport et commander un volume de rames/conducteurs pour pouvoir ensuite construire le plan de transport au cours de la phase tactique. La seconde phase tactique est réalisée à moyen terme *i.e.* un an en avance. Elle est aussi appelée *Train Timetabling Problem*. Elle consiste à construire la grille horaire, à allouer les rames et à allouer les conducteurs à des courses [10]. La troisième phase opérationnelle, à court terme et en temps réel, est la gestion opérationnelle des circulations. Le processus de planification ferroviaire se fait en entonnoir au sens où plus on est proche de l'heure de départ du train moins il est possible de changer des éléments. Nous évoquons ici les différentes phases de la planification ferroviaire en insistant particulièrement sur les phases opérationnelle et post-opérationnelle.

Phase stratégique

La phase stratégique consiste à imaginer et concevoir un mode d'exploitation d'une ligne de transport. La première étape est de prévoir le plus fidèlement possible la demande, appelée demande stratégique. Cette étape est détaillée dans la Section 2.3.1. Une fois cette demande estimée, la seconde étape consiste à adapter le réseau (rails, caténaires, gares, etc.) éventuellement en construisant de nouvelles infrastructures ce qui est toujours longs et coûteux. L'infrastructure et la demande stratégique permettent de définir la ligne commerciale qui comprend un projet d'offre, *i.e.* une fréquence de trains souhaitée aux différentes gares par sens et par heure. Ce projet d'offre permet ensuite de commander des rames et d'embaucher les conducteurs nécessaires à sa réalisation. Les caractéristiques physiques des rames (accélération et freinage, nombre de portes, largeur des portes, etc.), ainsi que leurs performances, ont ensuite un impact direct sur la planification et l'exploitation ferroviaire. Dans cette thèse, ces éléments seront considérés comme fixes. Davantage de détails sur cette phase peuvent être trouvés dans l'article de Caprara et al. [2007] ou dans les chapitres 10 à 15 du livre de référence de Ceder [2016].

Phase tactique

La phase tactique consiste à construire un plan de transport. Ce dernier est composé de trois éléments : la grille horaire, le roulement des rames et le roulement

10. Une course est un trajet de train entre une gare origine et une gare terminus avec une liste des arrêts intermédiaires.

des conducteurs. La grille horaire est la formalisation du projet d'offre émis pendant la phase stratégique. Elle est partagée avec l'autorité organisatrice et elle comprend l'ensemble des courses prévues. La grille horaire définit pour chaque course k l'ensemble des heures d'arrivée et de départ à chaque gare s. La grille horaire fige donc les temps de stationnement théoriques, c'est pourquoi nous développons dans la Section 2.2.2, les méthodes de conception de la grille horaire à Transilien. Nous n'abordons pas ici la problématique de la construction du roulement des rames et des conducteurs. Davantage de détails sur cette phase peuvent être trouvés dans Alfieri et al. [2006] pour le roulement des rames et dans Caprara et al. [1997] pour le roulement des conducteurs.

Phases opérationnelle et post-opérationnelle

La phase opérationnelle. La phase opérationnelle consiste à s'assurer de la bonne exécution du plan de transport ainsi qu'à gérer les aléas pour assurer le plan de transport le moins dégradé possible en cas de perturbations, qui surviennent tous les jours en zone dense. Une perturbation est un événement imprévu comme une panne d'infrastructure, une panne d'un train, un malaise de voyageur, un blocage de portes ou un signal d'alarme qui peut être mineure ou majeure. Une perturbation est majeure si elle nécessite une ré-allocation des rames et/ou des conducteurs.

Une perturbation, mineure ou majeure, entraine systématiquement du retard pour au moins un train, c'est-à-dire une déviation par rapport à son sillon théorique. Ce retard a deux conséquences. La première est que le train ne circule plus sur son sillon ce qui va potentiellement causer des conflits de circulations. En effet, le train qui précède le train en retard vient rattraper son sillon. Dans ce cas, le train en retard cause des retards secondaires. La deuxième est que ces retards vont entrainer une accumulation de voyageurs dans les gares avals, du fait de l'allongement de l'espacement entre les trains. Cet afflux inhabituel de voyageurs entrainera mécaniquement un allongement des temps de stationnement, ce qui enclenchera une boucle de retro-action causant encore davantage de retards. C'est l'effet boule de neige.

Le travail de cette thèse consiste à développer des modèles capables de recalculer des temps de stationnement en temps réel en fonction des flux de voyageurs afin d'anticiper et de prévenir ces effets boule de neige. La recherche en transport s'intéresse de près au sujet de la re-planification en temps réel Cacchiani et al. [2014].

L'analyse post-opérationnelle. L'analyse post-opérationnelle est considérée comme faisant partie de la phase opérationnelle toutefois les flèches de la Figure 2.4 indiquent que les résultats de l'analyse post-opérationnelle peuvent servir à la construction de grilles horaires (phase tactique) ou à la gestion opérationnelle des circulations (phase opérationnelle).

L'analyse post-opérationnelle des données de circulation est un sujet de recherche d'actualité en lien avec la croissance du volume de données dans le ferroviaire [Ghofrani et al., 2018]. Goverde and Meng [2011] montrent que l'analyse post-opérationnelle des circulations est essentielle à la planification ferroviaire pour trois grandes raisons : d'abord, pour estimer précisément certains paramètres nécessaires à la construction de la grille horaire (temps de stationnement, temps de parcours, etc.) ; ensuite, pour identifier des problèmes structurels dans la planification ou l'exploitation comme Graffagnino [2013] et Corman and Henken [2022] ; enfin, pour déterminer les causes des retards afin d'isoler les responsabilités, cf. voyageurs, infrastructure, rames, etc. Par exemple, Daamen et al. [2009], Rößler et al. [2021] distinguent les retards primaires des retards secondaires et leurs causes. L'analyse post-opérationnelle permet également d'entrainer des modèles d'*apprentissage statistique*[11] utiles à la gestion opérationnelle des circulations comme le propose Kecman [2014] dans sa thèse et dans ses travaux de recherche [Kecman and Goverde, 2015, Corman and Kecman, 2018].

Notre objectif représenté par deux flèches de renvoi sur la Figure 2.4, est d'exploiter les données post-opérationnelles pour aider à la gestion opérationnelle des temps de stationnement, et idéalement à leur dimensionnement lors de la construction de la grille horaire.

2.2.2 Grille horaire et analyse post-opérationnelle

Dans cette section, nous abordons conjointement la problématique de la construction de la grille horaire et de l'analyse post-opérationnelle à Transilien. L'objectif est de montrer que ces deux étapes sont complémentaires et que Transilien a intérêt à renforcer les allers-retours entre ce qui a été prévu et ce qui a été effectivement réalisé, en particulier en ce qui concerne les temps de stationnement.

Nous présentons d'abord les principes de base de la construction de grille horaire à Transilien, en insistant sur le dimensionnement des temps de parcours et des temps de stationnement. Nous présentons ensuite les différentes mesures des heures d'arrivée et de départ à Transilien. Nous concluons cette partie sur les principes d'analyse post-opérationnelle des temps de stationnement à Transilien.

Construction d'une grille horaire

Une grille horaire comprend un ensemble de courses ou numéros de train k uniques par jour. Un numéro d'un train est défini par le numéro de sillon qu'il emprunte, que ce soit partiellement ou totalement. Un numéro de sillon vérifie les contraintes suivantes : il comprend 6 chiffres au plus ; il est unique par jour ; il respecte la parité du sens de circulation : pair (vers Paris) et impair (vers la banlieue). Par la suite, nous noterons les gares desservies s.

11. Machine learning

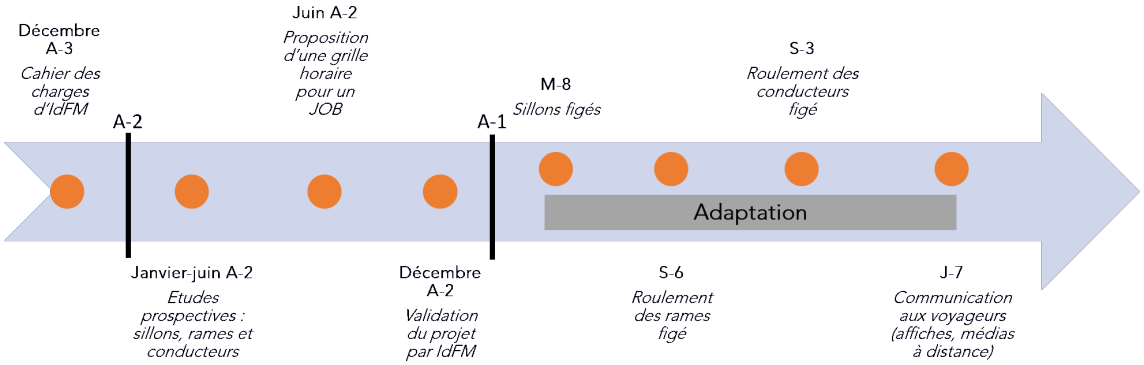


FIGURE 2.5 – Étapes de la planification ferroviaire à Transilien.

À Transilien, les grilles horaires sont cycliques, au sens où, d'une part, elles sont identiques pour un même type de jour, et d'autre part, elles sont construites à partir d'un ensemble de missions répétées suivant un même motif en heures creuses ou en heures de pointe. La représentation de la grille horaire sous plusieurs formes, pour les voyageurs comme pour les agents, montre qu'elle est cruciale pour la planification et l'exploitation des trains de banlieue. Du côté des voyageurs, elle apparait le plus souvent sous la forme d'un tableau à double entrée, avec les horaires de départ (en colonne) pour chaque gare (en ligne) comme sur la Figure 2.6. Du côté des agents, sa forme emblématique est le graphe espace-temps (GET) comme sur la Figure 2.7. Un graphe espace-temps, aussi appelé graphe de Marey, permet de rapidement identifier des conflits de circulations.

Nous décrivons succinctement la construction du plan de transport à Transilien en particulier lors de la construction du service annuel. Un service annuel est le résultat d'intenses discussions entre Transilien, SNCF-Réseau et IdFM. Nous représentons les principales étapes de ces négociations sur la Figure 2.5. La proposition de grille horaire, pour un jour de semaine type de SNCF-Réseau, se fait plus d'un an et demi avant le déploiement effectif du service annuel, le deuxième dimanche de décembre. Par ailleurs, les grilles horaires sont figées huit mois avant cette date. Ainsi, pour modifier un élément important de la grille horaire, comme les temps de stationnement, il faut anticiper ce changement très en amont. Les temps de stationnement sont revendiqués par le demandeur [Transilien] et attribués par SNCF-Réseau en tenant compte des contraintes d'exploitation. Les agents de Transilien en charge de la construction du plan de transport *i.e.* de l'étude prospective (5 ans en amont) à l'adaptation (jusqu'à 7 jours avant) en passant par la conception (jusqu'à 3 mois avant), sont réunis au sein des plateaux de planification unique (PPU).

Les horaires de la grille horaire sont précis à la déca-seconde côté opérateur et à la minute pour les voyageurs. Au Japon, Tokyo Metro trace ses horaires par pas de 5 secondes. On associe un temps de trajet à une course qui comprend l'ensemble des temps de parcours inter-gares plus les temps de stationnement. Par la suite, nous présentons les stratégies à Transilien pour tracer les temps de parcours et les temps de stationnement.

Nom du train Notes à consulter	APOR	APOR	APOR	APOR	ADDO LV	APOR SD	ADDO LV	ADDO LV	APOR SD	ADDO LV	ADDO LV	APOR SD	ADDO LV
Pontoise	04:36	05:06	05:36	06:06	06:22	06:36	06:37	06:52	07:06	07:07	07:22	07:36	07:37
Saint-Ouen l'Aumône	04:39	05:09	05:39	06:09	06:24	06:39	06:39	06:54	07:09	07:09	07:24	07:39	07:39
Saint-Ouen l'Aumône Liesse	04:42	05:12	05:42	06:12	06:28	06:42	06:43	06:58	07:12	07:13	07:28	07:42	07:43
Pierrelaye	04:45	05:15	05:45	06:15	06:31	06:45	06:46	07:01	07:15	07:16	07:31	07:45	07:46
Montigny Beauchamp	04:49	05:19	05:49	06:19	06:35	06:49	06:50	07:05	07:19	07:20	07:35	07:49	07:50
Franconville Plessis Bouchard	04:52	05:22	05:52	06:22	06:39	06:52	06:54	07:09	07:22	07:24	07:39	07:52	07:54
Cernay (Val d'Oise)	04:55	05:25	05:55	06:25	06:42	06:55	06:57	07:12	07:25	07:26	07:41	07:55	07:56
Ermont-Eaubonne	04:58	05:28	05:58	06:28	06:45	06:58	07:00	07:15	07:28	07:30	07:45	07:58	08:00
Champ de Courses d'Enghien	05:00	05:30	06:00	06:30	|	07:00	|	|	07:30	|	|	08:00	|
Enghien les Bains	05:03	05:33	06:03	06:33	06:49	07:03	07:04	07:19	07:33	07:34	07:49	08:03	08:04
La Barre Ormesson	05:05	05:35	06:05	06:35	|	07:05	|	|	07:35	|	|	08:05	|
Epinay Villetaneuse	05:07	05:37	06:07	06:37	06:53	07:07	07:07	07:22	07:37	07:37	07:52	08:07	08:07
Saint-Denis	05:11	05:41	06:11	06:41	06:57	07:11	07:11	07:26	07:41	07:41	07:56	08:11	08:11
Gare du Nord (Surface)	05:17	05:47	06:17	06:47	07:03	07:17	07:18	07:33	07:47	07:48	08:03	08:17	08:18

Nom du train Notes à consulter	ADDO LV	APOR SD	ADDO LV	ADDO LV	APOR SD	ADDO LV	ADDO LV	APOR	APOR	APOR	APOR	APOR	APOR
Pontoise	07:52	08:06	08:07	08:22	08:36	08:37	08:52	09:06	09:36	10:06	10:36	11:06	11:36
Saint-Ouen l'Aumône	07:54	08:09	08:09	08:24	08:39	08:39	08:54	09:08	09:38	10:08	10:38	11:08	11:38
Saint-Ouen l'Aumône Liesse	07:58	08:12	08:13	08:28	08:42	08:43	08:58	09:12	09:42	10:12	10:42	11:12	11:42
Pierrelaye	08:01	08:15	08:16	08:31	08:45	08:46	09:01	09:15	09:45	10:15	10:45	11:15	11:45
Montigny Beauchamp	08:05	08:19	08:20	08:35	08:49	08:50	09:05	09:18	09:48	10:18	10:48	11:18	11:48
Franconville Plessis Bouchard	08:09	08:22	08:24	08:39	08:52	08:54	09:09	09:22	09:52	10:22	10:52	11:22	11:52
Cernay (Val d'Oise)	08:11	08:25	08:26	08:41	08:55	08:56	09:11	09:25	09:55	10:25	10:55	11:25	11:55
Ermont-Eaubonne	08:15	08:28	08:30	08:45	08:58	09:00	09:15	09:27	09:57	10:27	10:57	11:27	11:57
Champ de Courses d'Enghien	|	08:30	|	|	09:00	|	|	09:30	10:00	10:30	11:00	11:30	12:00
Enghien les Bains	08:19	08:33	08:34	08:49	09:03	09:04	09:19	09:32	10:02	10:32	11:02	11:32	12:02
La Barre Ormesson	|	08:35	|	|	09:05	|	|	09:34	10:04	10:34	11:04	11:34	12:04
Epinay Villetaneuse	08:22	08:37	08:37	08:52	09:07	09:07	09:22	09:37	10:07	10:37	11:07	11:37	12:07
Saint-Denis	08:26	08:41	08:41	08:56	09:11	09:11	09:26	09:40	10:10	10:40	11:10	11:40	12:10
Gare du Nord (Surface)	08:33	08:47	08:48	09:03	09:17	09:18	09:33	09:47	10:17	10:47	11:17	11:47	12:17

FIGURE 2.6 – Grille horaire de la branche Paris Gare du Nord - Pontoise sur la ligne H.

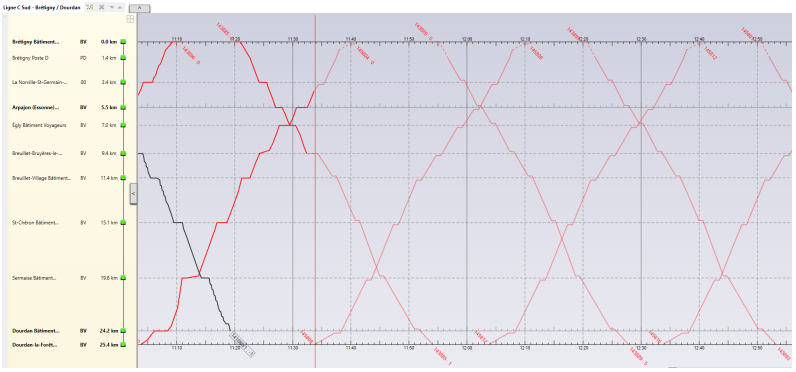


FIGURE 2.7 – Graphe espace-temps sur la ligne C.

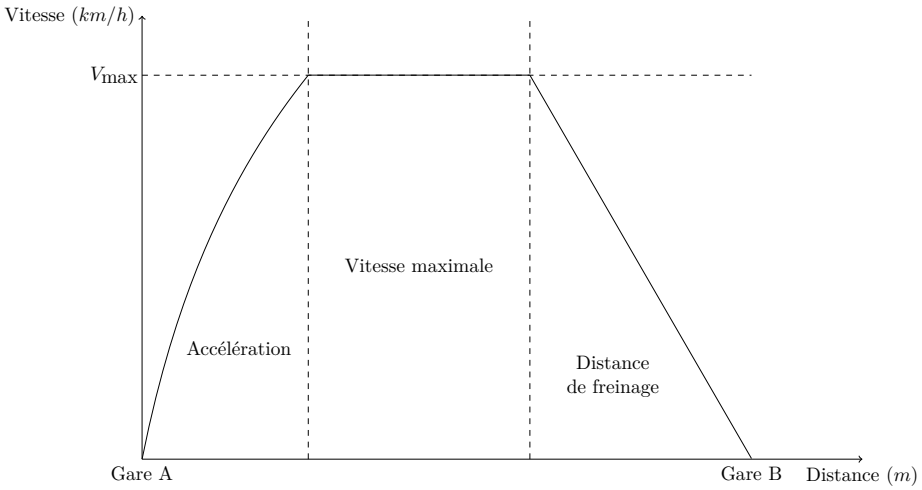


FIGURE 2.8 – Exemple fictif d'une marche tracée entre deux gares.

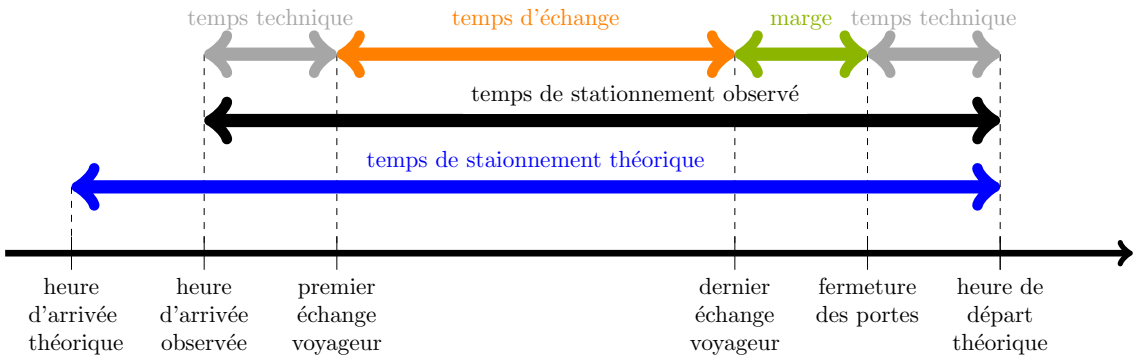


FIGURE 2.9 – Décomposition du temps de stationnement.

Le temps de parcours (running time en anglais) est calculé par SNCF-Réseau en coopération avec Transilien en fonction des caractéristiques de la voie (limites de vitesse, implantation de la signalisation, puissance électrique disponible, etc.) et des rames (vitesse d'accélération, performances de freinage). Il s'agit d'abord de tracer une marche avec un temps de parcours minimal entre deux gares *i.e.* une marche tendue. La Figure 2.8 en donne un exemple simplifié avec en abscisse la distance en mètre et en ordonnée la vitesse en km/h. La marche tendue est tracée avec une accélération maximale en sortie de gare A pour atteindre le plus rapidement possible la limite de vitesse, notée $V_{\max}$. La vitesse maximale est conservée jusqu'au dernier moment pour freiner et arriver à la gare B. Le temps de parcours de cette marche tendue n'est pas réaliste vis-à-vis de la variabilité des comportements de conduite.

La marche tendue est assouplie en ajoutant une marge d'exploitation. L'union internationale des transports publics (UITP) recommande d'ajouter entre 5 et 10 % de marge à la marche tendue pour éviter l'effet de propagation des retards entre les trains. L'idée est qu'un retard primaire (causé par un événement extérieur) n'entraine pas de retards secondaires. Une autre manière de positionner des marges, sans détendre les marches, est d'ajouter des sillons fantômes entre deux sillons. Ainsi, si un train est en retard sur son sillon, il n'empiétera pas tout de suite sur le sillon du train suivant. Le temps de parcours associé à cette marche détendue (temps de parcours de la marche tendue majorée de 5 à 10 %) est le temps de parcours planifié. Le calcul des temps de parcours et plus largement de la construction de la grille horaire, est un équilibre à trouver entre *robustesse* et nombre de trains. En effet, une grille horaire est considérée comme robuste si elle permet d'absorber de petits aléas, donc plus il y a de marge plus elle est robuste. Cependant, en contrepartie, une augmentation de la robustesse diminue la capacité du réseau.

Les temps de stationnement sont importants en Mass Transit car ils représentent entre 15 et 30 % du temps de trajet entre une origine et une destination. Cette part des temps de stationnement dans le temps de trajet est due au grand nombre d'arrêts sur le réseau Transilien, en moyenne 10 arrêts par trajet, ainsi qu'à leur durée théorique entre 40 et 60 secondes. On rappelle que l'heure de départ est stricte : un conducteur de train ne peut théoriquement pas partir en avance même si l'échange voyageur est terminé. Nous décomposons les temps de stationnement

(en noir) en trois sous-parties sur la Figure 3.1 : le temps technique (en gris), le temps d'échange (en orange) et la marge (en vert). Le temps technique est le temps nécessaire à l'ouverture et la fermeture des portes, auquel s'ajoute le temps de mise en marche du train. Les temps techniques sont bien connus à Transilien au travers des chronogrammes qui permettent au conducteur d'enclencher sa séquence de départ pour partir à l'heure. Cette séquence comprend le déclenchement du signal sonore (ronfleur), la fermeture des portes, le desserrage des freins et la mise en marche du train. Les temps techniques varient entre 12 secondes pour les modèles de rame Z50000/NAT et 15 secondes pour les Z57000/Regio 2N.

Le temps d'échange est le temps nécessaire à l'échange d'au moins 99 % des voyageurs. Nous excluons les retardataires du temps d'échange. Les temps d'échange sont très mal connus à Transilien car il n'y a aucune procédure interne permettant de relier le nombre de montées et de descentes au temps de stationnement. La marge est le temps restant une fois le temps technique et le temps d'échange écoulés. Cette marge existe pour les trains de banlieue car les conducteurs doivent attendre leur heure de départ stricte. Dans les réseaux de métros, tramways et bus, il n'y a pas de marges au niveau des temps de stationnement car ils fonctionnent à la fréquence. C'est pourquoi la marge est souvent oubliée dans les articles sur le sujet, elle est cependant déterminante dans notre travail.

Mesures des circulations

SNCF-Réseau et Transilien contrôlent les circulations avec plusieurs systèmes de mesure au sol ou à bord que l'on regroupe sous le terme d'*automatic vehicle localisation* (AVL). On rappelle que le train (le bord) et la voie (le sol) sont liés par le système de cantonnement. Nous résumons l'ensemble de ces systèmes de mesure dans la Table 2.1. Les données BREHAT sont les données les plus complètes, elles servent de référence notamment lorsque l'on souhaite avoir des taux de couverture en données de comptage [12]. Les données ATESS sont les plus précises, elles permettent d'étudier précisément les temps de stationnement. Les données CAVE sont les données de comptage reçues en temps réel à chaque départ de gare. Nous décrivons brièvement chacune de ces sources de données.

BREHAT. Le premier système de suivi des circulations est BREHAT (base des résultats de l'exploitation habiles à d'autres tâches). BREHAT est la source officielle du suivi de la régularité [13] des trains. Ce système repose sur des balises positionnées au niveau des points remarquables (PR) [14] permettant d'identifier et de localiser un train. Le flux de données BREHAT est acquis en temps réel à chaque passage de

12. Un taux de couverture est le pourcentage d'arrêts ayant été desservis par un train équipé de système de comptage mais dont l'information de comptage n'existe pas ou n'est pas utilisable.

13. La régularité est le respect par les trains des horaires prévus pour leur circulation.

14. Les points remarquables sont des points du réseau ferroviaire comportant des particularités d'exploitation (aiguille, gare, etc.).

TABLE 2.1 – Caractéristiques des différents systèmes de suivi des circulations à Transilien.

Nom	Localisation	Technologie	Point de mesure	Temps de stationnement	Fraicheur
BREHAT	sol	balises	PR	arrivée et départ projetés	temps réel
ATESS	bord	odométrie	2 s	dernier et premier tour de roue	J+23
CAVE	bord	odométrie	sortie de gare	création et envoi du fichier	temps réel
CAVE redressé	bord	odométrie	sortie de gare	création et envoi du fichier redressées	temps réel

balise. Les balises peuvent être éloignées de l'entrée ou de la sortie d'une gare ainsi BREHAT reconstruit des heures d'arrivée et de départ. Nous utiliserons les données BREHAT pour leur complétude et leur disponibilité en temps réel via l'*API Course* à Transilien.

ATESS. L'odométrie[15] alimente le système à bord ATESS (acquisition, traitement des événements sécurité en statique), remplissant une fonction analogue aux boîtes noires des avions. Ce système mesure la vitesse du train toutes les deux secondes en y associant un horaire. Ces mesures croisées avec les plans de voies permettent de reconstruire des heures de passage aux PR et des heures d'arrivée et de départ en gare. La précision d'ATESS est telle qu'à Transilien, il est admis que ce système mesure le dernier tour de roue et le premier tour de roue en gare, ce qui en fait la mesure de référence pour l'étude des temps de stationnement. Ce système a toutefois deux limites. La première est qu'il nécessite de récupérer les données manuellement enregistrées sur une cassette. La seconde est que cette cassette n'est récupérée en moyenne que tous les 23 jours. Ce sont donc des données très précises mais disponibles tardivement.

CAVE. Le système de comptage CAVE (comptage automatique des voyageurs embarqué) qui utilise des capteurs infra-rouges au niveau des portes, permet de compter le nombre de montées et de descentes par porte. Ces données sont stockées dans un fichier de comptage dont l'heure de création et d'envoi correspond à l'heure d'arrivée et de départ plus quelques secondes. Ce système a deux limites. La première est que seule une partie des rames ($\approx 50\%$ en 2022) sont équipées de CAVE. La seconde, plus fondamentale, est que la mesure de l'heure de départ dépend de l'émission du fichier de comptage qui ne se fait qu'après que le train a atteint une vitesse de 10 km/h. Cette vitesse minimale crée un décalage dans la

15. L'odométrie est une technique pour mesurer la vitesse et le distance parcourue à l'aide du nombre de tours de roues.

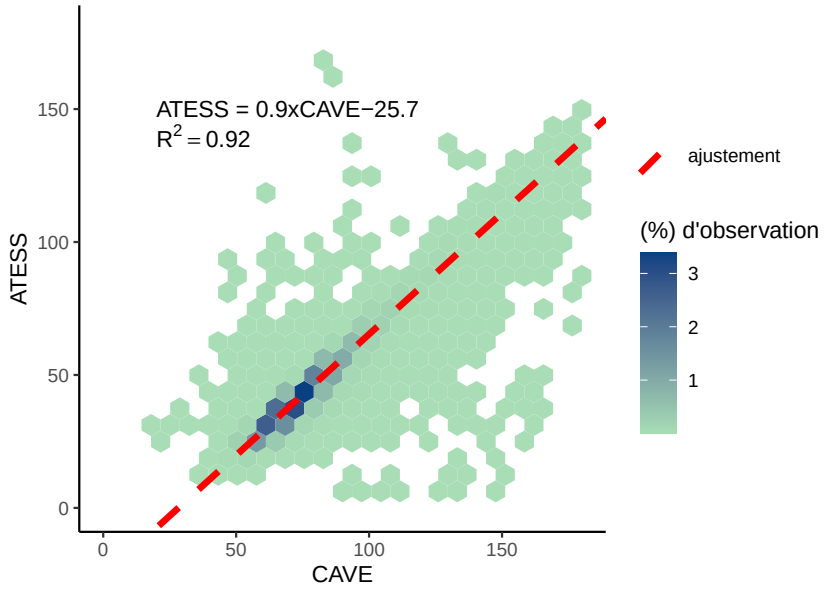


FIGURE 2.10 – Comparaison des mesures de temps de stationnement du CAVE et d'ATESS des arrêts des trains de la ligne H de janvier à mai 2018. Estimation d'une équation affine entre les deux mesures permettant de redresser les temps de stationnement CAVE.

durée des temps de stationnement mesurés du CAVE par rapport à ceux d'ATESS comme illustré sur la Figure 2.10.

CAVE redressé. La Figure 2.10 représente les observations des temps de stationnement de janvier à mai 2018 ($\approx$ 200 000 arrêts) des arrêts des trains dans toutes les gares de la ligne H. Ces données sont issues de deux logiciels internes : REXMICRO pour les données ATESS et CHATELET pour les données CAVE. Le but de cette figure est de montrer qu'il est possible de redresser en temps réel les données de temps de stationnement du CAVE. On constate qu'il suffit d'estimer une équation affine simple des temps de stationnement du CAVE sur ceux d'ATESS pour obtenir un modèle de bonne qualité avec un R^2 de 0.91. Ce modèle estimé peut ainsi être exploité en temps réel à partir de la seule mesure du temps de stationnement CAVE, facile d'accès comme présenté dans la Section 2.3.

Ces quatre mesures des heures d'arrivée et de départ sont cruciales pour la gestion opérationnelle des circulations, lorsqu'elles sont accessibles en temps réel, mais aussi pour l'analyse en phase post-opérationnelle.

Analyses post-opérationnelles des temps de stationnement

L'analyse post-opérationnelle pour Transilien se résume trop souvent au calcul d'indicateurs pour IdFM ou SNCF-Réseau : ponctualité, qualité de service et régularité. Les refontes d'offre globales sont rares mais permettent souvent

d'augmenter la qualité de service du plan de transport. Un document interne à Transilien décrit comment analyser *a posteriori* les temps de stationnement à l'aide des données ATESS. Cependant, les préconisations faites dans ce document paraissent peu performantes car elles consistent à se focaliser sur les temps de stationnement observés, qui dépassent les temps de stationnement théoriques, sans faire explicitement attention aux trains qui arrivent en avance. Or, on sait et on verra dans le Chapitre 3 que les trains en avance ont des temps de stationnement mécaniquement beaucoup plus longs car ils ne peuvent pas repartir avant leur heure de départ théorique. Dans ce document, il est également conseillé de placer le maximum de marge sur les temps de stationnement, ce qui a pour conséquence directe de compliquer encore la séparation entre le temps d'échange et la marge. En réalité, il serait plus utile de proposer une méthode de calcul des temps de stationnement aussi robuste que celle des temps de parcours. Cette dernière se baserait sur le calcul d'un temps de stationnement tendu auquel on ajouterait entre 5 et 10 % de marge. Ces deux préconisations impliquent une tendance à l'augmentation des temps de stationnement et donc aux rallongements des temps de trajet, sans remise en cause de l'équilibre robustesse / temps de trajet. Nous proposons dans le Chapitre 3, une nouvelle méthode pour estimer des temps de stationnement suivant les flux de voyageurs auquel on peut ensuite ajouter une marge si besoin.

2.2.3 Offre en opérationnel : des flux de trains

L'offre en opérationnel consiste essentiellement à suivre la bonne exécution du plan de transport, c'est la gestion opérationnelle des circulations. Nous rappelons qu'en Mass Transit il y a systématiquement des aléas qui viennent le perturber. Nous présentons dans la Section 2.2.3 les enjeux de la gestion opérationnelle des circulations effectuée par le centre opérationnel Transilien (COT) avec l'appui du centre opérationnel de la gestion des circulations (COGC). Nous montrons dans la Section 2.2.3 comment les trains automatiques permettraient une gestion plus fine des circulations en cas de perturbations.

Centre opérationnel Transilien

La re-planification en temps réel du plan de transport est effectuée manuellement par des agents. Elle est assurée par le COT qui effectue en situation perturbée les mêmes tâches de planification que celles de la phase tactique mais avec moins de temps et moins de ressources rames et conducteurs libres. Le COT est le chef d'orchestre de la re-planification en temps réel, en ce sens il coordonne la commande de nouveaux sillons ainsi que la ré-allocation des rames et des conducteurs déjà en circulation.

La re-planification est une tâche laborieuse que peu d'outils viennent faciliter. La plupart du temps les COT suivent des scénarios typiques de reprise du trafic après perturbations. Leur objectif est de retrouver rapidement le plan de transport planifié. Des outils d'aide à la re-planification destinés aux COT ont été testés à

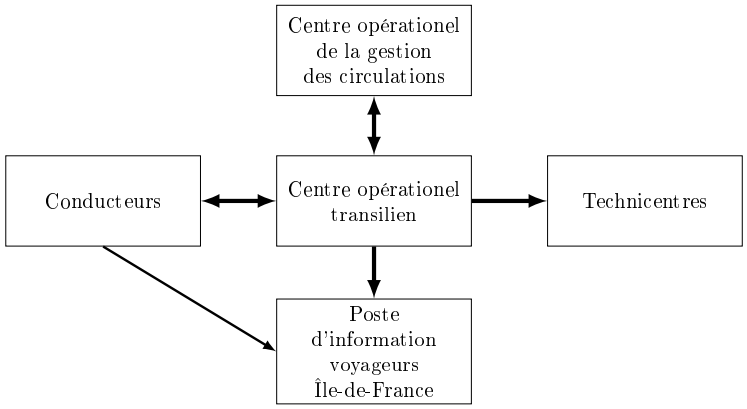


FIGURE 2.11 – Relations entre acteurs de la gestion opérationnelle des circulations à Transilien.

Transilien, [Altazin et al., 2020]. Ils sont pour partie en cours de déploiement. La re-planification opérationnelle est un processus lourd qui n'est activé qu'en cas de perturbations importantes. Le reste du temps, le COT a pour vocation de s'assurer que le plan de transport s'exécute comme prévu. La re-planification fine des temps de stationnement n'est pas une priorité pour la gestion opérationnelle des circulations pour deux raisons. Premièrement parce que lors de petites perturbations, modifier les temps de stationnement nécessite la coordination d'un grand nombre d'acteurs (conducteurs, COT et gares), qui sont difficiles à mobiliser rapidement. Deuxièmement parce que lors de grandes perturbations la priorité du COT est de commander de nouveaux sillons, trouver des rames et des conducteurs, la justesse des temps de stationnement planifiés est alors secondaire. Les trains automatiques permettraient une re-planification plus fine pour de petites comme pour de grandes perturbations car le système de régulation automatique modifie les temps de stationnement sans demander l'autorisation au COT.

Apports des trains automatiques

La recherche sur les trains ou métros automatiques est principalement tirée par les industriels, ce qui freine la diffusion des savoirs, notamment concernant la re-planification des temps de stationnement en temps réel. Le train automatique est un débouché naturel pour appliquer les résultats de cette thèse sur la modélisation et la prévision des temps de stationnement. Les systèmes de commande automatique sont surtout déployés dans les métros mais ont aussi été déployés pour le réseau S-Bahn à Copenhague ou sur l'Elizabeth Line à Londres. Les trains et les métros sont exploités selon différents niveaux d'automatisme (Grades of Automation : GoA) définis par l'UITP. Le plus bas niveau est la marche à vue avec conducteur (GoA 0) utilisée dans l'exploitation des tramways. Le plus haut niveau est l'automatisme sans personnel à bord (GoA 4). Les différents projets de métros automatiques dans le monde sont référencés par l'UITP dans un document très riche [UITP, 2019].

Les trains automatiques circulent avec des systèmes appelés contrôle automatique du trafic ferroviaire (CBTC). Un système CBTC est composé de trois grands

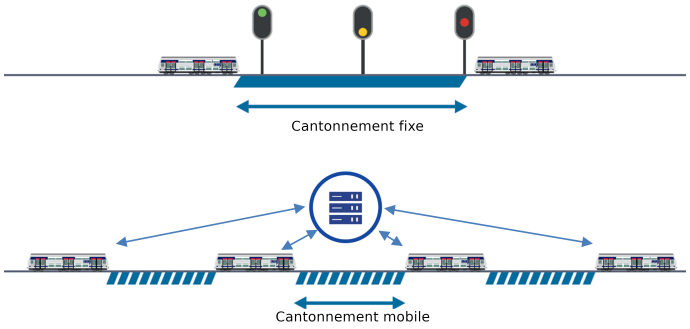


FIGURE 2.12 – Du cantonnement fixe au cantonnement mobile.

blocs : la protection des trains (ATP pour *automatic train protection*), le pilotage automatique (ATO pour *automatique train operation*) et la supervision des trains (ATS pour *automatic train supervision*). La protection automatique des trains (ATP) repose sur la communication entre les trains. Un des composants de l'ATP est le cantonnement mobile, illustré sur la Figure 2.12. Le cantonnement mobile consiste à contrôler de façon dynamique la distance de freinage d'un train. Il permet de réduire la distance minimale entre les trains par rapport au cantonnement fixe et donc d'augmenter la capacité de la ligne. La protection automatique des trains comprend également le contrôle de la vitesse et des portes. La conduite automatique (ATO) regroupe l'ensemble des fonctions permettant de substituer partiellement ou totalement les tâches du conducteur : accélération, freinage, commande des portes et commande du temps de stationnement. La supervision automatique (ATS) se rapproche des missions du centre opérationnel car elle permet de re-planifier les circulations en temps réel. Cette brique peut agir très facilement sur la vitesse, les temps de stationnement et avec l'accord du COT sur le nombre de trains en circulation. L'ATS permet de faire de la re-planification en temps réel de façon beaucoup plus fine et globale que ce qui est possible actuellement dans les centres opérationnels. Malgré ces avancées technologiques, la gestion des temps de stationnement dans les systèmes de métros automatiques est souvent simpliste [Assis and Milani, 2004, Wang et al., 2015] en ce qu'elle repose sur un temps de stationnement minimal qui dépend au mieux des flux habituels [Wang et al., 2015] ou, au pire, complètement dé-corrélé des flux de voyageurs. Un des trois piliers du projet NExTEO (GoA 2) de trains automatiques de la SNCF est de gérer de façon dynamique les temps de stationnement en opérationnel en fonction des flux de voyageurs.

Le projet NExTEO est la brique d'automatisation du tronçon central de l'extension de la ligne E (EOLE) vers l'ouest. NExTEO sera déployé entre les gares de Nanterre-La Folie et Rosa-Parks. Les trains seront en conduite automatique avec un conducteur à bord dont la mission est de gérer les séquences d'ouverture et de fermeture des portes sur 6 gares de la future ligne E et circuleront en conduite assistée sur le reste de la ligne. Les circulations seront denses dans Paris (28 trains par heure) mais plus diffuses en dehors de Paris. Un des enjeux de NExTEO est de faire cohabiter sur une même ligne des trains équipés de NExTEO, et des trains non équipés. L'automatisation par NExTEO permettra en théorie de gérer finement les temps de stationnement en les communiquant directement aux

conducteurs *via* un compte à rebours en cabine [Deau, 2015]. Cette idée de compte à rebours est essentielle pour pouvoir re-planifier en temps réel les temps de stationnement et communiquer facilement aux conducteurs les nouvelles heures de départ. Un tel compte à rebours n'est pas envisageable en dehors de NExTEO car il y a un risque lié à la sécurité si le système n'interagit pas avec la signalisation. Les rames équipées de NExTEO sont les rames Z58000/RER NG qui seront également dotées d'un système de comptage en temps réel des flux de voyageurs. Cette complémentarité entre l'automatisme NExTEO et le CAVE temps réel laisse espérer qu'au moment de la re-planification de l'ATS, les temps de stationnement soient calculés en fonction des flux de voyageurs anticipés.

Résumé de la section offre ferroviaire. La planification ferroviaire est un ensemble d'étapes imbriquées. La modification d'un composant à une étape a des conséquences sur toutes les autres étapes. Par exemple, l'arrivée d'une nouvelle génération de rame (phase stratégique) peut entrainer une réduction des temps techniques (phase tactique) et communiquer en temps réel l'affluence grâce au CAVE (phase opérationnelle). Dans cette thèse, nous ne nous intéressons pas directement à la phase stratégique et à la phase tactique de telle sorte que les grilles horaires et les ressources sont fixes. Nous faisons tout de même une excursion dans la phase tactique, en proposant d'appliquer le modèle statistique des temps de stationnement opérationnel pour tracer différemment les temps de stationnement théoriques. L'objectif est d'adapter la méthode de calcul de temps de parcours au temps de stationnement : un temps de stationnement tendu en fonction des flux de voyageurs auquel on ajoute une marge de 5 à 10 %. Le cœur de cette thèse porte sur la modélisation et la prévision conjointe des flux de trains et des flux de voyageurs. La mesure des heures d'arrivée et de départ et la mesure des flux de voyageurs en temps réel sont des briques essentielles pour ce projet. L'arrivée des trains automatiques est aussi une opportunité pour re-planifier en temps réel les temps de stationnement en fonction des flux de voyageurs.

2.3 Demande voyageurs

Les déplacements en transports en commun de millions de voyageurs quotidiens forment ce que nous appelons la *demande voyageurs*. La demande voyageurs est intrinsèquement liée à l'offre : une offre trop faible par rapport à la demande entrainant de la congestion qui amène des voyageurs à se détourner de l'itinéraire. Nous insistons particulièrement sur l'apport des données automatiques dans l'estimation de la demande en phase tactique et opérationnelle. Nous posons le contexte industriel et scientifique de l'estimation de la demande à toutes les étapes de la planification ferroviaire. Nous reprenons le découpage en phases de la planification ferroviaire pour présenter la demande.

Dans la Section 2.3.1, nous présentons les enjeux d'estimation de la demande stratégique pour dimensionner les infrastructures et les ressources de transports.

Dans la Section 2.3.2, nous insistons sur l'estimation de la demande tactique lors de la construction de la grille horaire et son amélioration continue grâce à l'analyse post-opérationnelle des données de comptage. Nous présentons dans la Section 2.3.3 la demande en opérationnel et son importance pour la gestion opérationnelle des flux de voyageurs.

2.3.1 Demande stratégique

Il s'agit de présenter succinctement les modèles de demande stratégique utilisés pour la conception et la modification d'infrastructures de transport. Comprendre l'estimation de la demande durant la phase stratégique permet de mieux comprendre les sources de données de la demande aux autres phases. L'estimation de la demande en phase stratégique est importante pour objectiver un projet d'investissement routier ou de transports en commun.

Modèles à quatre étapes

La demande stratégique est souvent modélisée à partir d'un modèle à quatre étapes dont le principe a été défini par McNally [2007]. Les quatre étapes du modèle, illustrées sur la Figure 2.13, sont : (1) la génération des déplacements, qui consiste à fixer le nombre de déplacements [16] que chaque individu fera dans une journée. Par exemple, un Francilien fait en moyenne 3,8 déplacements par jour d'après l'enquête globale transport OMNIL [2018]. (2) la distribution des déplacements, qui consiste à projeter spatialement les déplacements d'une zone d'origine à une zone de destination. Cette projection spatiale dépend des caractéristiques socio-démographiques de la zone et de son attractivité. (3) le choix modal, qui consiste à définir le ou les modes associés aux déplacements (marche à pied, vélo, voiture, transports en commun). Les modes disponibles varient en fonction des caractéristiques du foyer (détention d'une voiture, distance à une gare, etc.). (4) le choix d'itinéraire [17], qui consiste à allouer les déplacements à un itinéraire (lignes de transports en commun, routes). L'allocation passe, entre autres, par la minimisation successive du temps de trajet, auquel on peut ajouter des contraintes comme le confort à bord [Tirachini et al., 2013]. L'ajout d'une nouvelle offre de transport ou d'un nouveau projet d'infrastructure se fait soit à l'étape de choix modal, soit à l'étape de choix d'itinéraire. Ce modèle a une limite intrinsèque : le choix modal, et parfois même la distribution des déplacements, dépendent de la fréquence de l'offre. Par exemple, s'il y a peu de trains le soir, un individu prendra sa voiture le matin, même si ce n'est pas optimal, pour être sûr de pouvoir rentrer le soir.

16. Un déplacement peut être formé de plusieurs trajets, chaque trajet étant associé à un mode.
17. Un itinéraire ou une route est un chemin, entre une origine et une destination, emprunté par un voyageur avec un mode de transport.

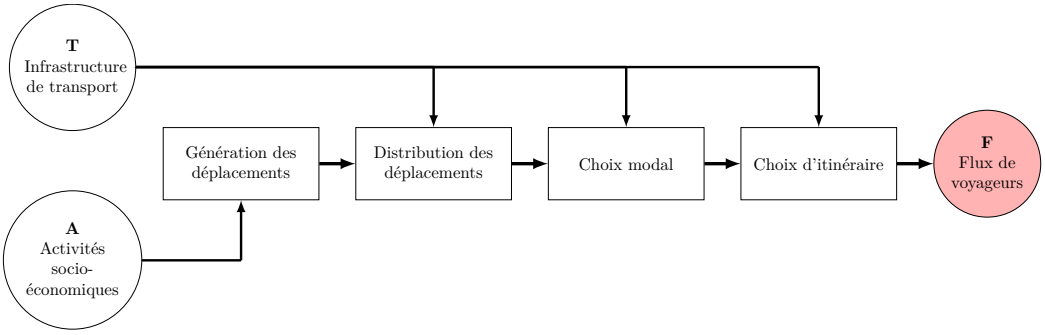


FIGURE 2.13 – Version simplifiée d'un modèle à quatre étapes inspirée librement du chapitre 3 du livre McNally [2007]. En rouge, la partie intéressante pour cette thèse.

Sources de données des modèles à quatre étapes

En Île-de-France, il y a quatre modèles à quatre étapes différents : MODUS pour la Direction régionale et interdépartementale de l'environnement, de l'aménagement et des transports (DRIEAT), GLOBAL pour la RATP, ARES pour SNCF-Transilien et ANTONIN pour IdFM. Le rapport de Massoni et al. [2015], commandé par la préfecture de Paris réalise un audit complet des 4 modèles à quatre étapes d'Île-de-France pour en conclure qu'ils sont globalement convergents mais qu'ils nécessiteraient des investissements massifs pour être plus précis. Ce rapport pose également la question de la rationalisation de l'investissement dans la modélisation de la demande stratégique en Île-de-France. Dans cette thèse, nous étudions l'impact de la demande sur l'offre dans le cas où l'infrastructure et la demande sont déjà figées. Nous nous intéressons aux flux de voyageurs, notés **F**, de la Figure 2.13.

Les données en entrée des modèles à quatre étapes sont de trois natures : des données de demande définies par les habitudes de déplacements et les caractéristiques socio-démographiques des habitants, des données de maillage territorial définies par la géographie du territoire et des données d'offre définies par la performance de l'infrastructure de transport. La connaissance de la demande comprend des données socio-démographiques, un volume d'activité pour chaque foyer dans une journée et le taux de possession de moyen de transport individuel (vélos, voitures). Les données d'activité proviennent principalement de l'enquête globale transport en Île-de-France ou les enquêtes ménages déplacements dans le reste de la France. Ces dernières ont lieu tous les 10 ans car elles sont extrêmement coûteuses. Les données de demande sont obtenues par l'INSEE, l'institut Paris Région, la DREIAT et les enquêtes globales transport. Le maillage territorial est défini par la région. L'offre de transport nécessite de devoir tracer les routes et les lignes de transports auxquelles on associe des capacités et des vitesses. Les données d'offre sont obtenues par la DREIAT, l'institut géographique national, IdFM, la SNCF et la RATP.

2.3.2 Demande tactique

La construction de la grille horaire nécessite des données de demande pour anticiper la fréquentation des trains planifiés. La connaissance de la demande est principalement issue d'enquêtes origine-destination (OD) et d'enquêtes de comptage réalisées tous les quatre ans sur chaque ligne de Transilien. Depuis plus de dix ans, Transilien et IdFM recueillent de plus en plus de données automatiques de demande, dont les comptages automatiques et les validations, qui sont autant d'opportunités pour une meilleure adaptation de l'offre à la demande.

Grille horaire et demande

Aujourd'hui, la grille horaire de Transilien répond à l'offre de référence demandée par IdFM. Cette offre de référence se fonde sur l'estimation de la demande stratégique effectuée à partir des modèles à quatre étapes. Les variations de la demande sont prises en compte par IdFM car le nombre de trains varie en fonction des heures creuses et des heures de pointe ainsi qu'entre les JOB, samedis ou dimanches. Toutefois, Transilien, au moment de construire la grille horaire, ne connait pas la demande associée à chaque train *i.e.* le nombre de montées, de descentes et le taux d'occupation. Or, ces informations sont essentielles pour calculer des temps de stationnement adaptés aux flux de voyageurs. Le projet de recherche Optimum Brethomé [2018], mené en collaboration avec Transilien, propose de construire une grille horaire et d'estimer la demande tactique conjointement. La construction de la grille horaire avec estimation de la demande tactique est ce que Robenek et al. [2016] appellent *Passenger Centric Train Timetabling Problem* (PCTTP). Par exemple, Wang et al. [2018] optimisent les intervalles entre métros en fonction de la demande afin de supprimer les situations de congestion. Parbo et al. [2016] notent que l'optimisation d'indicateurs centrés voyageurs peut toutefois pénaliser l'opérateur.

Des indicateurs adaptés aux besoins des voyageurs. Historiquement, la performance est mesurée par le gestionnaire d'infrastructure en calculant le retard du train par rapport à son sillon tracé pour chaque point remarquable (PR). La *régularité* est le pourcentage de trains avec moins de 5 minutes de retard. Le but de SNCF-Réseau est donc de tracer une grille horaire la plus robuste possible vis à vis de cet indicateur. Le problème est que cet indicateur ne prend pas en compte le volume de voyageurs impacté par les retards. Pour résoudre ce problème, IdFM fixe un objectif contractuel fondé sur la *ponctualité des voyageurs* qui est le pourcentage de voyageurs qui arrivent à leur destination avec moins de 5 min de retard par rapport à leur heure d'arrivée théorique. Cet indicateur est mieux adapté au Mass Transit ouvert car il correspond au ressenti des voyageurs. Par ailleurs, le nombre de montées, descentes et le taux d'occupation ne sont pas suffisamment bien connus au moment de la construction de la grille horaire. Il semblerait que Transilien et IdFM tracent des horaires afin que les taux d'occupation ne dépassent pas 80 % de la capacité assise dans la mesure du

possible. Cependant, ni Transilien, ni IdFM ne sont en mesure de donner une estimation fiable de ces taux d'occupation à la phase tactique. Pourtant, l'indicateur du *niveau de confort*, défini comme un niveau de taux d'occupation à ne pas dépasser, devient un enjeu de plus en plus important pour les voyageurs notamment après l'épidémie de COVID-19. L'estimation de la demande à la phase tactique passe généralement par l'allocation de flux de voyageurs issus de matrices origine-destination grâce à un algorithme de choix d'itinéraire plus ou moins sophistiqué.

Les matrices origine-destination. Pour concevoir une grille horaire en fonction de la demande, il est nécessaire d'avoir une connaissance de la demande indépendamment de l'offre de transport. Les matrices origine-destination à l'échelle du trajet (un mode) sont agrégées par heure pour un type de jour. Pour relier cette demande de trajets à des horaires de trains, il faut ajouter un modèle de choix d'itinéraire. C'est d'ailleurs ce que proposent Noursalehi et al. [2021] pour prédire les taux d'occupation à partir des données de télébillettique.

Temps de stationnement en phase tactique. Nous n'avons pas connaissance de documents internes à Transilien décrivant précisément comment calculer des temps de stationnement en fonction d'une demande estimée à la phase tactique. Il n'y a d'ailleurs aucun document conseillant de lier temps de stationnement et demande, à toutes les étapes de la planification ferroviaire. Toutefois, cette idée de lier demande et dimensionnement des temps de stationnement a été explorée par D'Acierno et al. [2017] qui combinent deux simulateurs, l'un pour la demande, l'autre pour l'offre. Le fait de coupler les deux modèles permet de faire varier conjointement les temps de stationnement et la demande afin de tester la robustesse d'un système de type métro. Buchmüller et al. [2008], Cornet et al. [2019] proposent des modèles statistiques pour les temps de stationnement prenant en compte un certain niveau de demande sans pour autant préciser comment ils obtiennent ce niveau de demande lors de la phase tactique. Pour notre part, nous nous concentrons sur la phase opérationnelle en couplant un modèle de modélisation des temps de stationnement et un modèle de prévisions des flux de voyageurs à court terme. Toutefois, la modélisation des temps de stationnement que nous proposons s'adapterait à la phase tactique si les flux de voyageurs étaient connus.

Mesures de la demande

Dans la Table 2.2, il y a trois grandes sources de données pour l'analyse post-opérationnelle : les *enquêtes* et les données automatiques qui regroupent les données de validation, appelées *automatic fare counting* (AFC), et les données de comptage, appelées *automatic passenger counting* (APC).

Les enquêtes. Il y a deux types d'enquête dans la Table 2.2. Les enquêtes de comptage mesurent le nombre de montées et de descentes par type de jours pour chaque arrêt s de chaque train k. Les enquêtes origines-destinations mesurent les caractéristiques socio-démographiques et les habitudes de déplacement des voyageurs sur les lignes de Transilien. Ces enquêtes sont réalisées tous les 4 ans par un institut de sondage externe. Elles sont très fiables et très matures. Transilien les utilise pour construire et analyser la performance de ses plans de transport. Les enquêtes OD mesurent des variables difficilement accessibles autrement mais sont incapables de capter les changements brutaux de la demande. Les méthodes d'enquêtes ont deux grandes limites : leur échantillonnage est réduit (3 jours tous les 4 ans) ; leur coût est rédhibitoire (plusieurs millions d'euros par enquête). Ces limites posent problème pour la re-construction de grilles horaires après une perturbation de longue durée comme l'épidémie de COVID-19. Il est donc pertinent que Transilien se tourne vers l'utilisation plus systématique des données automatiques pour mesurer la demande.

TABLE 2.2 – Présentation des données de comptage à Transilien. En rouge, les données utilisées dans cette thèse.

Informations générales				Informations techniques				Qualité des données			
	Nom	Phase de la planification associée	Périmètre	Technologie	Lieu d'installation	Quantité	Unité	Précision	Disponibilité	Niveau de fiabilité	Niveau de maturité
Enquêtes	Enquête de comptage (volume)	tactique et post-opérationnelle	une ligne de Transilien	enquêtes manuelles	dans les trains	nombre de montées et descentes	voyageurs	arrêt (k, s)	tous les 4 ans	élevé	élevé
	Enquête origine-destination (OD)	stratégique et tactique		enquêtes manuelles	sur les quais	habitudes de déplacement	x	individuel	tous les 4 ans	élevé	élevé
AFC	Validations (volume)	tactique et post-opérationnelle	gares Transilien	cartes à puce et billets magnétiques	entrée/sortie de gare	nombre d'entrées et sorties	voyageurs	individuel, 1-15-60 min	+2 J	élevé	élevé
	Validations (OD)	stratégique et tactique	réseau Île-de-France	cartes à puce	entrée/sortie de gare	matrices OD	voyageurs	15-30-60 min	+ 30 J	élevé	faible
	CAVE	tactique et opérationnelle	matériel compteur	capteurs infra-rouges/vidéos	portes	nombre de montées et de descentes	voyageurs	zones par arrêt (k, s)	temps réel	élevé	élevé
APC	Vidéo à quai	opérationnelle	x	caméras	à quai	affluence	affluence faible, moyenne et forte	zones par arrêt (k, s)	temps réel	faible	moyen
	Vidéo à bord	opérationnelle	x	caméras	à bord	affluence		zones par arrêt (k, s)	temps réel	moyen	faible
	Traces GPS (OD)	stratégique	région Île-de-France	géo-localisation du téléphone	SNCF Connect	matrices origine-destination	trajectoires de voyageurs	bout en bout	+6 min	moyen	faible
	Pesée	tactique et opérationnelle	matériel compteur	capteurs de pression	bogies	charge	kilogrammes	zones par arrêt (k, s)	ponctuelle	élevé	faible
	Traces GPS (volume)	opérationnelle	réseau Transilien	géo-localisation du téléphone	SNCF Connect	charge à bord	voyageurs	train	+6 min	faible	faible

Les validations. Dans la Table 2.2, nous présentons les données de validation sous forme de volumes et de matrices OD. Le volume de validations, *i.e.* de titres télébillettiques [18] et magnétiques, permet de suivre l'évolution de la fréquentation sur le réseau *a posteriori* (2 jours après). Elles ont été largement utilisées pour suivre la reprise après l'épidémie de COVID-19. Elles contiennent le volume d'entrées et de sorties validées [19] à chaque gare par plage temporelle (1-15-60 min). Le volume de titres permet de partager les recettes entre les différents opérateurs, il doit être fiable et mature. Les validations permettent de reconstruire des matrices origine-destination à partir des données de télé-billettiques Navigo. La stratégie de poursuite d'utilisateur présentée dans Trépanier et al. [2007] s'applique au réseau de transport d'Île-de-France avec quelques adaptations à cause de la validation non obligatoire en sortie de gare. Le système d'information des données de validation (SIDV) d'IdFM s'occupe de la reconstruction des matrices OD. Ces matrices OD reconstruites ont l'avantage d'être peu coûteuses et beaucoup plus précises que les données d'enquêtes. En effet, elles sont définies pour tous les jours de l'année à un pas de temps de 15, 30 ou 60 min. Les données d'OD sont cruciales pour la phase tactique de construction de la grille horaire. Pour en savoir plus sur les données de validation et leurs usages, nous nous référons à l'état de l'art de Pelletier et al. [2011]. Dans cette thèse, nous nous intéressons aux situations où les voyageurs sont déjà montés dans un train. Les données de validation ne sont pas au cœur de cette thèse mais elles sont très riches et ouvrent beaucoup d'opportunités. Par exemple, elles nourrissent un projet de recherche, piloté par le Lab' Mass Transit, sur l'estimation de charge à bord là où il n'y a pas de données de comptage automatiques à bord des trains.

Les comptages automatiques. Les données de comptage automatique, automatic passenger counting (APC), constituent une famille de systèmes permettant de mesurer, dans certains cas le nombre de montées et de descentes, dans d'autres cas le taux d'occupation ou l'affluence dans les trains. Pour un état de l'art mondial, voir [Darsena et al., 2020]. Nous décrivons ici et dans la Table 2.2 les différents systèmes APC qui sont en cours d'expérimentation ou utilisés à Transilien. Les systèmes expérimentaux utilisent principalement des caméras à bord ou à quai pour mesurer la densité de voyageurs par zone du train. En parallèle des expérimentations sur la vidéo, Transilien s'intéresse aussi aux traces GPS laissées par les utilisateurs de l'application *SNCF Connect*. Ces données ne sont pas fiables car le volume de voyageurs dépend du nombre d'utilisateurs de l'application SNCF Connect ayant autorisé la géo-localisation. De plus, elles ont une résolution spatio-temporelle très insuffisante pour permettre une estimation fiable et précise de la charge à bord des trains. Cependant les traces GPS des téléphones portables sont intéressantes car elles sont multimodales et permettent de mesurer des trajectoires de porte à porte d'un déplacement. Elles permettent de mesurer concrètement les choix de modes au-delà des affirmations déclaratives des enquêtes origine-destination. Les données automatiques de demande, comme nous le verrons dans le reste de la thèse, permettent souvent d'objectiver les comportements des voyageurs. Parmi les données APC, celles qui sont au cœur de

18. Smart card data en anglais
19. Pour les gares dont les équipements de validation le permettent.

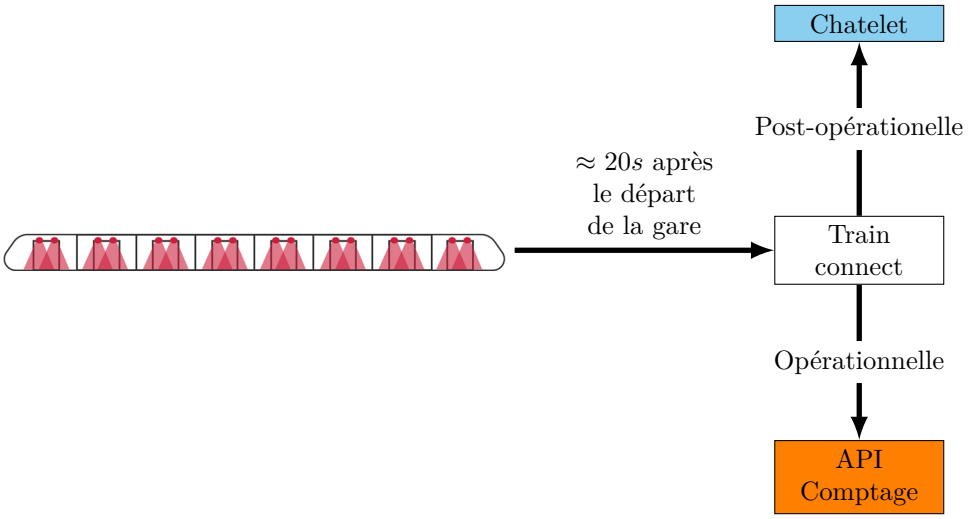


FIGURE 2.14 – Chaîne d'information simplifiée des données CAVE à Transilien.

la thèse sont les données de comptage automatique voyageur embarqué (CAVE) que nous décrivons en détail ci-dessou.

Données CAVE. Le système de *comptage automatique voyageurs embarqué* (CAVE), déjà présenté pour la mesure des temps de stationnement, est avant tout conçu pour mesurer le nombre de montées et descentes pour chaque arrêt et chaque porte des rames dites « compteuses ». Ces comptages sont issues de capteurs infra-rouges passifs [20] ou vidéo installés au-dessus des portes. Les rames équipées sont les rames Z50000/NAT (lignes H, K, L, J, P et E), Z57000/Regio 2N (lignes N et R) et prochainement le RER NG (lignes de RER D et E) et le MI20 (ligne de RER B). Elles représentent environ 50 % du parc Transilien en 2022. D'ici 2030, les rames équipées de CAVE représenteront plus de 70 % du parc de Transilien dont les trois lignes de RER B, D et E. Les données CAVE sont d'une suffisamment bonne qualité pour remplacer les enquêtes de comptage pour les lignes 100 % équipées. Les données CAVE ont l'avantage d'être précises et facilement accessibles en temps réel grâce à une API *Comptage*, voir Figure 2.14.

La chaîne d'information de la Figure 2.14 s'articule autour de quatre briques : le train transmet au sol un fichier de comptage 20 s après le départ ; le service *Train Connect* met sous un même format tous les fichiers ; la société *Acorel* actualise les données CAVE dans CHATELET tous les deux jours (délai nécessaire pour recevoir et redresser [21] les données) ; en parallèle, Train Connect alimente en temps réel l'API Comptage. L'API Comptage a été développée durant l'année 2021. Nous utilisons les données CHATELET pour les Chapitre 3-4 et nous utilisons les données brutes de l'API Comptagedans le Chapitre 5.

20. Les capteurs infra-rouges passifs utilisent la réflexion du faisceau lumineux pour détecter la présence d'un voyageur. Il y a deux faisceaux lumineux permettant de déterminer le sens de passage d'un voyageur, voir Pinna and Dalla Chiara [2010] pour plus de détails.

21. Les redressements sont le filtre des courses aberrantes et égaliser le nombre de montées et de descentes pour chaque course.

Évaluation des grilles horaires sous le prisme des comptages automatiques

L'analyse post-opérationnelle de l'offre et de la demande grâce aux comptages automatiques se développe. À Transilien, la pandémie de COVID-19 a accéléré le besoin de suivre finement la demande et son interaction avec les nouveaux plans de transport successivement déployés. Pour faire face à cette demande de suivi, Transilien a mis en place un suivi des volumes de validation. Nous pensions que les lignes équipées de CAVE pouvaient faire mieux, ainsi nous avons développé au DataLab' l'outil *ActuCharge* qui permet de visualiser, deux heures après la circulation, d'un train son taux d'occupation. Cette dynamique de rapprochement de l'analyse de l'offre et de la demande n'est pas originale. Barry and Cardl [2014] représentent les retards à l'aide d'un graphe de Marey et le nombre d'entrées sur le réseau Massachusetts Bay Transit Authority (MBTA) durant le mois de février 2014 à l'aide d'une carte de chaleur. Ces visualisations permettent de confirmer que le maintien des intervalles entre les métros est d'autant plus difficile que le nombre de voyageurs est grand. Le logiciel BusViz [Anwar et al., 2016] illustre que cette dynamique de visualisation de l'affluence est aussi suivie par les réseaux de bus notamment sous forme de graphes de Marey des taux d'occupation. Ces graphes révèlent précisément les points de congestion. Cette congestion est souvent peu ou mal connue par les voyageurs. En effet, un voyageur connait bien le réseau : ses gares et ses dessertes. Il lui est plus difficile d'appréhender le phénomène de la congestion qui est dynamique et non planifiée. Selon Burch et al. [2020], la visualisation de la congestion serait une information importante lors du choix d'itinéraire. Cette information vient s'ajouter aux représentations classiques du temps de trajet, Zeng et al. [2014]. Une bonne visualisation de la demande permet d'évaluer la qualité de la grille horaire et de donner aux voyageurs des informations pour leur choix d'itinéraire.

Adapter la demande à l'offre. L'analyse post-opérationnelle de la demande grâce aux comptages automatiques permet d'identifier les zones et les périodes de forte congestion. Il y a deux manières de réduire la congestion : augmenter l'offre ou réduire la demande, la première solution n'étant pas toujours souhaitable ou possible. L'idée de déplacer ou réduire la demande a été pensée pour l'exploitation des routes, cela s'appelle le *travel demand management* (TDM). Nous développons deux leviers du TDM : économique par la tarification, social par l'incitation. Le levier économique s'appuie sur l'élasticité prix de la demande [Litman, 2004]. Ce levier ne pourrait être activé par Transilien seul car la tarification des transports en commun est une des prérogatives d'IdFM. Par ailleurs, IdFM propose aujourd'hui un abonnement unique qui est antinomique avec l'idée de prix variables. Par ailleurs, une telle tarification nécessiterait de gros investissements d'adaptation des contrôleurs automatiques de billets (CAB). Le levier social est déjà activé par Transilien, il consiste à décaler les heures d'arrivées des grands pôles d'activité. Un exemple réussi de TDM social est celui de Rennes où le décalage des heures de début des cours de l'université Rennes 2 de 15 minutes a fortement diminué la congestion de la ligne [Briand et al., 2017]. Toutefois, pour de petites situations de congestion, la gestion opérationnelle des flux de voyageurs est une solution efficace.

2.3.3 Demande opérationnelle : des flux de voyageurs

Mon travail de thèse exploite la relative maturité des données APC de Transilien due au renouvellement massif des rames en Île-de-France. Le système de comptage en temps réel, par porte est intégré à une chaîne d'information industrielle de bonne qualité, précise et facilement accessible est une chance.

Affluence en temps réel pour guider les flux voyageurs

Transilien a résolument décidé d'inscrire la gestion opérationnelle des flux de voyageurs au cœur de sa stratégie de la gestion de la congestion à bord. L'entreprise a lancé un projet ambitieux d'information des voyageurs autour de l'affluence à bord, appelé *IV affluence*. La stratégie de Transilien et d'IdFM est d'informer en priorité les voyageurs sur l'affluence à bord.

L'information voyageurs d'affluence en temps réel. Le moteur du développement de l'API Comptage est le programme IV affluence[22]. Le programme IV affluence s'articule autour de trois grands axes.

1. Intégrer la notion de confort dans le choix d'itinéraire des voyageurs. Ce service est accessible sur `https://www.transilien.com` au travers de trois niveaux d'affluence qui s'affichent au moment de la recherche d'itinéraire, voir Figure 2.15. L'affluence est dite actuelle si elle est obtenue à partir des données CAVE temps réel. Elle est dite habituelle si elle provient des données d'enquêtes.
2. Lisser les échanges à l'interface quai-train en informant les voyageurs sur l'affluence par zone du train, voir la Figure 2.16. Ce service est en cours de développement sur les lignes H et N d'ici à la fin de l'année 2022. Il consistera à afficher sur les écrans en gare différents niveaux d'affluence par zone du train. Ce service est très largement inspiré du PoC Hector, voir Section 2.4.3, ainsi que d'autres initiatives étrangères dont celle proposée par Zhang et al. [2017].
3. Conseiller les voyageurs sur des itinéraires alternatifs en cas de congestion d'une partie du réseau, une sorte d'aide à la navigations dans les transports en commun en fonction de l'affluence.

Ces différents services aux voyageurs doivent permettre aux voyageurs d'éviter les zones et périodes de congestion. Ils posent les bases pour améliorer l'adéquation des flux de trains et des flux de voyageurs en opérationnel que nous développons dans la Section 2.4. Il reste néanmoins un important travail à mener pour faire en sorte que la demande en opérationnel soit aussi à disposition des agents. Une étape vers

22. Pour Transilien, l'affluence est la discrétisation en 4 niveaux ($[0, 25\%)$, $[25\%, 50\%)$, $[50\%, 75\%)$, $[75\%, 200\%)$) du taux d'occupation à l'échelle du train ou de la zone

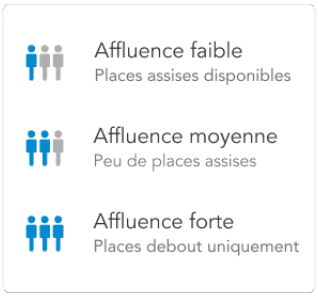


FIGURE 2.15 – Affluence à l'échelle du train affichée sur https://www.transilien.com.

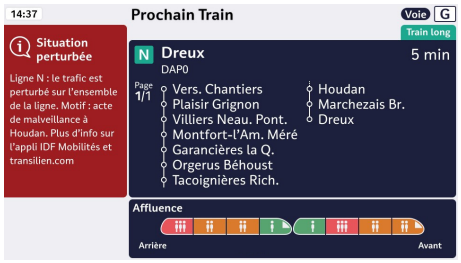


FIGURE 2.16 – Affluence à l'échelle de la zone bientôt affichée sur les écrans à quai.

cet objectif sera l'intégration de l'affluence dans l'outil *Solférino* utilisé par les COT pour la re-planification en temps réel.

Limites de la mesure des flux de voyageurs en temps réel

Les données CAVE sont une réelle opportunité pour Transilien et il ne sera bientôt plus question de faire rouler des trains sans connaitre leur taux d'occupation. Plusieurs défis doivent néanmoins être relevés concernant la fiabilité des données CAVE. Le premier concerne le taux de couverture du système de comptage CAVE. Une ligne avec 100 % de rames équipées CAVE ne remonte des flux de voyageurs que pour 90 % des arrêts en temps réel. La vitesse d'envoi au sol des fichiers de comptage est parfois largement plus longue que les 25 secondes évoquées sur la Figure 2.10. Ces trous de données récurrents seront identifiés plus rapidement grâce à une chaîne de supervision en cours de déploiement. Le second est qu'aujourd'hui, les lignes équipées de CAVE représentent environ 50 % du trafic, aucune ligne de RER n'en est équipée. Transilien cherche des alternatives pour ces situations. Une piste de recherche est l'utilisation des matrices OD du SIDV allouées aux trains. Le troisième est d'obtenir une mesure fiable de l'occupation par zone du train alors que les rames sont communicantes [23] et que les flux de voyageurs sont mesurés au niveau des portes. Ces déplacements sont à prendre en compte dans le calcul de la charge à bord, c'est le sujet du Chapitre 5 de cette thèse.

23. Une rame communicante est une rame qui permet aux voyageurs de se déplacer librement d'une voiture à l'autre. On peut les appeler aussi rames BOA ou traversantes. La conséquence est double : les voyageurs s'étalent sur toute la longueur du train ce qui est plus confortable pour eux ; mais ils ne restent pas forcément à proximité de la porte par laquelle ils sont montés, donc on ne sait plus quelles sont les portes les plus chargées à l'arrêt suivant.

Prévision à court terme des flux de voyageurs

Les services d'information d'affluence en temps réel pour les voyageurs comme pour les agents doivent anticiper l'évolution de l'affluence à court terme pour être performants. En effet, l'objectif est d'informer le voyageur sur l'affluence qu'il expérimentera au moment où il montera dans le train et non comment elle est actuellement. L'affluence en temps réel repose donc sur deux briques : une acquisition à l'instant t des flux de voyageurs et une brique de prévision à court terme que nous aborderons dans le Chapitre 4.

Prévision et boucle d'information. La prévision et l'information des voyageurs destinées à changer les comportements de mobilités ne sont pas toujours efficaces. Nous prédisons ce qui va se passer, puis les voyageurs choisissent un itinéraire en fonction de cette information prédite. Si au moment de la prévision, nous n'anticipons pas comment l'information sera suivie par les voyageurs cela peut entrainer des incohérences dans les prévisions. C'est ce que soulignent Koutsopoulos et al. [2019] dans leur article sur l'effet boucle de l'information. Ils incitent à anticiper à quel point les utilisateurs de calculateur d'itinéraire vont suivre le conseil pour éviter les reports de congestion.

Prévision en situation normale et perturbée. Nous proposons une méthode de prévision à court terme dans le Chapitre 4 pour des situations sans perturbation. L'enjeu de prédire l'affluence en cas de situations très perturbées reste donc ouvert. Pour résoudre ce problème de prévision en situation perturbée, il est probable que l'utilisation de données à l'échelle du train ne soit pas suffisante. Il serait pertinent d'utiliser une approche similaire à celle utilisée lors de la re-planification en temps réel en exploitant des matrices OD agrégées puis en allouant la demande aux nouveaux itinéraires.

Une nécessité de prévoir la demande pour les agents. La re-planification en temps réel est facilitée par l'acquisition en temps réel de données de demande voyageurs. L'action des conducteurs, des gestionnaires de flux ainsi que des agents gare serait plus adaptée si ces personnes avaient à leur disposition une évolution des flux de voyageurs dans les trains et dans les gares. Une meilleure anticipation des phénomènes de retard et d'affluence permettrait aux agents d'être plus efficaces dans leur métier de régulation.

Résumé de la section demande voyageurs. La demande en phase stratégique permet de dimensionner les infrastructures. La demande en phase tactique permet d'aider à la construction de la grille horaire. Les données des matrices OD fournies par le SIDV d'IdFM sont une richesse sous exploitée à Transilien pour imaginer en phase tactique des solutions réduisant la congestion dans les trains. L'adaptation de l'offre à la demande n'est pas toujours possible, le travel demand management

(TDM) permettrait d'améliorer l'adaptation de la demande à l'offre. Les données de cette thèse sont le nombre de montées et de descentes par porte à chaque arrêt qui sont mesurés par le système CAVE. Ces données sont accessibles en temps réel grâce à une API Comptage performante que Transilien a développée. Aujourd'hui, l'exploitation de ces données de demande en opérationnel est surtout dirigée vers les voyageurs. Toutefois, il est clair qu'une meilleure prise en compte de la demande, dans la mesure du possible, par Transilien permettrait de fluidifier l'exploitation du Mass Transit ouvert. Dans cette thèse, nous proposons d'améliorer l'exploitation en prenant davantage en compte les flux de voyageurs dans la gestion des flux de trains.

2.4 Synchronisation des flux de trains et des flux de voyageurs

Les trains automatiques ainsi que les données de comptage voyageurs acquises en temps réel sont des opportunités pour une meilleure synchronisation des flux de trains et de voyageurs. Dans cette thèse, nous nous intéressons à la modélisation et la prévision en phase opérationnelle de l'offre et de la demande à l'interface quai-train. Loin des algorithmes de re-planification en temps réel tout en un, nous construisons des briques utiles à ces algorithmes et à l'information des voyageurs. Ce sujet se décline en trois sujets qui forment les trois chapitres de cette thèse que nous tâchons d'inscrire dans le contexte scientifique et industriel de Transilien.

Nous montrons dans la Section 2.4.1 que la modélisation des temps de stationnement à partir des données post-opérationnelles doit être rigoureuse en distinguant la part des flux de voyageurs et des variables d'exploitation ferroviaire (temps de stationnement théorique, retard à l'arrivée, etc.). Pour passer d'une modélisation descriptives des temps de stationnement à un modèle opérationnel, nous abordons dans la Section 2.4.2 les enjeux de prévision à court terme des flux de voyageurs et des variables d'exploitation ferroviaire. Nous montrons dans le Chapitre 3 que les échanges à la porte critique (la porte avec le plus grand nombre de montées et descentes) sont importants dans le calcul des temps de stationnement. Nous montrons dans la Section 2.4.3 comment agir sur la répartition des voyageurs au moment de l'échange à l'interface quai-train à l'aide de canaux d'information de l'information des voyageurs.

2.4.1 Temps de stationnement : modèles et calcul de marge

Dans cette partie, nous inscrivons d'abord les résultats du Chapitre 3 dans le contexte industriel et scientifique de Transilien, pour ensuite proposer plusieurs applications, non développées dans la thèse, de la modélisation des temps de stationnement des trains en retard pour le calcul de marge.

Modélisation des temps de stationnement (Chapitre 3)

La modélisation des temps de stationnement à partir des données de flux de voyageurs et d'exploitation ferroviaire post-opérationnelles n'est pas un réflexe à Transilien. Nous avons indiqué dans la Section 2.2.1 que les méthodes utilisées pour l'analyse post-opérationnelle des temps de stationnement à Transilien ne prennent pas en compte des flux de voyageurs. À l'inverse Kuipers et al. [2021b] dans leur article de revue négligent la contrainte des heures de départ théoriques sur la longueur des temps de stationnement. Buchmüller et al. [2008], Pedersen et al. [2018], Medeossi and Nash [2020] remarquent qu'il est préférable de se concentrer sur les trains en retards, *i.e.* qui arrivent après leur heure de départ théorique, pour analyser les données post-opérationnelles de temps de stationnement. Dans ce travail de modélisation, nous tâchons de séparer l'effet du respect de la grille horaire, d'une part, et des flux de voyageurs, d'autre part. Pour cela, nous avons à disposition à la fois une mesure fine des temps de stationnement (ATESS) et des flux de voyageurs (CAVE). Nos contributions à fort impact industriel sont triples :

1. Les temps de stationnement des trains en avance ne sont pas contraints par les flux de voyageurs, ils sont contraints par l'heure de départ théorique. Ainsi, en première approche, pour analyser les temps de stationnement trop longs, il est souhaitable d'éliminer les trains en avance qui ont une faible chance d'être impactés par les flux de voyageurs ;
2. Les temps de stationnement des trains en retard sont impactés par les flux de voyageurs, ils peuvent être considérés comme une approximation de la notion d'un temps de stationnement adapté aux flux de voyageurs ;
3. Les flux de voyageurs à la porte critique, c'est-à-dire à la porte avec le plus d'échange voyageurs, est une variable importante dans la détermination les temps de stationnement.

Ce travail sur la modélisation statistique des temps de stationnement a fait progresser Transilien dans deux directions. D'une part, vers le croisement de l'analyse de la régularité et de celle des flux de voyageurs. D'autre part, vers l'analyse plus fine des flux de voyageurs pour identifier les zones critiques le long du quai à l'interface quai-train.

Calcul des marges des temps de stationnement

Transilien fait circuler des trains qui s'arrêtent toutes les 3 à 5 minutes en moyenne. Entre une origine et une destination, il y a entre 15 et 30 % du temps de trajet qui est passé à l'arrêt, voir Figure 2.17. Les temps de stationnement théoriques précis à la déca-seconde sont historiquement dimensionnés en phase tactique par une méthode itérative qui a tendance à rallonger les temps de stationnement théoriques d'année en année. Dans cette thèse, nous proposons deux stratégies pour distinguer les temps d'échange des marges lors des temps de stationnement, ce qui permet d'objectiver la part des voyageurs dans les retards. Une telle méthode permettrait de tracer les temps de stationnement en suivant la même philosophie que pour les temps de

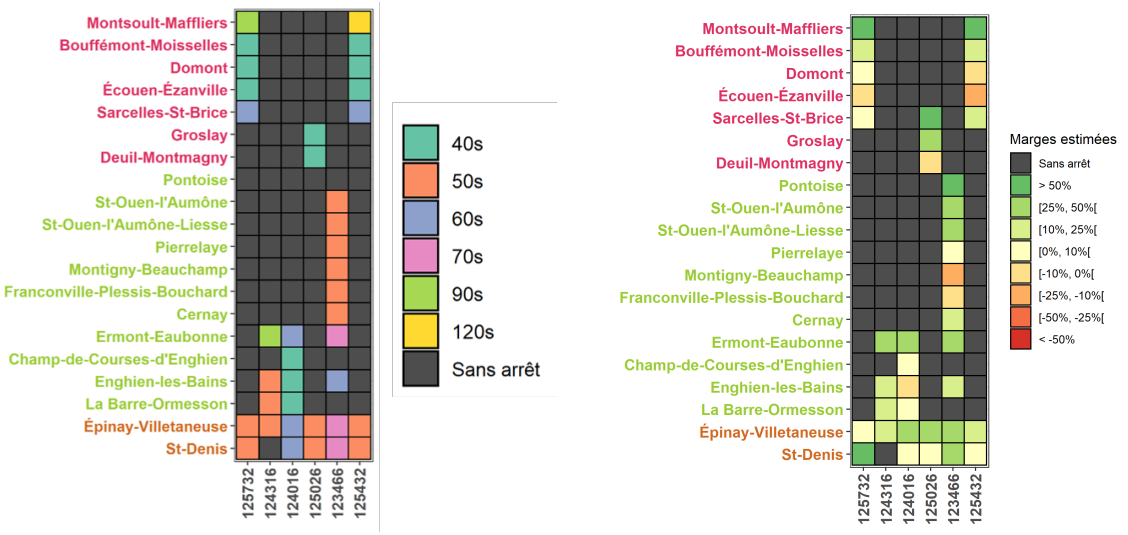


FIGURE 2.17

FIGURE 2.18 – Exemple de calcul des marges pour 5 trains de la ligne H allant vers Paris en heures de pointe du matin de janvier à septembre 2019 pour les jours ouvrés. À gauche, les temps de stationnement théoriques. À droite, les marges estimées par la méthode des trains en retard.

parcours : un temps de stationnement tendu auquel on ajouterait une marge. Elle permettrait d'objectiver l'affirmation que 1 % de fréquentation en plus serait égal à 0.5 % de régularité en moins.

Calcul des marges à partir des trains en retard. Nous montrons dans le Chapitre 3 que les temps de stationnement des trains en retard sont expliqués principalement par les flux de voyageurs. Une des raisons est que les conducteurs arrivant en retard doivent repartir dès que possible. Nous faisons ici l'hypothèse que la marge est négligeable pour les temps de stationnement des trains en retard. Dans ce cas, le temps de stationnement est composé uniquement d'un temps d'échange, dont les variations s'expliquent par les flux de voyageurs, et d'un temps technique, déterministe et connu. Dans le cadre d'un projet avec l'École des Ponts, nous avons testé une méthode en deux temps permettant de calculer les marges effectivement réalisées sur tous les temps de stationnement théoriques, à partir des trains en retard. Cette étude consiste en une analyse post-opérationnelle des données CAVE et de circulation de la ligne H en heure de pointe du matin pour les trains allant vers Paris de janvier à septembre 2019. Nous utilisons d'abord seulement les trains en retard pour estimer un modèle linéaire simple des temps de stationnement sans marge en fonction des flux de voyageurs : le nombre de montées, de descentes et le taux d'occupation. Ensuite, nous utilisons l'ensemble des trains qui ont circulé, pour estimer des marges de la manière suivante : d'abord, nous estimons pour tous les trains des temps de stationnement sans marge à partir des flux de voyageurs, notés $\widehat{y}^{\text{obs}}$, puis nous calculons la différence relative entre les temps de stationnement sans marge estimés et les temps de stationnement théoriques de la Figure 2.18 à gauche, notés y^{theo}. :

$$\frac{y^{\text{theo}} - \widehat{y}^{\text{obs}}}{y^{\text{theo}}}.$$

Nous calculons ces marges sur la Figure 2.18 de droite pour 5 trains en heures de pointe du matin vers Paris. Il apparait que les temps de stationnement théoriques pour la plupart des gares de la ligne H ont des marges relatives allant de 10 à 50 % du temps de stationnement théorique. Ce chiffre est bien supérieur aux 5 à 10 % recommandé par l'UITP. Nous remarquons que les temps de stationnement théoriques sur la Figure 2.18 à gauche, sont particulièrement longs pour la gare de Montsoult-Maffliers qui est une gare située à une convergence entre deux branches de la ligne H. Ainsi, d'autres raisons que les flux de voyageurs peuvent expliquer la longueur des temps de stationnement théoriques. Nous avons appliqué la même méthode aux arrêts de la ligne L qui ont des marges relatives négatives notamment en hyper-pointe du matin (8h-9h), ce qui signifie que les temps de stationnement théoriques sont trop courts par rapport aux flux de voyageurs. Cette première stratégie de calcul des marges est pratique car elle permet de calculer simplement des marges sur tout le périmètre équipé de CAVE. Cependant, elle repose sur une hypothèse forte qui est que tous les conducteurs respectent le geste métier de repartir dès que possible lorsqu'ils sont en retard. C'est pourquoi, nous présentons une autre méthode plus fine mais moins générale ci-dessous.

Calcul des marges à partir des données plus précises des Regio 2N. Nous proposons une deuxième méthode dans l'Annexe D qui consiste à exploiter les données quasiment continues des CAVE, accessibles exceptionnellement sur les rames Regio 2N. Ces données permettent de mesurer presque toutes les secondes le nombre de montées et de descentes pour chaque porte. Ces mesures des flux de voyageurs par porte amènent à poser la question de la définition du temps d'échange. Quand est-ce qu'il commence et quand est-ce qu'il se termine. Nous définissons pour cela la notion de cluster de voyageurs comme Wiggenraad [2001]. Un cluster est un ensemble d'au moins deux voyageurs séparés au plus par deux secondes. Le temps d'échange pour une porte est donc le temps séparant le début du premier cluster et la fin du dernier cluster. Le temps d'échange à l'échelle du train est le maximum du temps d'échange par porte, *i.e.* le temps d'échange à la porte critique. Ces temps d'échange permettent de définir une marge pour chaque porte. Cette méthode a permis de montrer que les trains en retard ont des marges non nulles à la porte critique contrairement à ce que voudrait l'hypothèse selon laquelle les conducteurs repartiraient dès que possible. Le volume de marge est cependant bien décroissant en fonction de la déviation à l'heure d'arrivée. L'étude des marges à l'aide de ces comptages continus permet d'observer que sur les lignes R et N où circulent les Regio 2N, entre 30 et 60 % du temps d'échange planifié correspond à de la marge au niveau des temps de stationnement. L'utilisation des comptages continus est prometteuse pour Transilien cependant leur collecte n'est pas automatisée, contrairement au CAVE qui a une chaîne de transmission dédiée, voir Figure 2.14. La méthode exploitant les trains en retard est une approximation qui est cependant facilement généralisable. Elle permettra de questionner plus largement la marge à l'arrêt sur les lignes équipées de CAVE.

2.4.2 Prévision à court terme (Chapitre 4)

La modélisation des temps de stationnement est utile pour le dimensionnement des temps de stationnement à la phase tactique mais aussi pour anticiper la bonne tenue des temps de stationnement théoriques en opérationnel. Transilien transporte des flux de voyageurs conséquents, il est donc intéressant pour cet opérateur d'anticiper des dépassements de temps de stationnement liés à un nombre de voyageurs particulièrement important. Pour les trains automatiques ou pour les trains en retard, anticiper le retard susceptible d'être rattrapé par le train ou le conducteur en fonction des flux de voyageurs permettrait une meilleure gestion opérationnelle. Dans cette thèse, nous proposons de passer d'une modélisation des temps de stationnement dans le Chapitre 3 à une modélisation opérationnelle en prédisant à court terme les variables explicatives (le retard à l'arrivée et les flux de voyageurs). Pour prédire ces variables, nous avons travaillé dans deux directions. Le premier axe consiste à comprendre les modèles utilisés par Transilien, pour la prévision à court terme, détaillés dans le Chapitre 4. Ils fonctionnent grâce à l'actualisation de valeurs de référence futures à partir des déviations actuellement observées par rapport aux valeurs de référence. Le deuxième axe consiste à exploiter la structure particulière d'un réseau de transport. Un réseau de trains de banlieue est composé d'une infrastructure et d'un plan de transport rigide sur lesquels les événements sont ordonnés au niveau d'une gare et d'un train. Un train circule sur ce réseau en accumulant de l'information d'arrêt en arrêt. Ainsi, nous avons proposé dans le Chapitre 4, une méthode générale de prévision des différentes variables : flux de voyageurs ou exploitation ferroviaire. Ce travail nous a permis de mieux comprendre la structure du réseau et de la grille horaire pour modéliser les interactions entre les arrêts. La méthode de prévision proposée est meilleure que les modèles de Transilien et rivalise avec les performances d'une méthode à l'état de l'art pour la prévision des retards [Corman and Kecman, 2018]. Ce deuxième axe a l'ambition de construire un pont entre la littérature sur la prévision de retard et celle sur la prévision des flux de voyageurs. La prévision des flux de voyageurs étant trop souvent abordée sous l'angle des flux en gares ou des matrices origines-destinations [Toqué, 2019]. La prévision des flux de voyageurs et de la charge à bord à court et moyen terme s'inscrit dans la stratégie de Transilien de proposer des services aux voyageurs sur l'affluence à bord au moment de leur choix d'itinéraire ou avant de monter dans le train.

2.4.3 Hector et la modélisation des flux de voyageurs

La mauvaise répartition des voyageurs à l'interface quai-train est un enjeu pour Transilien et les voyageurs. Pour Transilien, une meilleure répartition à l'interface quai-train permettrait de réduire le temps nécessaire à l'échange des voyageurs et donc de pouvoir réduire les temps de stationnement théoriques. Pour les voyageurs, une meilleure répartition à bord est synonyme de plus de confort, ce d'autant que la congestion multiplie le temps de trajet perçu [Tirachini et al., 2013]. Tout d'abord, nous présentons Hector qui est un projet de recherche et développement

que nous avons développé pendant cette thèse. Nous montrons, ensuite comment la problématique des déplacements à bord des rames est vite apparue comme déterminante. Enfin, nous concluons avec l'analyse de la stratégie en sortie des voyageurs à l'interface quai-train au moment de descendre.

Hector et l'information voyageurs

Pourquoi ? L'épidémie de COVID-19 a entrainé 3 mois de confinements strict durant lesquels la mobilité des personnes était quasiment figée. Lors du premier dé-confinement, il fallait pouvoir garantir que les taux d'occupation dans les trains de banlieue ne dépassaient pas 30 % de la capacité totale. Pour piloter ces taux d'occupation, Transilien a amélioré son flux de données CAVE pour le rendre accessible en temps réel. Le Lab' Mass Transit par mon intermédiaire s'est saisi de cette avancée pour proposer un nouveau service aux voyageurs destiné à améliorer leur confort.

Comment ? Hector (Estimation de la Charge à bord en Temps réel pour l'Optimisation de la Répartition à quai sur la ligne H) permet d'informer les clients de la ligne H de l'affluence par zone du train. Il y avait un site web dédié `https://hector.transilien.sncf.fr/` où il suffisait de sélectionner une gare puis un sens de circulation (vers Paris ou autres directions) pour voir s'afficher l'affluence à bord par zone[24] des trois prochains trains comme sur la Figure 2.19. Ce service a été déployé pendant plus de 3 mois de mars à juin 2021 puis 6 mois supplémentaires, à la demande d'IdFM, de septembre 2021 à mars 2022. Il a touché plus de 7 000 voyageurs de la ligne H et a pu être déployé sur un écran en gare d'Épinay-Villetaneuse, voir Figure 2.19.

Et après. Nous avons fait le choix de développer Hector rapidement *i.e.* en dehors de l'écosystème des médias numériques de Transilien. Nous avons décidé, dès le début d'utiliser les données CAVE temps réel qui avaient vocation a être pérennisées, ce qui va permettre au service Hector d'être déployé sur tous les écrans à quai des gares des lignes H et N d'ici à la fin d'année 2022 selon la Figure 2.16.

Modélisation des déplacements à bord (Chapitre 5)

Hector a donné naissance à un sujet mathématiquement intéressant autour du déplacement des voyageurs à bord. Le défi scientifique est que les rames équipées de CAVE temps réel, qui mesurent le nombre de montées et de descentes par porte, sont des rames communicantes. C'est-à-dire que les voyageurs, après être montés, peuvent se déplacer librement à l'intérieur de la rame sans être à nouveau mesurés.

24. Une zone est un ensemble de deux voitures *i.e.* deux portes pour les NAT.

FIGURE 2.19 – Exemple d'information voyageur sur l'affluence issue de l'expérimentation Hector (Estimation de la Charge à bord en Temps réel pour l'Optimisation de la Répartition à quai sur la ligne H). À gauche, une incrustation sur un smart-phone. À droite, une photo d'un écran installé en gare d'Épinay-Villetaneuse de mars à juin 2021.

La question du déplacement des voyageurs dans la rame a amené une réponse rapide et naïve présentée dans l'Anexxe B, que nous avons développés dans le Chapitre 5. Une voie explorée que nous ne présentons ni dans le résumé, ni dans un chapitre de thèse est de croiser l'affluence d'Hector avec une mesure de masse à l'essieu pour les Regio 2N. Il s'avère que l'écart entre les deux mesures est négligeable au global mais que pour les affluences moyennes ou fortes, l'écart est entre 20 et 30 % avec déplacements et entre 30 et 40% sans déplacements. Ces résultats prouvent, d'une part, l'importance de prendre en compte les déplacements des voyageurs y compris pour des rames comme les Regio 2N à deux niveaux, mais d'autre part, que le modèle à l'échelle du trajet que nous avions proposé en 2021 peut encore être amélioré. Il serait nécessaire, si un jour le modèle d'Hector devait être industrialisé, de confirmer la pertinence des redressements soit grâce à la masse à l'essieu, soit grâce à des mesures manuelles.

Stratégies des voyageurs à la descente du train

L'efficacité d'Hector a été relative sur la meilleure répartition des voyageurs à l'interface quai-train à cause de l'accès à l'information difficile ainsi que des autres motivations des voyageurs au moment de se positionner sur le quai. Le confort durant le voyage est un des objectifs poursuivis par les voyageurs de Transilien. Cependant, il est aussi admis dans la littérature scientifique que les voyageurs peuvent être stratégiques en sortie, c'est-à-dire qu'ils peuvent chercher à minimiser leur temps de marche une fois qu'ils sont arrivés à destination. Kim et al. [2014] estime à plus de 50 % la part de ces voyageurs stratégiques en sortie. Il est vraisemblable que ces voyageurs seront moins enclins à se déplacer le long du quai en fonction de l'affluence à bord. Nous avons proposé une méthode dans l'Annexe C exploitant les données CAVE par porte, notamment le nombre de descentes par porte, pour calculer la part de voyageurs stratégiques en sortie pour

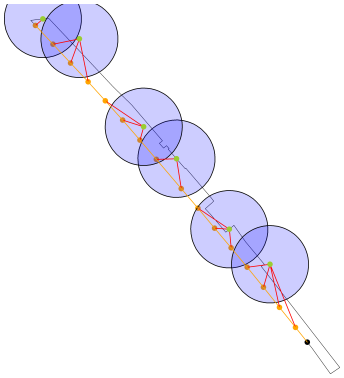


FIGURE 2.20 – Epinay-Villetaneuse

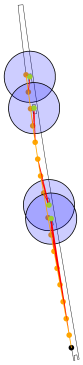


FIGURE 2.21 – Saint-Denis

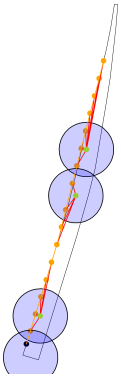


FIGURE 2.22 – Paris gare du Nord

FIGURE 2.23 – Graphe des quais avec un cercle de rayon de 20 mètres autour des sorties de quai pour trois gares de la ligne H. Nous affichons plusieurs éléments : • la position des sorties ; • cercle de 20 mètres de rayon ; • position de l'arrêt du train ; • positions des portes ; – distance pour chaque porte à la plus proche sortie.

6 gares de la ligne H. Pour cela, nous avons géo-localisé chaque porte du train pour chaque quai, les points orange sur la Figure 2.23. Parallèlement, nous avons géo-localisé les entrées/sorties de quai, en vert sur la Figure 2.23. Nous définissons pour chaque entrées/sorties de quai un cercle d'attractivité de rayon donné. L'ensemble des portes du train comprises dans ce rayon sont dites stratégiques pour un rayon donné. La conclusion est que la part de voyageurs stratégiques en sortie est variable suivant les gares mais qu'il y a en moyenne assez peu de voyageurs, moins de 30 % qui se placent à moins d'une porte des entrées/sorties de quai. Ce taux monte à 50 % lorsqu'on étend le seuil à plus ou moins deux portes. Ce travail a permis d'identifier des zones de quai sur-utilisées à cause de points chauds liés à des gares attractives.

Résumé de la synchronisation des flux de trains et des flux de voyageurs. La synchronisation des flux de voyageurs et des flux de trains est au cœur de cette thèse. Nous avons abordé ce problème en exploitant systématiquement les données de comptage automatique du nombre de montées et de descentes par porte. Ces données nous ont permis, tout d'abord, de se convaincre que les temps de stationnement réalisés à Transilien étaient premièrement déterminés par les déviations par rapport aux heures d'arrivée et non pas par les flux de voyageurs. Il a été crucial de montrer que pour les situations critiques, c'est-à-dire pour les trains en retard, c'est bien le nombre de montées et de descentes à la porte critique qui est déterminant. Ensuite, la modélisation des temps de stationnement n'étant pas une fin en soi, nous avons voulu rendre opérationnel ce modèle en ajoutant des prévisions des variables explicatives. La méthode générale développée pour ces prévisions s'avère être utile notamment pour prédire le retard ou l'affluence à bord des trains. Enfin, l'épidémie de COVID-19 a motivé un projet de recherche et développement au service des voyageurs pour qu'ils puissent être le plus confortablement installés possible. Ce service est un succès industriel car il devrait être déployé dans les mois à venir. C'est aussi un succès

scientifique car il a inspiré l'idée d'un chapitre de thèse sur la modélisation des déplacements des voyageurs à bord des trains et deux résumés pour des conférences en transport.

3 Dwell time modeling

We model trains dwell times based on a rich data set containing both railway operations and passenger flows variables, which is rare in the literature. Our models are either linear regressions or machine-learning methods like random forests. While railway operations variables remain key for the modeling of dwell time, we characterize the added value of passenger flows variables. Overall, they lead to a reduction of the global modeling error by about 0.5 s, with up to 5 s – 10 s improvements in challenging situations consisting of late arrivals or associated with high passenger volumes.

Contents

3.1 Introduction and literature review

We model dwell times for trains subject to a possibly dense timetable (up to 24 trains per hour during peak hours) in the greater Paris area (SNCF operator). We do so based on two sets of variables : railway operations and timetable (scheduled dwell time, deviation to scheduled arrival time, train length, etc.), on the one hand ; passenger flows (numbers of alighting and boarding passengers, occupancy rate), on the other hand. We consider two railway lines, one significantly more dense than the other, but with a common point : the vast majority of their trains is equipped with automatic passenger counting (APC) device at each door. We may therefore use the breakdown of alighting and boarding numbers by door, with a particular interest on the critical door.

Only few earlier references could use such a combination of variables based on railway operations and on passenger flows. Among these, Cornet et al. [2019] rely on similar data (same greater Paris area, same SNCF operator) and model some excess dwell time with respect to some minimum dwell time (see below), solely based on passenger flows and not using the available railway operations variables. Also, Palmqvist et al. [2020] model dwell times in a setting with a more flexible and less precise timetable (main line trains in Sweden), relying on passenger flows (alighting, boarding, and crowding factor) and on the deviation to scheduled arrival time ; their contribution has, however, cannot leverage a variable like the scheduled dwell time y^{theo}, as the latter equals a single value of 42s for all stations and trains. Thus, the quality of railway operations data prevents a direct comparison of their results to ours ; in particular, they had not exhibited any effect of deviation to scheduled arrival time on dwell time, which is, on the contrary, one of the main determinants of our models for dwell time.

Our approach is to build a single statistical model for dwell times at all stations and in all contexts, which the literature does not often offer. A key variable to be considered to that end is the regime of punctuality of a train at a given station (early arrival, i.e., arrival before the theoretical arrival time ; late arrival, i.e., arrival after the theoretical departure time ; arrival on time, i.e., arrival between the theoretical arrival and departure times). An important note at this stage is that we directly tackle the dwell time (the difference between the departure and the arrival times), and not some notion of "minimum" dwell time (given, e.g., by the alighting and boarding time, or by restricting the attention to the dwell time in constrained situations like late arrival).

We process the variables described above using linear regressions with or without interactions, as well as standard machine-learning methods (random forests, gradient boosting with trees, neural networks), as in Kecman and Goverde [2015]. Models inspired by the latter will form our benchmark, as they tackle dwell times in all situations, including early arrivals—which most references do not offer. As in Kecman and Goverde [2015], these benchmark models will be built solely on operations variables : with our data set we will then be able to characterize the added value of passenger flows variables to model dwell time in a railway context.

In particular, we want to determine when and how much passenger flows impact railway operations.

We now detail several streams of the literature alluded at in the overall view provided above.

Dwell time modeling solely based on railway operations and timetable constraints. On a data set of railway circulation between the Hague and Rotterdam with scheduled stop, Hansen et al. [2010] exhibited a piece-wise linear relationship between dwell time and arrival delay : trains that are early or on time experience average dwell times that decrease with the earliness factor Δa, while the average dwell times of trains out of schedule are independent of how late these trains are. Of course, an explanation is that train drivers must wait the theoretical departure time even if the alighting and boarding process is over. We obtain a similar relationship on our data set, see Figure 3.3. Kecman and Goverde [2015] also consider data collected on trains circulating between the Hague and Rotterdam : their scheduled dwell times y^{theo}, deviations Δa to scheduled arrival time, and train types. These variables are also available on our data set and we formally define them in Section 3.2. They process these variables using linear regression-type methods or random forests and note that the thus constructed models for dwell times should still be improved.

Of interest is also the literature that rather aims to forecast dwell times, e.g., in some auto-regressive manner by considering past dwell times as features (at the same station for earlier trains or at earlier stations for the same train). We may cite Pritchard et al. [2021] for a UK railway network, though they only discuss delayed trains. (See also Li et al. [2016], for a Dutch railway network without a strict theoretical departure time at short stops.)

Modeling of (lower bounds on) dwell time based on passenger flows. The impact of passenger flows on dwell time was first and mostly studied for transportation means without a strict theoretical departure time, like bus, metro or light railway (as these passengers flows are then the only source of information to model dwell time). The seminal work of Levinson [1983] for buses exhibited an affine relationship between dwell time at the bus scale and passenger affluence, i.e., the sum $A + B$ of the number of alighting A and boarding passengers B. Lin and Wilson [1992] for light railway in Boston and Puong [2000] for metro also in Boston (MBTA Red line) studied a multiple linear regression modeling with variables A and B considered separately and together with a crowding factor C. All these references were based on small-scale data obtained by human observations.

For the mass transit modes described in the previous paragraph, the dwell time equals the time for alighting and boarding (also known as the exchange time—depicted in orange in Figure 3.1), i.e., the time between the first passenger exchange and the last passenger exchange, plus a technical time around the latter (e.g., to open and close the doors and to arrive or leave the stop—in grey in Figure 3.1). In

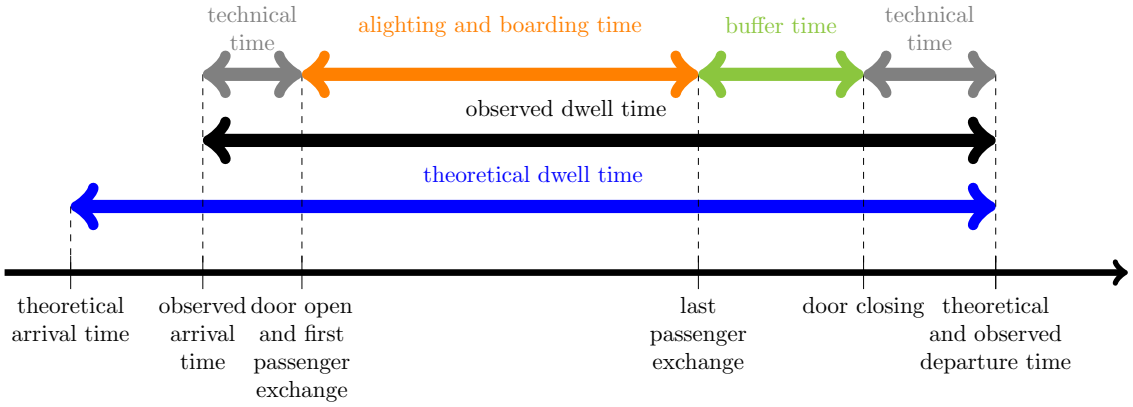


FIGURE 3.1 – Decomposition of the total dwell time in railway context.

a railway context with strict theoretical departure time, dwell time contains a third component : a buffer time (in green in Figure 3.1), corresponding to some additional waiting time (till the strict theoretical departure time) when the train is early. This buffer exists by design, as some operational margin is usually added when timetables are conceived, for the sake of robustness. The literature thus rather focused on some lower bounds on the dwell time, or on the dwell time in constrained situations like late arrivals when there is no buffer time.

Among them, Buchmüller et al. [2008] studied the alighting/boarding time only for train stops without theoretical departure time constraints. Pedersen et al. [2018] and Medeossi and Nash [2020] reduced their attention to delayed trains (for which it is essentially assumed that their dwell times equal the alighting/boarding time plus a technical time, as in the case of buses and metros). The intuition behind the descriptive study by Pedersen et al. [2018] was indeed that passenger flows variables should be useful in these situations. Finally, Cornet et al. [2019] introduces some concept of empirically minimal dwell time, which they then model (essentially in some affine way). Their concept stems from running a PCA on the dwell time based on the numbers A and B of alighting and boarding passengers and the load L of the train ; it turns out that the scatterplot of dwell time on the first principal component of this PCA reveals an affine lower bound.

We will propose a complete modeling of dwell time (i.e., for all stations and all trains) using passenger flows variables on top of railway operations variables. In the presence of a strict theoretical departure time, passenger flows variables provide useful additional information for dwell time modeling on top of railway operations variables (which remain the most critical variables to be used).

A specific discussion of passenger flows by door. It is intuitively clear (and was later demonstrated) that alighting/boarding time discussed above depends on the passenger affluences $A^i + B^i$ by door i and not only on the total passenger affluence $A + B$. However, these passenger affluences $A^i + B^i$ are not uniform at all and strongly depend, for each station, on the closeness to the entry/exit of the platform (see the studies by Wirasinghe and Szplett [1984] and Wiggenraad [2001]).

Yet, most data sets with passenger flows measure them only at the train scale and not at the door scale ; their treatment then has to rely on an unrealistic assumption of uniform distribution of passenger affluence by door, i.e., $A^i + B^i = (A + B)/I$, where I is the number of doors. For recent examples, see Palmqvist et al. [2020] and Medeossi and Nash [2020]. (We note that Wirasinghe and Szplett [1984] proposed a theoretical model based on Gumbel's distribution for boarding numbers by door based on the location of exit/entry platforms and the number of doors.)

On the contrary, our data set is richer than the one of Cornet et al. [2019] as it also contains door-by-door measures A^i and B^i of alighting and boarding numbers. We may then define the critical passenger affluence M, which is the maximum of the $A^i + B^i$ over the doors i, and see its added value on the modeling. If there is some, then, somehow, it is proven that the assumption of uniform distribution of passenger affluence by door is unrealistic. However, as discussed in Section 3.5 based on the survey by Kuipers et al. [2021a], there is room for further exploration of ways for defining critical passenger flows.

We cannot define a meaningful notion of crowding factor at door scale as the trains considered have corridor connections between coaches.

We note that in a different context with no timetable (Beijing subway Line 13) and thus a modeling solely based on passenger flows, Chu et al. [2015] already modeled the dwell time (equal to alighting/boarding time plus a fixed technical time in this context) based on boarding numbers per door B^i, together with global alighting numbers A and crowding factor C (which they turn into per-door quantities by diving by the number I of doors, i.e., using the unrealistic assumption of uniform distribution). Their data set was however of small scale (it was obtained by human observations).

Outline of the chapter and summary of the results obtained

We describe the available data set and the railway context in Section 3.2 : as discussed above, unlike most previous studies in the literature, it offers both railway operations variables and passenger flows variables. We then explain in Section 3.3 (and Appendix 3.A) which machine-learning methods we consider to build, in a data-driven way, models for dwell time that can be used for all stations, all working days, all hours, and all trains. The modeling performance obtained by these models is discussed in Section 3.4, both at a global and at a "local" level. As summarized in our conclusions in Section 3.5, the main findings of our study are the following.

1. On average at a global level, the consideration of passenger flows variables on top of railway operations variables (only) decreases by about 0.5 s the modeling error on the observed dwell time based on mere railway operations variables.
2. However, it locally improves this modeling error in critical situations (while never deteriorating performance in non-critical situations) by sometimes up to 5 – 10 s on average : most notably, for late arrivals or for dense situations

(when passenger affluence is large).

3. More generally, the most influential variables to model the observed dwell time are the passengers flows for late arrivals, and the scheduled dwell time and deviation to scheduled arrival time for early trains.
4. The model built to obtain the global and local improvements mentioned above are based on a fully data-driven machine-learning technique called random forests, but it turns out that a closed-form linear regression model with multiplicative effects (by stations, ways and regimes of punctuality—trains that are early, on time, or late), that is also fully data driven, obtains a modeling performance that is only slightly worse.

We note that Appendix 3.B discusses how robust our findings are, by considering variations around the main results.

3.2 Methodology : description of the data set

We consider a suburban railway network located in the Greater Paris area, and operated by Transilien SNCF. More precisely, we are interested in two different branches of lines H and L, featuring respectively 13 stations (11 without origin/terminus) and 11 stations (9 without origin/terminus) ; see Figure 3.2. We picked them because they are completely or almost completely run with Z50000-type rolling stocks equipped both with on-train monitoring recorder (OTMR) systems, which measure speed, arrival and departure times more precisely than track circuits, and with an automatic passenger counting (APC) system, which measures, for each door of the rolling stocks, the numbers of passengers boarding and alighting at each stop. Z50000-type rolling stocks on lines H and L are composed of 8 and 7 communicating coaches, respectively. For both lines, the mean seating and total capacities by coach equal respectively 59 seats and 119 passengers. The doors width is 1.96m. We are primarily interested in line L and postpone the study of line H in Appendix 3.B.2 ; we explain in depth at the end of this section why we do so.

The data set spans 18 months, from March 15, 2018 to September 16, 2019. Each daily train ride comes with a unique ID, which we will refer to as the train ID. Three primary keys will therefore be used to refer to individual data points : the train number k, station s and day d : see Table 3.1. We merge the two data sources (OTMR data and APC data) by matching the triplets (k, s, d). We keep all triplets present in both data sources and delete the other ones. We do not impose further restrictions, like the availability of all triplets (k, s', d) for a given day d and a given train ride k when s' spans the set of stations. The further pre-processing steps carried out are described below.

Description of the variables. The variables initially available for each triplet (k, s, d) are summarized in Table 3.2. Table 3.3 lists the variables created based on the ones of Table 3.2.

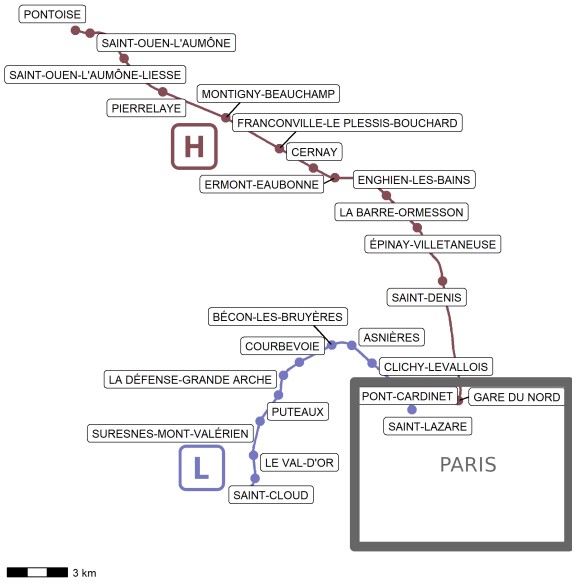


FIGURE 3.2 – Branches of interest in lines H and L of the suburban railway network of the Greater Paris area.

TABLE 3.1 – Primary-key variables.

Variable	Notation
Train number	k
Station	s
Day	d

The railway operations variables of Table 3.2 consist first of observed and scheduled (theoretical) arrival times a^{obs} and a^{theo}, and actual (observed) and scheduled (theoretical) departure times d^{obs} and d^{theo}. Dwell times (observed and theoretical ones) are defined as the differences $y^{\text{obs}} = d^{\text{obs}} - a^{\text{obs}}$ and $y^{\text{theo}} = d^{\text{theo}} - a^{\text{theo}}$. The variable of interest is the observed dwell time y^{obs}. All these variables are indexed by triplets (k, s, d). Two final variables are only indexed by (k, d) as they only depend on the train rides, not on the specific stations : the capacity c of the rolling stocks (the maximal passenger load allowed) and their types t. The type is "single" for single-unit trains and "double" for double-unit trains. The latter are mostly used during rush hours to increase capacity. Scheduled times were obtained from the timetables while other railway operations variables were picked in the OTMR data set.

Based on the variables just described, we may compute three other railway operations variables described in Table 3.3. The way w (that only depends on the train number k, i.e., on the ride) indicates whether the train goes from Paris to its suburbs, or from a suburban area to Paris [1]. The deviation to the scheduled arrival

1. This is an important variable in the Greater Paris area : at morning peak hours, trains from the suburbs to Paris are crowded and suffer more frequently from delays, while trains from Paris to the suburbs circulate in a smoother fashion. In the afternoon peak hours, the situation is the

TABLE 3.2 – Railway operations variables (*top and middle parts of the table*, lower case) and passenger flow variables (*bottom part of the table*, upper case).

Variable	Domain and units	Notation
Variable of interest		
– Observed dwell time	$\{0\text{ s}, 2\text{ s}, \ldots, 180\text{ s}\}$	$y^{\text{obs}}_{k,s,d} = d^{\text{obs}}_{k,s,d} - a^{\text{obs}}_{k,s,d}$
Railway operations [Timetable data]		
– Theoretical (scheduled) arrival time	10 s steps	$a^{\text{theo}}_{k,s,d}$
– Theoretical (scheduled) departure time	10 s steps	$d^{\text{theo}}_{k,s,d}$
– Theoretical (scheduled) dwell time	$\{0\text{ s}, 10\text{ s}, \ldots, 180\text{ s}\}$	$y^{\text{theo}}_{k,s,d} = d^{\text{theo}}_{k,s,d} - a^{\text{theo}}_{k,s,d}$
Railway operations [OTMR data]		
– Observed arrival time	2 s steps	$a^{\text{obs}}_{k,s,d}$
– Observed departure time	2 s steps	$d^{\text{obs}}_{k,s,d}$
– Capacity (maximal passenger load)	{720 ; 922 ; 1,520 ; 1,844}	$c_{k,d}$
– Type	{single, double}	$t_{k,d}$
Passenger flows [APC data]		
– Alighting (number of passengers alighting)	$\{0, 1, 2, 3, 4, \ldots\}$	$A_{k,s,d}$
– Boarding (number of passengers boarding)	$\{0, 1, 2, 3, 4, \ldots\}$	$B_{k,s,d}$
– Load of the train after departure	$\{0, 1, 2, 3, 4, \ldots\}$	$L_{k,s,d}$

time Δa is the difference $a^{\text{obs}} - a^{\text{theo}}$ between the actual arrival time a^{obs} and the scheduled one a^{theo}. Three situations may actually arise in terms of punctuality, and this leads to a final, categorical, variable called "Regime of punctuality" and denoted by z. Early trains, i.e., trains for which $a^{\text{obs}} < a^{\text{theo}}$, will be tagged with $z = 1$. Late arrivals are tagged with $z = 3$ and will refer to trains arriving after the scheduled departure time, i.e., for which $a^{\text{obs}} > d^{\text{theo}}$. (By definition, trains with a late arrival are not tied anymore by the constraint of not leaving before the scheduled departure time.) The third category $z = 2$ corresponds to trains on time, for which $a^{\text{theo}} \leqslant a^{\text{obs}} \leqslant d^{\text{theo}}$.

We only use some of the passenger flows variables available. Indeed, the APC data set reports the numbers of passengers alighting and boarding for each train at each station, globally (variables A and B) and for each door i (variables A^i and B^i). All these variables are indexed by triplets (k, s, d). The values available in the APC data set are not raw data but were obtained after some pre-processing ensuring consistency (e.g., total numbers A and B are the sums of the by-door quantities A^i and B^i ; the sums of the boarding numbers along the ride equal the sums of the

opposite one.

TABLE 3.3 – Processed variables (with the same breakdown as in Table 3.2).

Variable	Domain and units	Notation
Railway operations		
– Way	{0,1}	w_k
	$w_k = 1$ *if train* k *goes from Paris to suburbs,* $= 0$ *from suburbs to Paris*	
– Deviation to scheduled arrival time	$[-600\text{ s}, 600\text{ s}]$	$\Delta a_{k,s,d} = a^{\text{obs}}_{k,s,d} - a^{\text{theo}}_{k,s,d}$
– Regime of punctuality	{1, 2, 3}	$z_{k,s,d}$
	$= 1$ *if train is early,* $a^{\text{obs}}_{k,s,d} < a^{\text{theo}}_{k,s,d}$;	
	$= 2$ *if on time,* $a^{\text{theo}}_{k,s,d} \leqslant a^{\text{obs}}_{k,s,d} \leqslant d^{\text{theo}}_{k,s,d}$;	
	$= 3$ *if late,* $d^{\text{theo}}_{k,s,d} < a^{\text{obs}}_{k,s,d}$	
Passenger flows		
– Crowding factor	[0, 2]	$C_{k,s,d} = L_{k,s,d}/c_{k,s,d}$
– Passenger affluence at the critical door	$\{0, 1, 2, 3, 4, \dots\}$	$M_{k,s,d}$

alighting numbers). Such a pre-processing is required because of the measurement noise due to the infra-red sensor.

To avoid considering too many variables, we only use the total numbers A and B of alighting and boarding passengers (Table 3.2), as well as the passenger affluence at the critical door, defined as the maximal number, over the I doors, of alighting and boarding passengers at a given door i :

$$M = \max\big\{A_i + B_i : i = 1, \dots, I\big\}\,. \tag{3.1}$$

The passenger affluence at the critical door M is thus a processed variable (Table 3.3). A second processed variable is the crowding factor $C = L/c$, defined as the ratio between the load L and the maximal capacity c. We observe some values of C larger than 1 in the data set.

All in all, our data set is a unique combination of typically accessible railway operations variables with rich passenger flows variables. The closest data set in the literature is the one of Cornet et al. [2019], which however does not contain by-door measures of passenger affluence.

Modeling vs. prediction. The focus of the present chapter is only on modeling dwell time based on the explanatory variables described above. The passenger flows variables are available in real time (i.e., right after the train leaves a station) while the railway operations variables are only known with some delay (they are not transmitted in real time). To move from modeling to prediction we would need to predict passenger flows variables for the next station and know the railway operations variables in real time (e.g., know the deviation Δa to scheduled arrival when the train stops and the passenger exchange starts taking place). It turns out that the APC data set actually contains some railway operations variables, measured in real time, but they are less reliable than the OTMR measurements. In any case, we

would need predictions for passengers flows. This is why the focus of the present chapter is only on modeling (i.e., explaining the determinants of dwell time) and not on forecasting.

Further pre-processing of the data / data volume. On top of the pre-processing described above, which consisted of keeping only observations and variables relative to triplets (k, s, d) present in both data sources (OTMR and APC), we performed some data cleaning. First, we deleted triplets (k, s, d) corresponding to anomalous situations : when the observed dwell time $y^{\text{obs}}_{k,s,d}$ is longer than 180 seconds (as Cornet et al. [2019], Dueker et al. [2004] did) or when current cumulative delays on the ride are larger than 10 minutes.

Doing so, we get more than 350,000 observations for line L and 416,000 for line H.

Railway contexts : line L is more important than line H. We mainly report results for line L in the next sections and will only briefly discuss line H later on, in Section 3.B.2. We explain here the several reasons why we favor line L over line H in our study.

First, the passenger flows on line L are more varied than on line H. On line H, most of the passenger flows take place at terminus, while on line L, there exist major intermediate stations (like La Défense Grande Arche) also generating major passenger flows. The average passenger volumes vary from 1,300 to 37,000 passengers per day on the considered branch of line L.

Second, the railway operations are more challenging on line L than on line H. On the one hand, the traffic on line L is more dense than the one of line H for stations distant from Paris, on peak hours : typically, a train every 5 minutes on line L versus every 15 minutes on line H. (For stations close to Paris, the density is similar, with about 22 to 24 trains per hour, that is, a train every 2 to 3 minutes.) On the other hand, lines L and H also differ in terms of punctuality, with a greater variety of situations for line L : most of the train rides on line H end up being on time or late, while there is also a significant fraction of early train rides on line L on top of on-time and late train rides.

Figure 3.3 depicts the observed dwell times, as well as the averages thereof, on the considered branch of line L, by deviations Δa to scheduled arrival times. The averages and standard errors were obtained by a generalized additive method modeling based on 10 cubic splines, see Wood [2006]. We build confidence intervals around the averages of half-widths ± 2 standard errors. Finally, boarding volumes on line L are sufficiently higher than on line H to result in larger crowding factors, despite the higher density of trains.

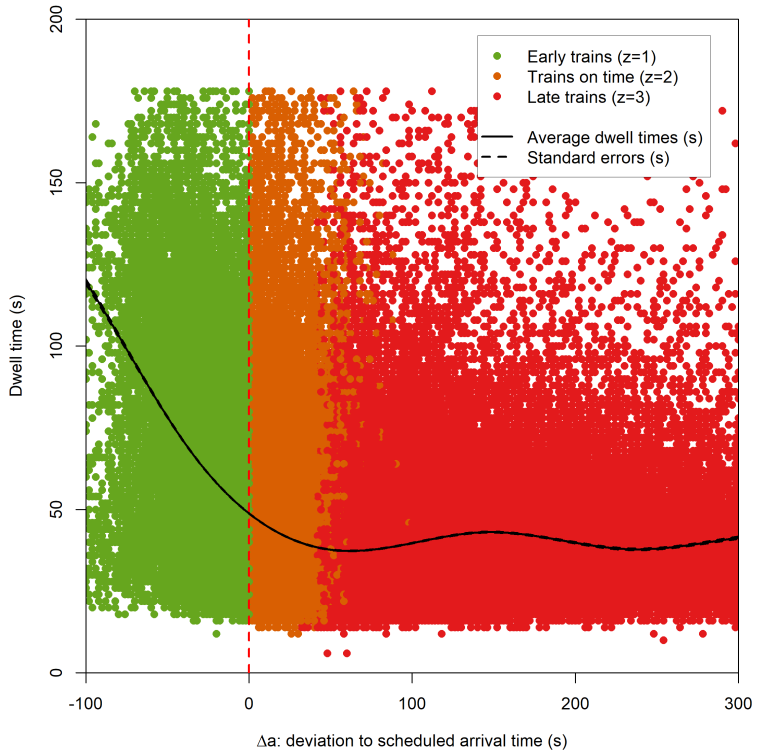


FIGURE 3.3 – Observed dwell times (y–axis, seconds) by deviations Δa to scheduled arrival times (x–axis, seconds) ; we also report average dwell times and standard errors thereof, based on some generalized additive modeling. (The corresponding lines are extremely close to each other.) Three regimes are considered : early trains, trains on time, late trains.

3.3 Methodology : regression models and machine-learning methods

We model the observed dwell times $y^{\text{obs}}_{k,s,d}$ as a stochastic function of *some* of the variables described in Tables 3.2 and 3.3, namely,

$$y^{\text{obs}}_{k,s,d} = f(s,\, w_k,\, t_{k,d},\, y^{\text{theo}}_{k,s,d},\, \Delta a_{k,s,d},\, z_{k,s,d},\, A_{k,s,d},\, B_{k,s,d},\, C_{k,s,d},\, M_{k,s,d}) + \varepsilon_{k,s,d}\,, \quad (3.2)$$

where f is some deterministic function and the additive residual terms $\varepsilon_{k,s,d}$ are random variables (assumptions thereon will depend on each method used, see below). We justify below the choice of the variables used in Equation (3.2).

We are interested in some statistical modeling and do not propose simulation or probabilistic models (as did D'Acierno et al. [2017] or Cornet et al. [2019]). Also, we model directly $y^{\text{obs}}_{k,s,d}$, and not $y^{\text{obs}} - y^{\text{theo}}$, as we want to separate the respective information provided by passenger flows variables and railway operations variables. See Appendix 3.B.1 for more details and a report of the performance obtained by rather modeling $y^{\text{obs}} - y^{\text{theo}}$.

3.3.1 Justification of the variables used

First, for each Transilien network branch, the combination of the station s and the way w_k indicates on which specific platform the train will stop. This is important in light of studies like the one by Daamen et al. [2008], who confirmed the major impact of platform design (stepping gap, height difference, etc.) on the alighting and boarding time.

Now, out of the many railway operations variables available, we only use $y^{\text{theo}}_{k,s,d}$, $\Delta a_{k,s,d}$, $z_{k,s,d}$, $t_{k,d}$. We do so because we want to build a model that can be easily grasped. First, the scheduled dwell time y^{theo} of course provides some benchmark on the expected dwell times ; this piece of information is typically used in modelings, in some direct or indirect way, see, among others, Kecman and Goverde [2015] and Li et al. [2016]. We already explained that Hansen et al. [2010] showed how important the deviation to scheduled arrival time Δa is to explain the dwell times, and Figure 3.3 illustrated it. We build regimes of punctuality z based on Δa (see Table 3.3) to isolate unconstrained dwell times (for late trains, $z = 3$) from dwell times constrained by the scheduled departure time. In particular, we expect that early trains which do not face too high a passenger affluence need to wait till the scheduled departure time and hence, have an observed dwell time equal to $y^{\text{theo}} + |\Delta a|$, the scheduled dwell time plus how early they were. On the contrary, we expect that drivers of late trains will try to shorten dwell times as much as possible.

Finally, the type t of train is also important : it provides an indirect idea of the expected passenger affluence, as double-unit trains are only used when necessary ; this idea is inspired from Kecman and Goverde [2015]. It may also help because double-unit trains and single-unit trains occupy different shares of the platform, and we already mentioned how important the design of the platform is. However, we chose not to consider the other railway operations variables, that should either be irrelevant (scheduled departure time, scheduled and observed arrival times should not convey any information beyond what is already contained in Δa and y^{theo}) or be future variables (the observed departure time is basically what is to be modeled).

As far as passenger flows variables are considered, we consider them all except the load L of the train, as the latter only has a meaning relative to the train capacity—hence the crowding factor C.

Remark : no auto-regressive modeling. Our aim is to model dwell times based on the current context (state of railway operations, passenger affluence, etc.) and determine which elements of this context have the most important influence on dwell time. Our aim is not to forecast dwell times. Therefore, we do not consider auto-regressive-type models, i.e., do not include variables like $y^{\text{obs}}_{k-1,s,d}$ or $y^{\text{obs}}_{k,s-1,d}$ in the modeling of $y^{\text{obs}}_{k,s,d}$. See Li et al. [2016] and Pritchard et al. [2021] for such modelings.

Subsets of variables : RO, PF, M. As we want to determine which variables are most influential, we group them in two groups and a half. We always use s and w_k and cluster the rest of the variables into

- Railway operations variables [short-hand notation "RO"] : $y^{\text{theo}}_{k,s,d}$, $\Delta a_{k,s,d}$, $z_{k,s,d}$, $t_{k,d}$;
- Passenger flows variables [short-hand notation "PF"], not taking into account the passenger affluence at the critical door : $A_{k,s,d}$, $B_{k,s,d}$, $C_{k,s,d}$;
- Passenger affluence at the critical door [short-hand notation "M"] : $M_{k,s,d}$.

We will actually run our methods with s, w_k and either just RO variables, or just PF variables, or RO+PF variables, or RO+PF+M variables.

One model for all stations, all working days, all hours, and all trains. We restrict our attention to working days (i.e., Mondays to Fridays that are not public holidays nor belong to school holidays). We do so because we want to assess the impact of passenger flows on dwell time, and these flows are limited on non-working days. Our second aim is to build models suitable for all stations, all working days, at all hours and for all trains simultaneously, using only variables s (the station) and w_k (the way of train k) to locally adapt the model.

We however do not try to provide a general model that would work for all train networks (as in Harris and Anderson [2007] and Li et al. [2016]), but rather provide a general methodology to adjust specific dwell-time models for each (sub)network suitably equipped in terms of monitoring devices (APC and OTMR ones).

Linear regression models (see, among others, Lam et al. [1998], Harris and Anderson [2007], Palmqvist et al. [2020]) are a popular such general methodology, that leads to easily interpretable models. Even with linear regression models, some non-linear modeling may be achieved by considering multiplicative effects, which we will do. Machine-learning models were later considered (see, among others, Kecman and Goverde [2015]) to improve the accuracy of the modeling based on linear regressions, at the cost of building black-box models which are highly non-linear per design. We describe linear regression models in Section 3.3.2, and then some machine-learning methods : random forests and gradient boosting in Section 3.3.3, and neural networks in Section 3.3.4. We provide concise such descriptions but refer interested readers to Hastie et al. [2009] for deeper expositions. Finally, we explain in Section 3.3.5 how to tune these methods (on a train set) and evaluate them in a fair way (on a test set).

3.3.2 Linear regression models (with additive or multiplicative effects)

The simplest version of linear regression models uses an affine function f in Equation (3.2). In f, the quantitative variables, namely, $y^{\text{theo}}_{k,s,d}$ and $\Delta a_{k,s,d}$ when RO variables are considered, $A_{k,s,d}$, $B_{k,s,d}$, $C_{k,s,d}$ when PF variables are considered, and $M_{k,s,d}$ for the M variable, are each associated with slope coefficients denoted by

$\beta^{(y)}$, $\beta^{(\Delta a)}$, $\beta^{(A)}$, $\beta^{(B)}$, $\beta^{(C)}$, and $\beta^{(M)}$, respectively. As for the categorical variables, namely, s and w_k in all cases, and $z_{k,s,d}$ and $t_{k,d}$ when RO variables are considered, we include them by considering $K-1$ indicator variables, where K denotes the number of modalities taken. For instance, we denote by S the number of stations and index them by $1, \ldots, S$; we take the last station as a reference modality and the regression function f thus features $S-1$ coefficients $\beta^{\text{station}}_{s'}$, where $s' \in \{1, \ldots, S-1\}$. Similarly, for t and w, which both only take two values, we take the modalities "single" and $w = 0$ (from suburbs to Paris) as reference values, and the regression function f features the coefficient β^{type} and β^{way}. Finally, for the variable z which take three modalities, we pick $z = 2$ (trains on time) as a reference value and thus have two coefficients β^{early} and β^{late} for inclusion in f. We denote the global intercept by β^0. All in all, with the simultaneous consideration of the RO, PF, and M variables, we use in Equation (3.2)

$$
\begin{aligned}
&f(s, w_k, t_{k,d}, y^{\text{theo}}_{k,s,d}, \Delta a_{k,s,d}, z_{k,s,d}, A_{k,s,d}, B_{k,s,d}, C_{k,s,d}, M_{k,s,d}) \qquad (3.3)\\
&= \left.\begin{aligned} &\beta^0 + \beta^{\text{way}} \mathbb{1}_{[w_k=1]} + \sum_{s'=1}^{S-1} \beta^{\text{station}}_{s'} \mathbb{1}_{[s=s']} \end{aligned}\right\} \text{ in all cases}\\
&\quad \left.\begin{aligned} &+ \beta^{(\Delta a)} \Delta a_{k,s,d} + \beta^{(y)} y^{\text{theo}}_{k,s,d} + \\ &\quad \beta^{\text{type}} \mathbb{1}_{[t_{k,d}=\text{double}]} + \beta^{\text{early}} \mathbb{1}_{[z_{k,s,d}=1]} + \beta^{\text{late}} \mathbb{1}_{[z_{k,s,d}=3]} \end{aligned}\right\} \text{ RO variables}\\
&\quad \left.\begin{aligned} &+ \beta^{(A)} A_{k,s,d} + \beta^{(B)} B_{k,s,d} + \beta^{(C)} C_{k,s,d} \quad + \beta^{(M)} M_{k,s,d} \end{aligned}\right\} \text{ PF+M variables}
\end{aligned}
$$

If we only use some of these variables, we suppress some terms in the equation above (e.g., the second line if we only use the PF+M variables; or the term $\beta^{(M)} M_{k,s,d}$ if we use the RO+PF variables).

Linear regression with additive effects. We call the model above the linear regression with additive effects. It features $S+1$ coefficients in all cases, plus 5 coefficients when RO variables are included, 3 when PF variables are used, and 1 for the M variable, respectively. This leads to the numbers of coefficient stated in the first line of Table 3.4.

Multiplicative effect of Δa by z. To take into consideration the special relation between the dwell time and the deviation to scheduled arrival time (see Figure 3.3), we provide a different affine modeling in terms of $\Delta a_{k,s,d}$ for each value of $z_{k,s,d}$. Put differently, instead of a single slope coefficient $\beta^{(\Delta a)}$ in front of $\Delta a_{k,s,d}$, we provide a breakdown by punctuality $z_{k,s,d} \in \{1, 2, 3\}$ and use three different slope coefficients $\beta^{(\Delta a)}_1$, $\beta^{(\Delta a)}_2$, $\beta^{(\Delta a)}_3$. We do this on top of setting different intercept levels through the consideration of β^{early} and β^{late}. That is, with the simultaneous consideration of

the RO, PF, and M variables, we use in Equation (3.2)

$$
\begin{aligned}
& f(s, w_k,\, t_{k,d},\, y^{\text{theo}}_{k,s,d},\, \Delta a_{k,s,d},\, z_{k,s,d},\, B_{k,s,d},\, A_{k,s,d},\, C_{k,s,d},\, M_{k,s,d}) \qquad (3.4)\\
&= \left.\begin{aligned} & \beta^0 + \beta^{\text{way}} \mathbb{1}_{[w_k=1]} + \sum_{s'=1}^{S-1} \beta^{\text{station}}_{s'} \mathbb{1}_{[s=s']} \end{aligned}\right\} \text{ in all cases}\\
&\quad \left.\begin{aligned} & + \sum_{z\in\{1,2,3\}} \mathbb{1}_{[z_{k,s,d}=z]}\, \beta^{(\Delta a)}_z\, \Delta a_{k,s,d} \end{aligned}\right\} \begin{array}{l}\text{RO variables,}\\ \text{new part : interaction between } \Delta a \text{ and } z\end{array}\\
&\quad \left.\begin{aligned} & + \beta^{(y)}\, y^{\text{theo}}_{k,s,d} + \beta^{\text{type}} \mathbb{1}_{[t_{k,d}=\text{double}]} + \beta^{\text{early}} \mathbb{1}_{[z_{k,s,d}=1]} + \beta^{\text{late}} \mathbb{1}_{[z_{k,s,d}=3]} \end{aligned}\right\} \begin{array}{l}\text{RO variables,}\\ \text{no change}\end{array}\\
&\quad \left.\begin{aligned} & + \beta^{(A)}\, A_{k,s,d} + \beta^{(B)}\, B_{k,s,d} + \beta^{(C)}\, C_{k,s,d} \qquad + \beta^{(M)}\, M_{k,s,d} \end{aligned}\right\} \text{PF+M variables}
\end{aligned}
$$

We call the model above the linear regression with a multiplicative effect of Δa by z. When RO variables are considered, it contains two additional coefficients with respect to the model with additive effects and does not differ from the latter when RO variables are omitted ; see the second line of Table 3.4.

Additional multiplicative effects. We may have the slope coefficients, as well as the intercepts, vary by pairs (s, z) or even, triplets (s, w, z) to locally tailor the model to the stations and to the regime of punctuality ; i.e., with PF variables, the regression function f would, for instance, feature terms like

$$
\begin{aligned}
& \sum_{\substack{s'\in\{1,\ldots,S\}\\ w\in\{0,1\}\\ z\in\{1,2,3\}}} \mathbb{1}_{\left[\begin{smallmatrix} s=s',\\ w_k=w,\\ z_{k,s,d}=z\end{smallmatrix}\right]}\, \beta^{(A)}_{s,w,z}\, A_{k,s,d} + \sum_{\substack{s'\in\{1,\ldots,S\}\\ w\in\{0,1\}\\ z\in\{1,2,3\}}} \mathbb{1}_{\left[\begin{smallmatrix} s=s',\\ w_k=w,\\ z_{k,s,d}=z\end{smallmatrix}\right]}\, \beta^{(B)}_{s,w,z}\, B_{k,s,d} +\\
& \qquad\qquad \sum_{\substack{s'\in\{1,\ldots,S\}\\ w\in\{0,1\}\\ z\in\{1,2,3\}}} \mathbb{1}_{\left[\begin{smallmatrix} s=s',\\ w_k=w,\\ z_{k,s,d}=z\end{smallmatrix}\right]}\, \beta^{(C)}_{s,w,z}\, C_{k,s,d}
\end{aligned}
$$

instead of $\beta^{(A)}\, A_{k,s,d} + \beta^{(B)}\, B_{k,s,d} + \beta^{(C)}\, C_{k,s,d}$. For multiplicative effects by triplets, we end up with models with at least $6S$ coefficients per quantitative variable considered (the total number of coefficient depending on the specific dependencies considered for the intercepts) ; see the third line of Table 3.4. We tried many formulations and all get a similar performance.

The linear models discussed in this section are reference models and are mostly of interest for the sake of comparison with more complex, machine-learning, methods, which often exhibit a better performance, at the cost of not leading to statistical models, i.e., closed-form relationships that may be interpreted. We consider two methods based on regression trees, which we describe now, and one on neural networks, which we describe later.

TABLE 3.4 – Numbers of coefficients of the various linear regressions considered, for line L (for which there are $S = 10$ stations).

	Variables used			
	PF	RO	RO+PF	RO+PF+M
Additive effects	14	16	19	20
Multiplicative effect of Δa by z	14	18	21	22
Multiplicative effects by (s, w, z)	$\geqslant 180$	$\geqslant 120$	$\geqslant 300$	$\geqslant 360$

3.3.3 Machine-learning methods based on regression trees : random forests and gradient boosting

Machine-learning methods based on regression trees were already considered by the literature on transportation systems : Kecman and Goverde [2015] used random forests to model dwell time for trains circulating between the Hague and Rotterdam, based on similar railway operations (RO) variables as we consider in this chapter ; Ding et al. [2016] used gradient boosting and were interested in short-term metro ridership forecasting (next 15 minutes) on three major Beijing stations. Zhang and Haghani [2015] considered both methods to forecast car travel time on a motorway section in Maryland ; in their study, boosting methods slightly outperformed random forests. All three references present in details random forests and gradient boosting and do so by first introducing regression trees. We follow the same path.

Concept of a regression tree. We denote by $\boldsymbol{X}_{k,s,d}$ the feature vectors, i.e., the vectors of variables available for each triplet (k, s, d). These variables were described in Section 3.3.1, except that we replace the non-binary categorical variables s and z, which have S and 3 modalities, by S and 3 binary variables, respectively [2]. Table 3.5 indicates the size of $\boldsymbol{X}_{k,s,d}$ depending on the subset of variables used, by distinguishing components that are quantitative variables and the ones that are given by binary categorical variables.

A regression tree relies on a (hierarchically organized) partition of the feature space into finitely many regions $\mathcal{R}_1, \ldots, \mathcal{R}_R$ defined by thresholds on the components of feature vectors $\boldsymbol{X}$. Indeed, the partition stems from a binary tree, where the two children of each node are defined by a threshold level on a quantitative variable, or the values 0 and 1 of a binary variable. A toy illustration with a two-level hierarchy and its associated partition is provided in Figure 3.4 : the threshold at the root node is based on the number A of passengers alighting and uses the value 150, and there is a second level for the left child, which is based on the number B of passengers boarding and uses the threshold value 200. A regression tree is built on

2. Doing so, we only consider additive effects, as in (3.3). We also tested—but do not discuss here—partially multiplicative effects, for instance, replacing the component $\Delta a_{k,s,d}$ of $\boldsymbol{X}_{k,s,d}$ by the three variables $\Delta a_{k,s,d}\, \mathbb{1}_{[z_{k,s,d}=z]}$, for $z \in \{1, 2, 3\}$, that were used in (3.4). We did not observe significant gains in performance and were not surprised : regression-tree-based methods are per se able to deal with complex interactions between features and output (dwell time).

TABLE 3.5 – Size of the feature vectors $\boldsymbol{X}_{k,s,d}$ for line L (for which there are $S = 10$ stations), depending on the subsets of variables considered.

	Variables used			
Size of $\boldsymbol{X}$	PF	RO	RO+PF	RO+PF+M
Quantitative components	3	2	5	6
Binary components	11	15	15	15
Total number	14	17	20	21

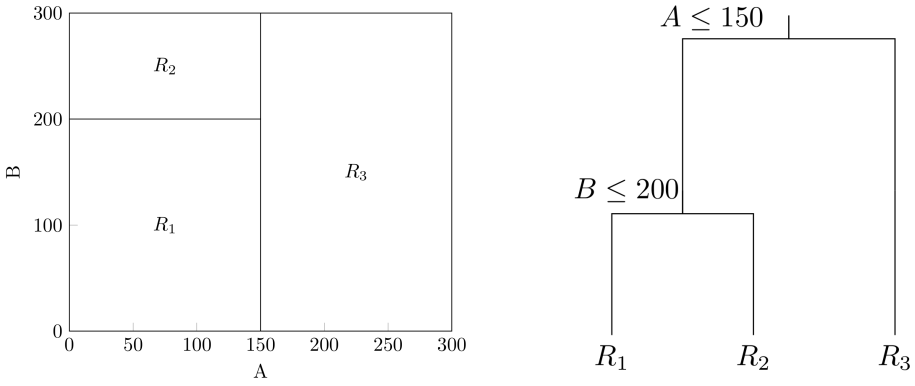


FIGURE 3.4 – Toy example of a regression tree : the partition with 3 elements in terms of values of the variables A and B (*left*) and the associated binary tree (*right*).

train data $\boldsymbol{X}_{k,s,d}$ in a greedy manner through successive refinements of the current binary regression tree until the refinement stops, i.e., when a node is declared a leaf. The variable and associated threshold at each node are determined by considering all possible choices thereof and by picking the pair that leads to the smallest in-sample square error for the corresponding augmented regression tree. We will consider two stopping rules, both aiming to avoid over-fitting the data. The first rule is that if one of the created children node contains fewer than 5 observations, the refinement actually does not occur, and the node at hand is declared a leaf. The second rule is to construct complete binary trees of a fixed depth (where the depth of a tree is defined by the number of nodes along the longest path from the root node down to the farthest leaf), i.e., stop refining when a certain depth is reached. This concludes the description of the construction of a regression tree ; we recall that it may be identified with a hierarchical partition $\mathcal{R}_1, \ldots, \mathcal{R}_R$ of the feature vectors $\boldsymbol{X}$.

Then, when a new feature vector $\boldsymbol{X}$ is to be handled, the method first identifies in which region $\mathcal{R}(\boldsymbol{X})$ of the partition $\mathcal{R}_1, \ldots, \mathcal{R}_R$ this feature vector lies. The modeled dwell time $f(\boldsymbol{X})$ for this new feature vector $\boldsymbol{X}$ finally equals the empirical average of the values $y^{\text{obs}}_{k,s,d}$ of those feature vectors $\boldsymbol{X}_{k,s,d}$ that lie in the same region $\mathcal{R}(\boldsymbol{X})$, if there is at most one such vector (otherwise, an arbitrary value is output) :

$$f(\boldsymbol{X}) = \frac{1}{\sum_{k,s,d} \mathbb{1}_{[\boldsymbol{X}_{k,s,d} \in \mathcal{R}(\boldsymbol{X})]}} \sum_{k,s,d} y^{\text{obs}}_{k,s,d} \, \mathbb{1}_{[\boldsymbol{X}_{k,s,d} \in \mathcal{R}(\boldsymbol{X})]} \, . \tag{3.5}$$

The response function f is piecewise constant (it is constant over each member $\mathcal{R}_r$

of the partition).

One major problem of regression trees comes from their instability, which is due to their hierarchical construction : small variations in data may affect the choices made in the higher nodes and result in drastically different final results. To overcome this issue, two methods were proposed by the machine-learning literature : random forests and gradient boosting with regression trees. Both are ensemble methods using many trees of small depth to avoid over-fitting and to reduce the variances of regression trees.

Random forests. Random forests were introduced by Breiman [2001], they consist of generating (partially at random) T regression trees $f^{(1)}, \ldots, f^{(T)}$ as described above with the first stopping rule, and by resorting to the response function given by the average of these trees :

$$f(\boldsymbol{X}) = \frac{1}{T} \sum_{t=1}^{T} f^{(t)}(\boldsymbol{X}) \,. \tag{3.6}$$

The number T of random trees is large, and the rationale behind the average is that model errors are therefore expected to compensate each others. Kecman and Goverde [2015] was the first to use random forests for dwell time estimation, using RO variables. We explain below how the random trees $f^{(t)}$ are generated and how their number T is chosen.

Two sources of randomness are introduced to construct each given tree $f^{(t)}$: first, the data sample used to build $f^{(t)}$ is obtained by bootstrapping (i.e., by sampling with replacement into the original data), and second, to grow the tree from this bootstrapped sample, only m variables (picked at random) out of the p variables are used. These artificial sources of randomness are useful to create independence between the trees $f^{(1)}, \ldots, f^{(T)}$.

We implemented random forests using the R package `ranger` (see Wright and Ziegler, 2017 ; it is better suited to large data sets than, e.g., the `randomForest` package). It uses two parameters, `ntree` for the the number of trees T and `mtry` for m the number of variables chosen at each split. Both are tuned by cross validation, see Section 3.3.5. The bootstrapped data samples are of the same size as the original data set.

Gradient boosting with regression trees. While random forests rely on a compensation of individual errors through bagging, gradient boosting (Friedman, 2001) iteratively builds weighted sums of regression trees by focusing on the observations with the highest model errors. The regression trees successively picked for the weighted sums are thus not independent from each other.

More precisely, the basic idea of gradient tree boosting is to consider a set $\mathcal{F}$ of possible binary trees and start with an arbitrary tree $f^{(1)} \in \mathcal{F}$; in the chosen

implementation, $\mathcal{F}$ is the set of all complete binary trees of depth 6. At each iteration $t \geqslant 2$, we then construct a weighted sum $f^{(t)}$ of regression trees by first considering the modeling errors

$$e_{k,s,d}^{(t-1)} = y_{k,s,d}^{\text{obs}} - f^{(t-1)}(\boldsymbol{X}_{k,s,d}) \tag{3.7}$$

associated with the weighted sum $f^{(t-1)}$ of the previous step, by picking the best tree $g^{(t)} \in \mathcal{F}$ to model these errors, i.e.,

$$g^{(t)} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{k,s,d} \left(e_{k,s,d}^{(t-1)} - f(\boldsymbol{X}_{k,s,d}) \right)^2 , \tag{3.8}$$

by picking the best step size $\alpha^{(t)} \in \mathbb{R}$ to model these errors given $g^{(t)}$, i.e.,

$$\alpha^{(t)} \in \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \sum_{k,s,d} \left(e_{k,s,d}^{(t-1)} - \alpha g^{(t)}(\boldsymbol{X}_{k,s,d}) \right)^2 , \tag{3.9}$$

and by finally outputting

$$f^{(t)} = f^{(t-1)} + \eta\, \alpha^{(t)} g^{(t)} , \tag{3.10}$$

where we consider a shrinkage parameter η. The optimizations on $f \in \mathcal{F}$ and $\alpha \in \mathbb{R}$ are performed successively and not simultaneously because of computational issues. The procedure stops after T rounds and the final modeling f equals

$$f = f^{(T)} = f^{(1)} + \eta \sum_{t=2}^{T} \alpha^{(t)} g^{(t)} . \tag{3.11}$$

Both η and T are parameters to be set by the user.

As indicated above, these are only the high-level ideas behind the specific method used, namely, XGBoost by Chen and Guestrin [2016], which relies on two decades of advances in boosting and tree methods. We use the R package `xgboost`. The XGBoost method may be finely tuned through a few dozens of parameters, including the choice of the tree set $\mathcal{F}$; we use the default values, except for T and η (which correspond to the parameters `nrounds` and `eta`, respectively), which we tune by cross validation, see Section 3.3.5. It is a common choice (both in machine learning competitions and in the transportation literature, see Ding et al., 2016) to focus mostly on these two parameters. They work hand in hand : on the one hand, large values of T and η lead to over-fitting (i.e., building a model too close to historical data with poor generalization guarantees), on the other hand, for XGBoost to "converge"—i.e., be such that $f^{(t)}$ does not change much as t approaches T—the shrinkage parameter η and T need to be small enough. All in all, a good balance between T and η should be achieved.

3.3.4 Feed-forward neural networks

Artificial neural networks (see Goodfellow et al., 2016) are a popular method for designing highly non-linear predictors, in all fields of science and engineering,

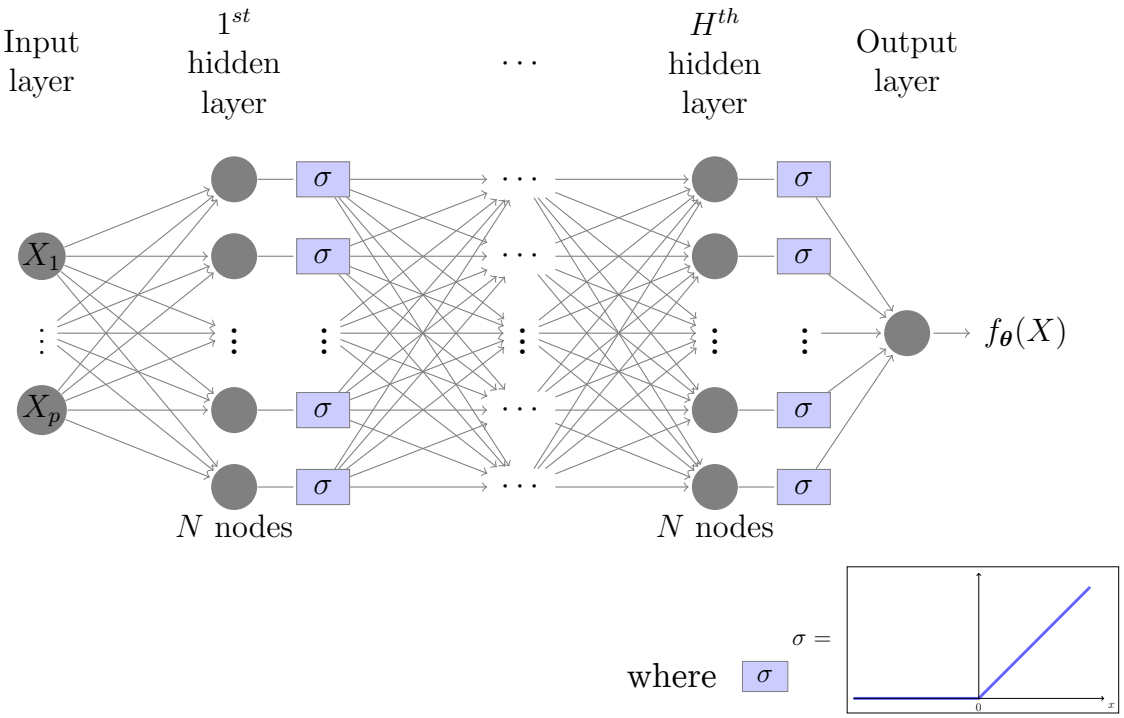


FIGURE 3.5 – Architecture of the feed-forward networks considered ; the σ boxes correspond to the application of the rectified linear unit (ReLU) activation function $\sigma(x) = \max\{x, 0\}$ depicted at the bottom.

including transportation research. They are considered in transportation research about public transports with relatively simple architectures, typically based on at most one hidden layer. For instance, Yaghini et al. [2013] used such simple neural networks to classify train delays for Iranian railways, while Amita et al. [2015] did so to predict bus running times in Dehli based on GPS data. In traffic literature more complex architecture are often considered, as did Li et al. [2018] for the forecasting of road traffic flows on two data sets from California highways. They compare two methods, a dense feed-forward neural network with two hidden layers of 256 nodes each and a more complex diffusion convolution recurrent neural network. As we face a public-transport application, we do not consider the latter method and only proceed with feed-forward neural networks.

The mentioned references all consider different architectures for their feed-forward neural networks : to a great extent, the choice of the architecture of a neural network is subjective and relies on engineering experience. However, in this work, we consider the number of hidden layers H and the number N of nodes per layer as tuning parameters (to be chosen through cross validation, see Section 3.3.5).

The architecture considered for our feed-forward networks is depicted in Figure 3.5 ; it is composed of an input layer, of H hidden dense layers (each with N nodes), and of an output layer. It corresponds to inductively constructing the modeling $f(\boldsymbol{X}_{k,s,d})$ as follows. The output function $f^{(1)}$ of the first layer $h = 1$ takes a p–dimensional vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ as argument, where p is provided by Table 3.5, and outputs a vector of length N, based on real weights $\theta_{j,n,0}$, on intercepts $b_{n,0}$, and on the so-

called rectified linear unit (ReLU) activation [3] function $\sigma(x) = \max\{x, 0\}$:

$$f^{(1)}(\boldsymbol{X}) = \left(\sigma\left(b_{n,0} + \sum_{j=1}^{p} \theta_{j,n,0} X_j\right)\right)_{n\in\{1,\dots,N\}}. \tag{3.12}$$

The hidden layers $h \in \{2, \dots, H\}$ are then each associated with a function $f^{(h)}$ based on the components $f_{n'}^{(h-1)}$ of $f^{(h-1)}$, on real weights $\theta_{n',n,h}$ of the arc connecting node n' of hidden layer $h-1$ and node n of hidden layer h, on intercepts $b_{n,h}$, and on the ReLU activation function σ :

$$f^{(h)}(\boldsymbol{X}) = \left(\sigma\left(b_{n,h} + \sum_{n'=1}^{N} \theta_{n',n,h}\, f_{n'}^{(h-1)}\right)\right)_{n\in\{1,\dots,N\}}. \tag{3.13}$$

The final function $f_{\boldsymbol{\theta}}$ is then based on $f^{(H)}$ and on a final series of real weights $\theta_{n,H+1}$ and on a final intercept b_{H+1} :

$$f_{\boldsymbol{\theta}} = b_{H+1} + \sum_{n=1}^{N} \theta_{n,H+1}\, f_n^{(H)}\,; \tag{3.14}$$

here, we collected all parameters (weights and intercepts, of all layers) into a vector denoted by $\boldsymbol{\theta}$.

The final function f is obtained by fitting $\boldsymbol{\theta}$ on data :

$$f = f_{\hat{\boldsymbol{\theta}}}\,, \qquad \text{where} \qquad \hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{k,s,d} \left(y_{k,s,d}^{\text{obs}} - f_{\boldsymbol{\theta}}(\boldsymbol{X}_{k,s,d})\right)^2. \tag{3.15}$$

Efficient gradient-descent techniques (the so-called gradient back-propagation algorithm) exist to perform the optimization leading to the value of $\hat{\boldsymbol{\theta}}$ (which is called "training the network"). We use the R package `keras` to build the architecture and the R package `tensorflow` to train the network. The model is trained with batch size 32 and mean absolute error as the loss function (see Section 3.3.5). We use the classical `Adam` optimizer, which is based on stochastic sampling, to compute gradients. We run 50 epochs, not more (to avoid over-fitting), not fewer (to train sufficiently the parameters).

3.3.5 Fair assessment of the performance : picking parameters on a train set, and evaluating performance on a test set

All methods above require some training on historical data. Fitting the coefficients of the linear regression models on such historical data is straightforward. Machine-learning techniques (random forests, gradient boosting with regression trees, and feed-forward neural networks) require a more sophisticated use of historical data : they need to pick some hyperparameters—two per method, which we recall in Table 3.6—and fit the model on data based on these hyperparameters.

TABLE 3.6 – Grids for picking the hyperparameters (a.k.a. tuning parameters) on the train set.

Method	Hyperparameter #1	Hyperparameter #2
Random forests	$m \in \{1, 2, \ldots, 15\}$	$T \in \{1, 10, 50, 100, 500, 1000, 5000\}$
Gradient boosting with regression trees	$\eta \in \{0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$	$T \in \{1, 200, 600, 1000, 2000, 6000, 8000\}$
Feed-forward neural networks	$H \in \{1, 2, 3, 4, 5, 6\}$	$N \in \{32, 64, 128, 256, 512\}$

TABLE 3.7 – Tuning parameters selected based on the considered sets of variables.

Methods	Pairs of parameters	Variables sets			
		PF	RO	RO PF	RO PF+M
Random forests	$(T, m) =$	$(10, 15)$	$(100, 10)$	$(500, 5)$	$(5000, 7)$
Gradient boosting with regression trees	$(T, \eta) =$	$(6000, 0.0005)$	$(8000, 0.0005)$	$(6000, 0.005)$	$(6000, 0.005)$
Feed-forward neural networks	$(N, H) =$	$(256, 5)$	$(32, 4)$	$(32, 4)$	$(64, 2)$

A popular solution in statistics, already considered in transportation research by, among others, Kecman and Goverde [2015], consists in separating the data set into two subsets : a train data set and a test data set. For machine-learning methods, the train data set is used both to select hyperparameters by (5-fold) cross-validation and fit the models accordingly. To do so, our procedure consists of two passes on the train data set, a first to select the hyperparameters, based on cross-validation, and a second to fit the corresponding model. A more detailed statement of this procedure may be found in Appendix 3.A.2. For linear regression models, we directly fit coefficients on the train data set. The test data set is used to evaluate the performance of the thus constructed and fitted methods. Doing so, we avoid favorable biases that would consist, for instance, of constructing and evaluating the methods on the same data subset ; in that case, we would be providing some in-sample error rather than an out-of-sample error.

Breakdown used. We consider a fixed 60%–40% breakdown of the data set into a train data set (data points from March 15, 2018 to March 15, 2019) and a test data set (data points from March 16, 2019 to September 15, 2019). We set the 60%–40% proportions in some arbitrary way.

3. Li et al. [2018] also use the ReLU activation function while Yaghini et al. [2013] and Amita et al. [2015] use instead a sigmoid activation function $x \mapsto 1/(1 + \mathrm{e}^{-x})$. We picked the ReLU activation function mostly because of its popularity, as asserted by Goodfellow et al. [2016].

A note on the hyperparameters considered. In the cross-validation procedure alluded at above, hyperparameters are selected based on grids of possible values, provided in Table 3.6. These grids had been determined ex ante and were constructed based on previous choices of these hyperparameters in the literature. For random forests, we built the grids around the default parameters of the `ranger` package (see Wright and Ziegler, 2017), which equal $m = \lfloor \sqrt{p} \rfloor$ for `mtry` (where p is the number of input variables considered) and $T = 500$ for `ntree`; given the values of p (see Table 3.4), this leads to $m = 4$ or $m = 5$. For gradient boosting with regression trees, we take the exact same grids as considered by Zhang and Haghani [2015, Section 3.2]. For the feed-forward neural networks, we built a reasonable grid based on the default values $N = 256$ units and $H = 2$ hidden layers chosen by Li et al. [2018, Annex E].

The hyperparameters selected by the (5-fold) cross-validation procedure are reported in Table 3.7. These are the hyperparameters we use in the rest of the chapter.

It turns out that on the data set considered, the performance of the machine learning methods is not too sensitive to the pairs of hyperparameters considered. More details are to be found in Appendix 3.A.1, where the performance of the methods are tabulated on the grids of Table 3.6 and where we observe that whenever these hyperparameters are large enough, a close-to-optimal performance is reached.

3.4 Main results

The previous section described machine-learning methods to build single data-driven models for dwell time valid for all stations, all working days, all hours, and all trains. In this section, we quantify their modeling performance, i.e., report the modeling errors, namely, the mean absolute modeling errors and the root mean squared modeling errors. We do so both at a global level (Section 3.4.1) and at a local level (Section 3.4.2), possibly also by considering an addition breakdown of the modeling performance by regimes of punctuality or passenger affluence (Section 3.4.3). By "global" results, we mean errors obtained by global averages over all stations, all working days, all hours, and all trains. By "local" results, we mean conditional averages of the form "average error suffered when some explanatory variable equals a given value". We of course define first more formally the concept of "local" performance Section 3.4.2. We conclude by a ranking of the explanatory variables depending on their modeling influence (Section 3.4.4).

Metrics for the assessment of performance. With the notation of Section 3.3, models $\widehat{f}$ are built on the train set and are evaluated on the test set $\mathcal{T}\text{est}$, whose cardinality is denoted by $N_{\mathcal{T}\text{est}}$. The mean absolute error (MAE) and root mean

squared error (RMSE) of such a model $\widehat{f}$ are respectively defined by

$$\text{MAE}\big(\widehat{f}\big) = \frac{1}{N_{\mathcal{T}\text{est}}} \sum_{(k,s,d)\in\mathcal{T}\text{est}} \left| y^{\text{obs}}_{k,s,d} - \widehat{f}(\boldsymbol{X}_{k,s,d}) \right| \tag{3.16}$$

$$\text{and} \qquad \text{RMSE}\big(\widehat{f}\big) = \sqrt{\frac{1}{N_{\mathcal{T}\text{est}}} \sum_{(k,s,d)\in\mathcal{T}\text{est}} \left(y^{\text{obs}}_{k,s,d} - \widehat{f}(\boldsymbol{X}_{k,s,d}) \right)^2}. \tag{3.17}$$

No metric seems preferred in the transportation literature, and each has its own advantages : MAE summarizes best the global performance while RMSE is sensitive to large errors.

3.4.1 Main table : "global" performance

Table 3.8 reports the global performance for the modeling of dwell time, i.e., the MAE and the RMSE achieved on the entire test data set, of the six methods presented in Sections 3.3.2–3.3.4 run on four possible subsets of variables described in Section 3.3.1.

We first comment how the modeling performance depends on the subsets of variables. Using passenger flows [PF] variables only is suboptimal, and railway operations [RO] variables seem key to achieve the best performance. We also observe that overall, using RO variables only is not as good as using RO and PF variables simultaneously, which is itself slightly outperformed by using RO and PF variables together with the M variable consisting of the passenger affluence at the critical door. The observations made above are consistent with previous observations in the literature, which deemed RO variables more important than PF variables for commuter trains (Hansen et al., 2010, Kecman and Goverde, 2015). We detail in subsequent subsections how PF variables, including the M variable, are valuable to consider on top of RO variables. This will, in particular, show the genuine interest of the PF and M variables, which, for now, seems modest on Table 3.8—while one could have expected a more dramatic effect based on the study by Wirasinghe and Szplett [1984].

We now comment the influence of the method. We first observe that the more complex the linear regression models, the better the performance. But linear regression models, which provide explainable relationships, exhibit suboptimal performance compared to the machine-learning methods (random forests, gradient boosting with regression trees, feed-forward neural networks), which do not offer explicit relationships and only provide black-box (highly non-linear) modelings. Among these machine-learning methods, gradient boosting with regression trees performs slightly better than random forests and feed-forward neural networks. All in all, the linear regression with a multiplicative effect of Δa by z probably offers the best trade-off, among all six methods considered, between simplicity, explainability and performance.

TABLE 3.8 – Modeling performance for each method and each set of variables, in MAE (*left part of the table*) and RMSE (*right part of the table*). Columns indicate which variables are used (see Section 3.3.1) : only passenger flows [PF] variables, only railway operations [RO] variables, both RO and PF variables, and all variables (RO, PF, and M, the passenger affluence at the critical door). Each line corresponds to a method to process data : linear regressions (Section 3.3.2), random forests and gradient boosting (Section 3.3.3), feed-forward neural networks (Section 3.3.4). Standard errors are smaller than 0.03 seconds.

Methods	MAE				RMSE			
	PF	RO	RO PF	RO PF+M	PF	RO	RO PF	RO PF+M
1. Linear regression with additive effects	13.7	10.5	10.2	10.1	18.4	14.8	14.5	14.3
2. Linear regression with a multiplicative effect of Δa by z	13.7	9.1	8.9	8.8	18.4	13.6	13.2	13.1
3. Linear regression with multiplicative effects by triplets (s, w, z)	13.3	8.8	8.3	8.3	18.0	13.2	12.6	12.5
4. Random forests	13.7	8.4	8.1	8.0	18.8	12.9	12.5	12.3
5. Gradient boosting with regression trees	12.9	8.5	8.0	7.9	17.9	13.0	12.4	12.2
6. Feed-forward neural networks	12.7	8.4	8.0	8.0	17.4	13.0	12.4	12.2

3.4.2 "Local" performance, depending on the level of explanatory variables

We now provide a more "local" study of performance : instead of reporting global measures of performance, we rather explain how performance varies as a given explanatory variable (passenger affluence, deviation to scheduled arrival time, etc.) varies. For the sake of concision, we will only consider one machine-learning method ; to allow comparison to earlier results, we select random forests : Kecman and Goverde [2015] ran random forests on RO variables, and we will be running them also on all variables (RO, PF, and M). We will refer to both instances of random forests by the short-hand notation RF–RO and RF–All.

Our main aim in this section is to highlight the added value of considering PF and M variables on top of RO variables : while the fourth line of Table 3.8 shows an extremely similar global performance of RF–RO and RF–All, we will demonstrate improvements in the "local" performance thereof. We first explain how we define and

compute the latter.

Concept of local performance. We merely describe here how Figures 3.6–3.9 were obtained and how they measure local performance. We comment below on the gain in efficiency brought by RF–All with respect to RF–RO, in a dedicated series of paragraph.

Figures 3.6–3.9 aim to illustrate the impact of passenger affluence $A + B$ (the sum of the numbers of passengers alighting plus the ones boarding), which is to be found in x–axis, on performance for the modeling of dwell time, which is to be found in y–axis. This performance may be measured in an absolute manner (for RF–RO or for RF–All, as in the left graph of Figure 3.7) or in a relative manner (improvement of RF–All over RF–RO, as in the right graph of Figure 3.7).

We explain first how local performance is measured in an absolute manner. We fix a given method, say, RF–All. The scatterplot underlying the left graph of Figure 3.6 consists of the pairs

$$\left(A_{k,s,d} + B_{k,s,d},\ \left| y^{\text{obs}}_{k,s,d} - \widehat{f}(\boldsymbol{X}_{k,s,d}) \right| \right) \tag{3.18}$$

as (k, s, d) varies in the test set $\mathcal{T}$est. We apply the same smoothing as in Figure 3.3. Doing so, we obtain a curve representing the average absolute error in the modeling of dwell time by the level of passenger affluence (solid line) ; this average is associated with a ± 2 times standard deviation (dotted lines). The right graph of Figure 3.6 is simply a cleaned version of the left one, where we erased the underlying scatterplot.

Now, the representation just described may be performed for RF–RO and for RF–All : see the left graph of Figure 3.7, where we also added a ± 1 s tube starting from the dotted lines. This tube measures the significant improvements : as dwell time is measured with 2 s steps (see Table 3.2), we are only interested in average improvements larger than 1 s. We may read on the left graph the range where RF–All improves significantly over RF–RO : the range where the lower part of the tube around RF–RO is higher than the upper dotted line for RF–All. This range accounts for 5.8% of the observations, as we write on the blue arrow under the curves.

We represent this comparison in an equivalent manner on the right graph of Figure 3.7 : the average difference by passenger affluence is the difference of the average absolute errors by passenger affluence between RF–RO and RF–All, and the associated standard deviations are the sums of the standard deviations associated with the average errors of RF–RO and RF–All. The same ± 1 s tube is depicted, around the value 0.

We may proceed similarly with square errors for the modeling of dwell time. The left graph of Figure 3.8 depicts the scatterplot of

$$\left(A_{k,s,d} + B_{k,s,d},\ \left(y^{\text{obs}}_{k,s,d} - \widehat{f}(\boldsymbol{X}_{k,s,d}) \right)^2 \right) \tag{3.19}$$

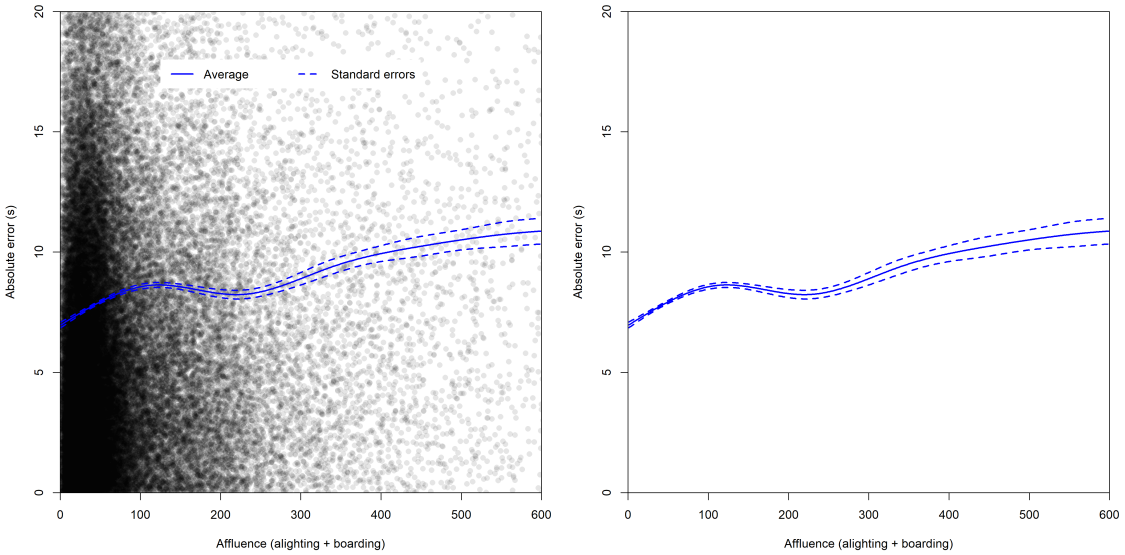


FIGURE 3.6 – Left graph : scatterplot of the absolute errors on the test set for dwell time modeling by RF–All plotted against passenger affluence, together with an estimation of the associated average absolute errors (solid line), and standard errors thereof (dotted lines). Right graph : left graph without the underlying scatterplot.

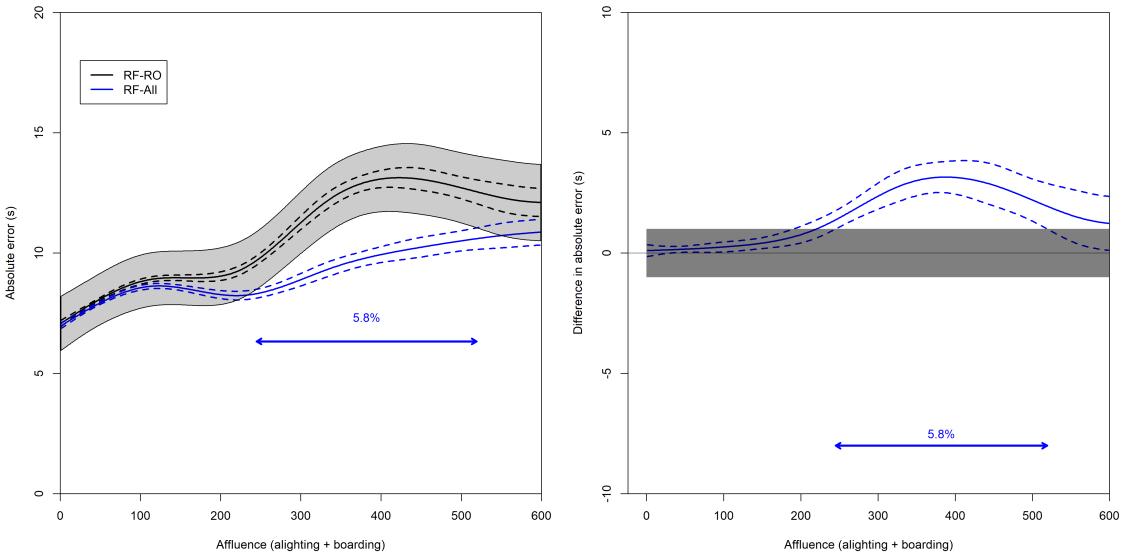


FIGURE 3.7 – Left graph : average absolute error for the modeling of dwell time (y–axis) by passenger affluence (x–axis) for RF–RO (black) and RF–All (blue). Right graph : difference of these average absolute errors, between RF–RO and RF–All. The horizontal arrows indicate the data ranges where significant improvements are achieved on average by RF–All over RF–RO ; the percentages below the arrows are the corresponding data shares.

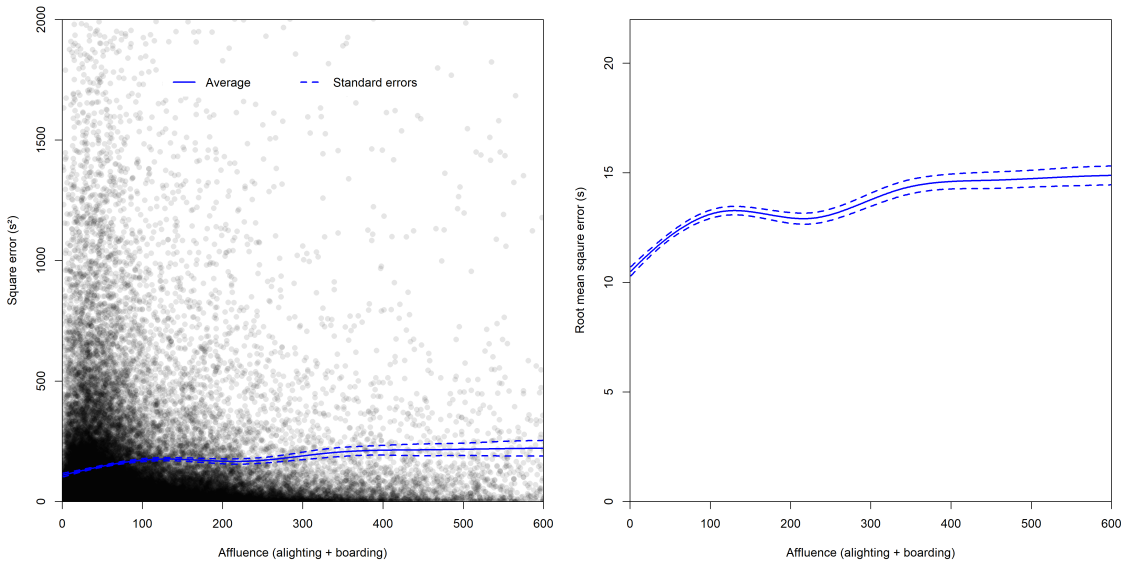


FIGURE 3.8 – Left graph : scatterplot of the squared errors on the test set for dwell time modeling by RF–All plotted against passenger affluence, together with an estimation of the associated average squared errors (solid line), and standard errors thereof (dotted lines). Right graph : root of the curves obtained in the left graph, corresponding to root mean square errors by passenger affluence.

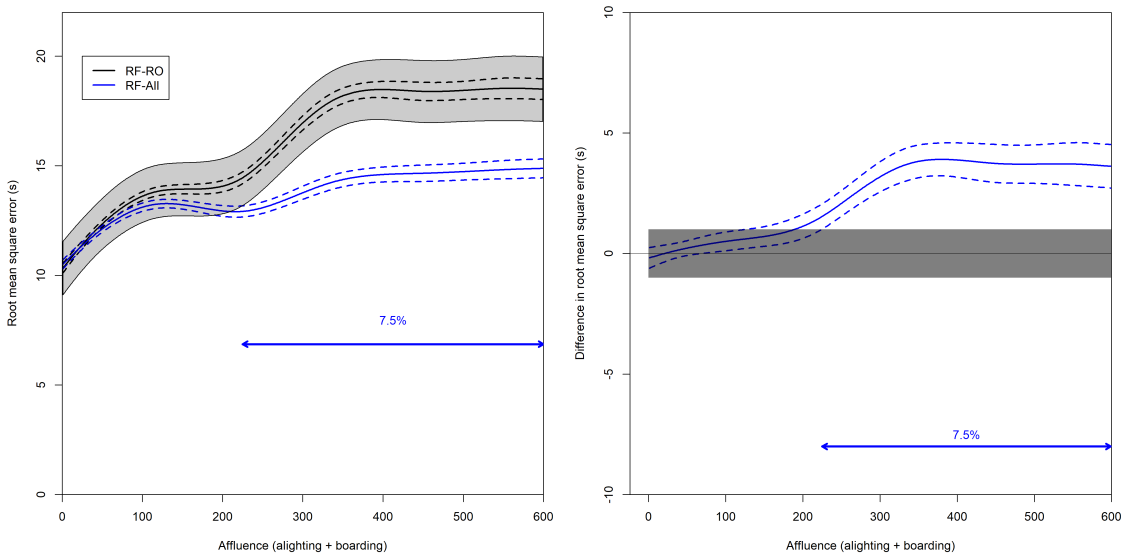


FIGURE 3.9 – Left graph : root mean square errors for the modeling of dwell time (y–axis) by passenger affluence (x–axis) for RF–RO (black) and RF–All (blue). Right graph : difference of these root mean square errors, between RF–RO and RF–All. The horizontal arrows indicate the data ranges where significant improvements are achieved on average by RF–All over RF–RO ; the percentages below the arrows are the corresponding data shares.

as (k, s, d) varies in the test set $\mathcal{T}$est. Average squared errors by passenger affluence and their associated ± 2 standard deviations may then be computed, exactly as in the left graph of Figure 3.6. The right graph of Figure 3.8 depicts the roots of the curves computed in the left graph of Figure 3.8 ; these root curves depict root mean square errors by the passenger affluence, associated with measures of deviations. The left graph of Figure 3.9 provides such root curves for RF–RO (together with a ± 1 s tube) and RF–All, while the right graph of Figure 3.9 is the difference between these curves, in the RF–RO minus RF–All direction.

Comments on local performance by passenger affluence (Figures 3.6–3.9). These figures generally show that the improvement in the dwell time modeling from RF–RO to RF–All, i.e., when taking PF and M variables into account, lies in situations with a high passenger affluence. These account for a limited share of the situations considered : around 5 to 7% of them. Yet, these are exactly the situations where the modeling of dwell time is challenging, as can be seen from the relatively large average errors made by the reference model RF–RO. In particular, the left graph of Figure 3.9 shows that RF–All enjoys a more steady performance, while the one of RF–RO worsens as passenger affluence increases.

Comments on local performance by observed dwell time (Figure 3.10). Figure 3.10 depicts how modeling errors vary with the observed dwell time y^{obs}. Errors for both RF–RO and RF–All methods follow U-shaped curves, with a low plateau in the 40 s – 90 s range, and with linear increases outside of this range. The RF–All method outperforms significantly the RF–RO method on average for extreme values of the observed dwell time : short dwell times (inferior to 22 s) and long dwell times (larger than 110 s). These values only account for 3 to 5% of all observations. No scheduled dwell time is shorter than 30 s, so, observed dwell times shorter than 22 s must correspond to trains with a delay and low passenger affluence, that attempt to leave the station as early as possible. This hints at the necessity of a study of local performance by deviation to scheduled arrival time, which we provide next. We have no convincing or consistent explanations for the improvements for longer dwell times.

Comments on local performance by deviation to scheduled arrival time (Figure 3.11). Figure 3.11 depicts how modeling errors vary with the deviation Δa to scheduled arrival time. As for Figure 3.10, errors for both RF–RO and RF–All methods follow U-shaped curves with a minimum reached at $\Delta a = 0$, i.e., when trains are perfectly on time. On the first part of the U-shaped curves, i.e., for early trains, the performance of RF–RO and RF–All exhibit is virtually indistinguishable. This is certainly explained by the fact that early trains wait longer than needed in a station ; therefore, passenger flows do not constrain dwell times. On the contrary, on the second part of the U-shaped curves, i.e., for late trains, RF–All consistently outperforms RF–RO, in a statistically significant manner as soon as Δa is larger than something of the order of 70 s, which accounts for 14% of the total observations in

TABLE 3.9 – Breakdown of the data set for line L by regimes of punctuality or passenger affluence.

Punctuality	Early : 54,697	On time : 34,388	Late : 28,242
Passenger affluence	Low : 58,813	High : 58,514	

MAE (and 7% in RMSE). The rough improvement in performance lies is of order 2 – 3 s, for errors of the order of 10 – 15 s. These observations thus show that for delayed trains, passenger flows are a key determinant of dwell time, as intuition commands : trains attempt to leave a station as fast as possible when they are late on schedule.

3.4.3 Breakdown of the performance by regimes of punctuality or passenger affluence

The local performance study above highlights the situations where taking passenger flows into consideration helps (i.e., where RF–All is superior to RF–RO) : in case of large delays to scheduled arrival time or high passenger affluence. We now clarify further these determinants by looking at their joint influence : we break down the performance by regimes of punctuality or passenger affluence.

The regimes of punctuality considered were already explained in Table 3.3 : trains may be early, on time, or late. We define two regimes of passenger affluence, high and low, setting as threshold a quasi-median of the passenger affluences $A_{k,s,d}+B_{k,s,d}$ observed on the test data set : we use 51–52 as thresholds (low passenger affluence is passenger affluence $\leqslant 51$ and high passenger affluence is passenger affluence $\geqslant 52$). All in all, the 117,327 triplets (k, s, d) of the test data set may be broken down as indicated in Table 3.9.

In this section, we follow somewhat the structure of the previous analysis and first report global numerical results factored by regimes of punctuality or regimes of passenger affluence (a table), and second, provide a more local, graphical, idea of the improvement in performance of RF–All over RF–RO factored by regimes of punctuality or passenger affluence. Finally, we analyse the importance of variables by regimes of punctuality , i.e., pin point the variables that are the most important to model dwell time for each regime of punctuality.

Global performance by regimes of punctuality or regimes of passenger affluence. Table 3.10 deepens the results of the fourth line of Table 3.8, which was devoted to random forests : the first line of Table 3.10 is a mere copy of the fourth line of Table 3.8. We then break down the performance achieved on the test data set by regimes of punctuality, i.e., compute the errors only over early trains, trains on time, or late trains. The first line of Table 3.10 is therefore a weighted average of its second, third, and fourth lines.

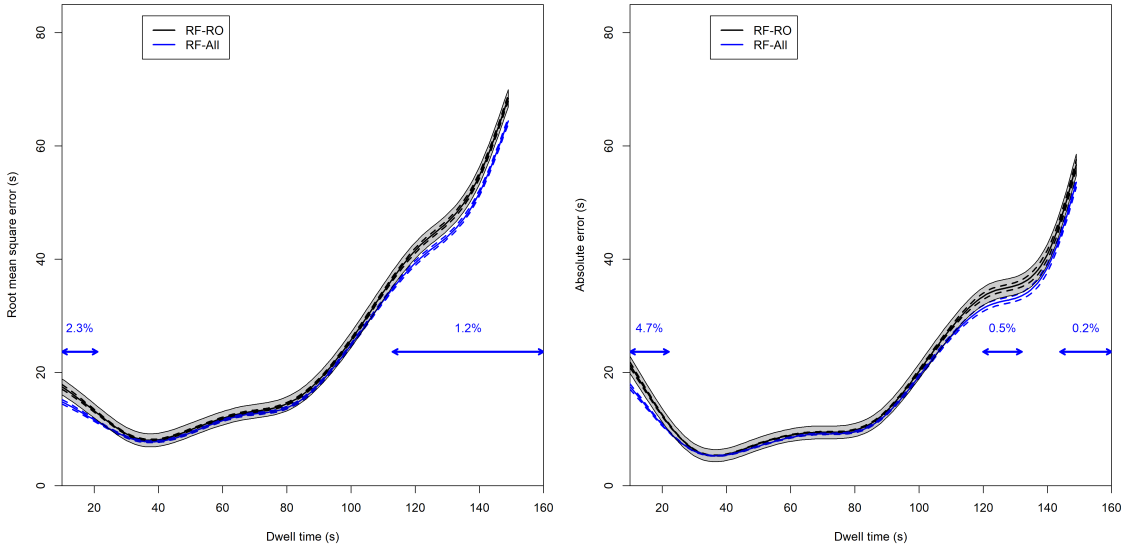


FIGURE 3.10 – Root mean square errors (left graph) and absolute errors (right graph) for the modeling of dwell time (y–axis) by observed dwell time (x–axis) for RF–RO (black) and RF–All (blue). The horizontal arrows indicate the data ranges where significant improvements are achieved on average by RF–All over RF–RO ; the percentages below the arrows are the corresponding data shares.

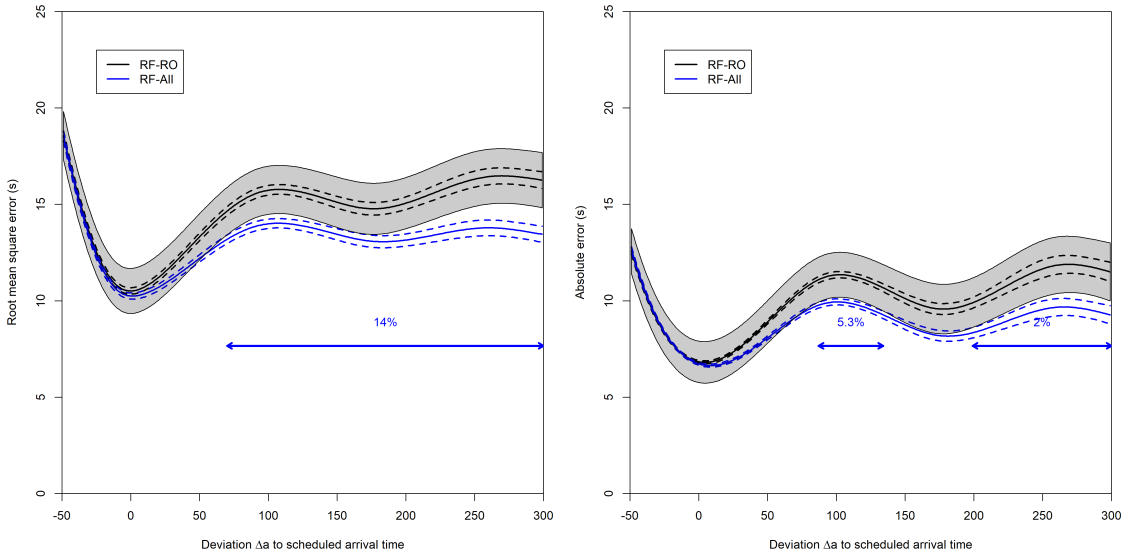


FIGURE 3.11 – Root mean square errors (left graph) and absolute errors (right graph) for the modeling of dwell time (y–axis) by deviation Δa to scheduled arrival time (x–axis) for RF–RO (black) and RF–All (blue). The horizontal arrows indicate the data ranges where significant improvements are achieved on average by RF–All over RF–RO ; the percentages below the arrows are the corresponding data shares.

TABLE 3.10 – Modeling performance for random forests by regimes of punctuality or regimes of passenger affluence (lines) and for each subset of variables (rows) ; see the legend of Table 3.8), in MAE (*left part of the table*) and RMSE (*right part of the table*). Standard errors are smaller than 0.03 seconds.

Random forests	MAE				RMSE			
	PF	RO	RO PF	RO PF+M	PF	RO	RO PF	RO PF+M
All trains	13.7	8.4	8.1	8.0	18.8	12.9	12.5	12.3
Early trains	15.4	7.9	8.0	7.9	21.0	12.5	12.6	12.4
Trains on time	11.8	8.3	7.8	7.8	16.2	12.4	11.9	11.7
Late trains	12.7	9.7	8.5	8.5	17.2	14.2	12.9	12.7
Low passenger affluence	12.4	7.6	7.5	7.5	16.9	11.4	11.3	11.4
High passenger affluence	15.0	9.2	8.7	8.5	20.5	14.3	13.5	13.1

We first comment the influence of the regime of punctuality on performance. The smallest errors are always observed for trains on time, then for early trains, while the largest errors are suffered for late trains. The influence of the subsets of variables considered is similar to what was observed already in Table 3.8 : RO variables in isolation are more useful that PF variables in isolation, while the simultaneous consideration of RO and PF variables is even better, with the consideration of the critical door data (subset M) not changing substantially performance.

For regimes of passenger affluence, similar observations may be issued concerning the subsets of variables, noting however that the added value of PF variables on top of RO variables is larger in the case of a high passenger affluence than for a low passenger affluence. Generally speaking, dwell time is more difficult to predict in situations of high passenger affluence, as intuition commands.

Local performance by passenger affluence factored by regimes of punctuality (Figure 3.12). On Figure 3.12, we break down by regimes of punctuality the differences in modeling errors between RF–RO and RF–ALL as functions of the passenger affluence already depicted in the right graphs of Figure 3.7 and 3.9. Therein, we had noticed a significant reduction of the errors in 5% to 7.5% of the cases. We observe that these cases all feature late trains or trains on time (in particular, that a significant reduction of the error is observed for none of the early trains). Also, the obtained improvements are larger for late trains than for trains on time and/or take place for lower values of passenger affluence.

Local performance by dwell time factored by regimes of punctuality (Figure 3.13). Figure 3.13 is the counterpart of Figure 3.10, where we broke

down the differences in modeling errors as functions of the observed dwell times by regimes of punctuality. In this case as well, RF–RO and RF–All obtain the same performance on early trains (which are the only ones with observed dwell times larger than 100 s). Improvements in performance are mostly due to late trains, for which between 20% and 30% of the data points are better modeled ; improvements due to trains on time are negligible. These improvements take place, as in Figure 3.10, in a U-shaped fashion, for small and large values of the observed dwell times.

Local performance by dwell time factored by passenger affluence regime (Figure 3.14). In Figure 3.14 we represent the differences in modeling errors between RF–RO and RF–ALL as functions of the dwell time, with a break down by regimes of passenger affluence. As in Figures 3.10 and 3.13 we observe improvements of RF–All over RF–RO for short (inferior to 20 s) or long (larger than 90 s) dwell times. But interestingly, there is a clear association of regimes : improvements for short dwell times only take place in the case of low passenger affluence, while improvements for long dwell times happen only in the case of high passenger affluence.

Local performance by deviation to scheduled arrival time factored by passenger affluence regime (Figure 3.15). We represent in Figure 3.15 the differences in modeling errors between RF–RO and RF–ALL as a function of the deviation Δa to scheduled arrival time and break it down by regimes of passenger affluence. This figure is the counterpart of Figure 3.11, where we observed modest improvements of RF–All over RF–RO for deviations larger than something of the order of 50 s. Figure 3.15 shows that these modest improvements are all associated with high affluence.

3.4.4 Most influential variables

We recall a general methodology to determine which variables are the most influential in a random-forest modeling, explain how we implemented it on our data set, and discuss the obtained results, with a special focus on the identification of the most influential variables by regimes of punctuality.

General methodology. Two main family of methodododologies exist to determine which explanatory variables are the most influential for random forests on a given data set : mean decrease accuracy [MDA] and mean decrease impurity [MDI]. Each of them may be implemented in several specific ways despite a common spirit for each methodology proposed by Breiman [2001]. We discuss below the specific implementations provided by the R package `ranger` already mentioned in Section 3.3.3 (see Wright and Ziegler, 2017), corresponding to the options

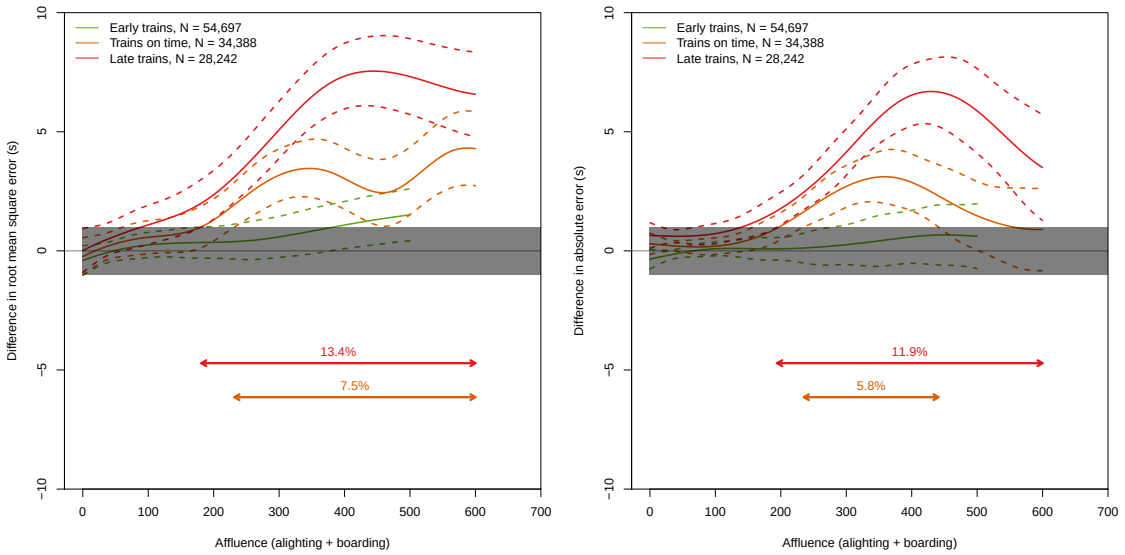


FIGURE 3.12 – Differences in mean square errors (left graph) and absolute errors (right graph) between RF–RO and RF–All for the modeling of dwell time (y–axis) by passenger affluence (x–axis), factored by regimes of punctuality. Positive numbers correspond to the superiority of RF–All over RF–RO.

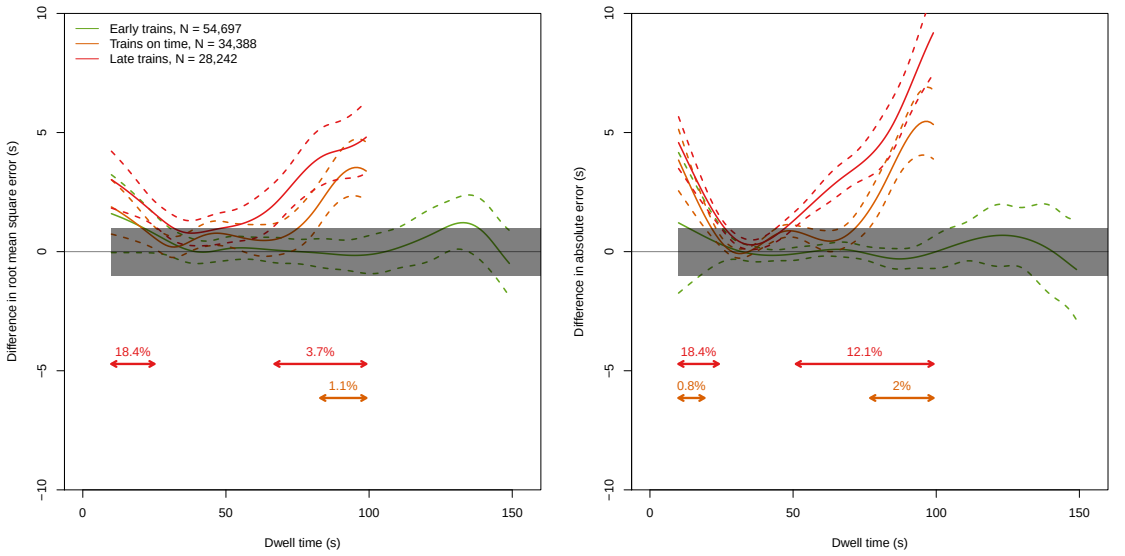


FIGURE 3.13 – Differences in mean square errors (left graph) and absolute errors (right graph) between RF–RO and RF–All for the modeling of dwell time (y–axis) by observed dwell time (x–axis), factored by regimes of punctuality. Positive numbers correspond to the superiority of RF–All over RF–RO. The horizontal arrows indicate the data ranges where significant improvements are achieved on average by RF–All over RF–RO ; the percentages below the arrows are the corresponding data shares.

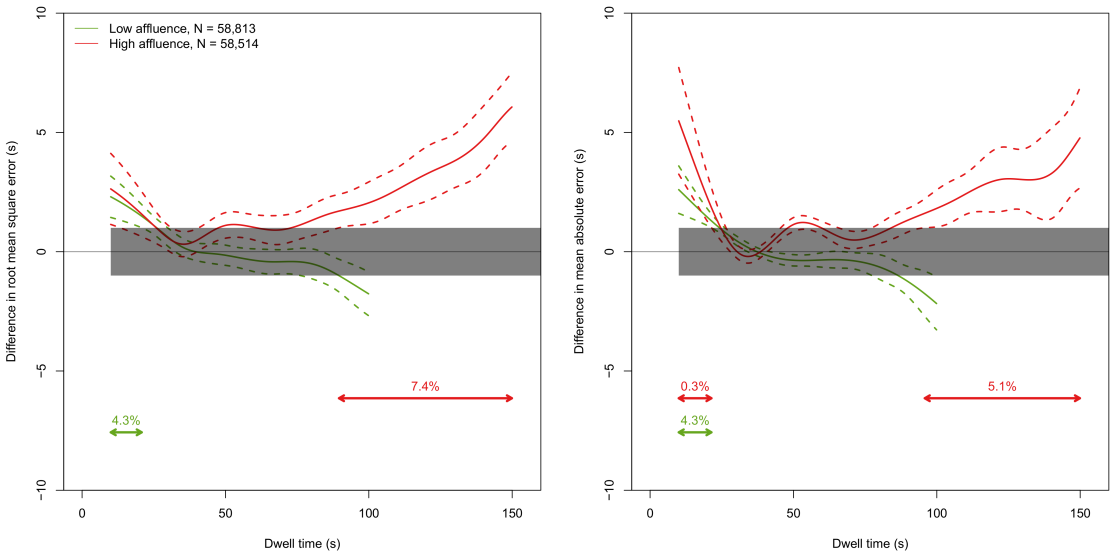


FIGURE 3.14 – Differences in mean square errors (left graph) and absolute errors (right graph) for the modeling of dwell time (y–axis) by observed dwell time (x–axis), factored by passenger affluence regime. Positive numbers correspond to the superiority of RF–All over RF–RO. The horizontal arrows indicate the data ranges where significant improvements are achieved on average by RF–All over RF–RO ; the percentages below the arrows are the corresponding data shares.

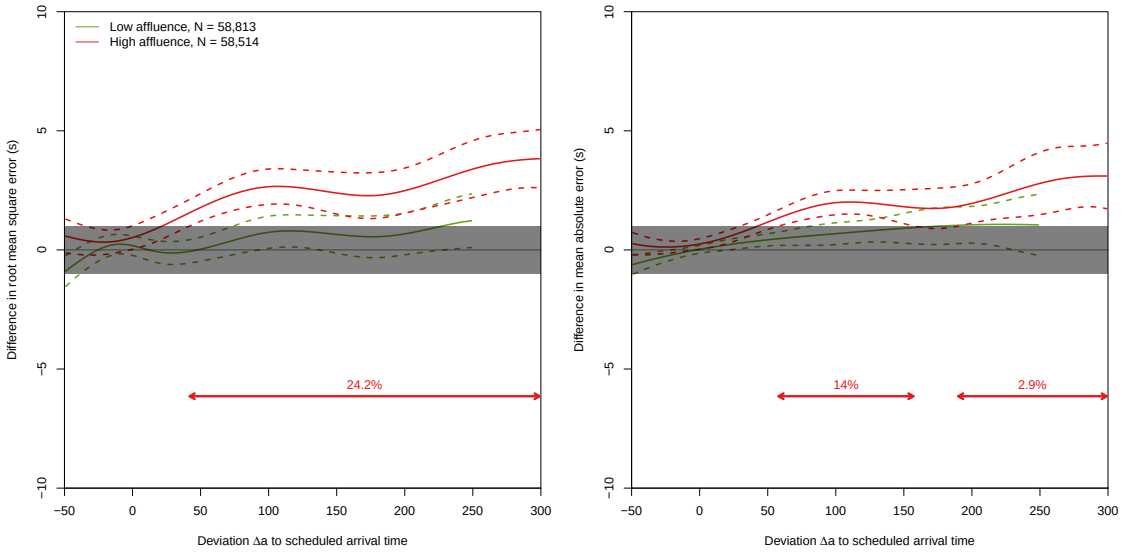


FIGURE 3.15 – Differences in mean square errors (left graph) and absolute errors (right graph) for the modeling of dwell time (y–axis) by deviation Δa to scheduled arrival time (x–axis), factored by passenger affluence regime. Positive numbers correspond to the superiority of RF–All over RF–RO. The horizontal arrows indicate the data ranges where significant improvements are achieved on average by RF–All over RF–RO ; the percentages below the arrows are the corresponding data shares.

`permutation` [MDA] and `impurity` [MDI]. The most popular criterion is probably MDA but we provide here the results obtained for both criteria.

We recall that random forests exploit instances $\boldsymbol{X}_{k,s,d}$ of vectors of variables $\boldsymbol{X} = (X_1, \ldots, X_p)$.

The spirit of MDA is the following : for each variable j, an index MDA_j is computed as follows. We first bootstrap data with replacement into T data sets (where T is the number of trees of Table 3.7) compute a random forest based on each of these T bootstrapped data sets, and evaluate an average difference of performance on the remainder observations of each of these data sets (the so-called out-of-bag observations) : the average squared error on modified out-of-bag observations, obtained by randomly permuting the values of the variable of interest, minus the average squared error on original out-of-bag observations. The larger this average difference MDA_j, the more crucial the variable under scrutiny.

As explained in Section 3.3.3, each tree $f^{(t)}$ of a forest is grown through refinements decided based on maximal reductions of in-sample errors, and MDI exploits this construction : $\text{MDI}_j^{(t)}$ is simply the weighted sum of the reductions associated with the same variable j, over all refinements leading to tree $f^{(t)}$, where the weights are the proportion of observations falling in the region to be refined. The final index MDI_j is then obtained by averaging out the $\text{MDI}_j^{(t)}$ over the T trees of the forest.

In both cases, we obtain non-negative families of indices $(\text{MDA}_j)_{1\leqslant j\leqslant p}$ and $(\text{MDI}_j)_{1\leqslant j\leqslant p}$ and we depict on Figure 3.16 the normalized vectors

$$\frac{\text{MDA}_j}{\sum_{i=1}^{p} \text{MDA}_i} \quad \text{and} \quad \frac{\text{MDI}_j}{\sum_{i=1}^{p} \text{MDI}_i}, \qquad \text{where} \quad j = 1, \ldots, p\,. \tag{3.20}$$

Specific application. The first line of Figure 3.16 provides the normalized MDA and MDI indices for RF–All run on the entire data set ; the 21 variables it relies on (see last column of Table 3.5) are ranked according to their normalized indices.

We also implement RF–All (still tuned with the hyperparameters of the last column of Table 3.7) on each of the three subsets defined by regimes of punctuality ; when we do so, we only feed RF–All with 18 variables, omitting the three binary categorial variables stemming from the regime of punctuality z (given that they would be constant anyway on each of the subsets). The bottom three lines of Figure 3.16 provide the normalized MDA and MDI indices computed for each regime of punctuality.

Results : at a global level. Be it for MDA or MDI indices, the top three influential variables are, globally, the scheduled dwell time y^{theo}, the passenger affluence at the critical door M, and the deviation to scheduled arrival time Δa. Then come the numbers A and B of alighting and boarding passengers, as well as

the crowding factor C and the fact that the train is early, i.e., $z = 1$. The importance of M may seem surprising given the modest differences in modeling performance read in Tables 3.8 and 3.10 : we comment this issue in detail below.

Results : early trains. This picture for early trains is somewhat similar to the picture at the global level, except that y^{theo} and Δa have an importance much superior to other variables : they gather about 60% of the total importance. This is likely to be due to the fact that early trains are supposed to depart at the scheduled time, i.e., as mentioned earlier in Section 3.3.1, after a dwell time equal to $y^{\text{theo}} - \Delta a$. (We recall that $\Delta a < 0$ for early trains.) Thus, we expect that the observed dwell time y^{obs} is close to $y^{\text{theo}} - \Delta a$. Machine-learning techniques like random forests spot this kind of rules in some automatic way, which explains why y^{theo} and Δa are the most two influential variables for early trains. We remind, however, that they do so while providing a single model for all stations, all working days, all hours, and all trains.

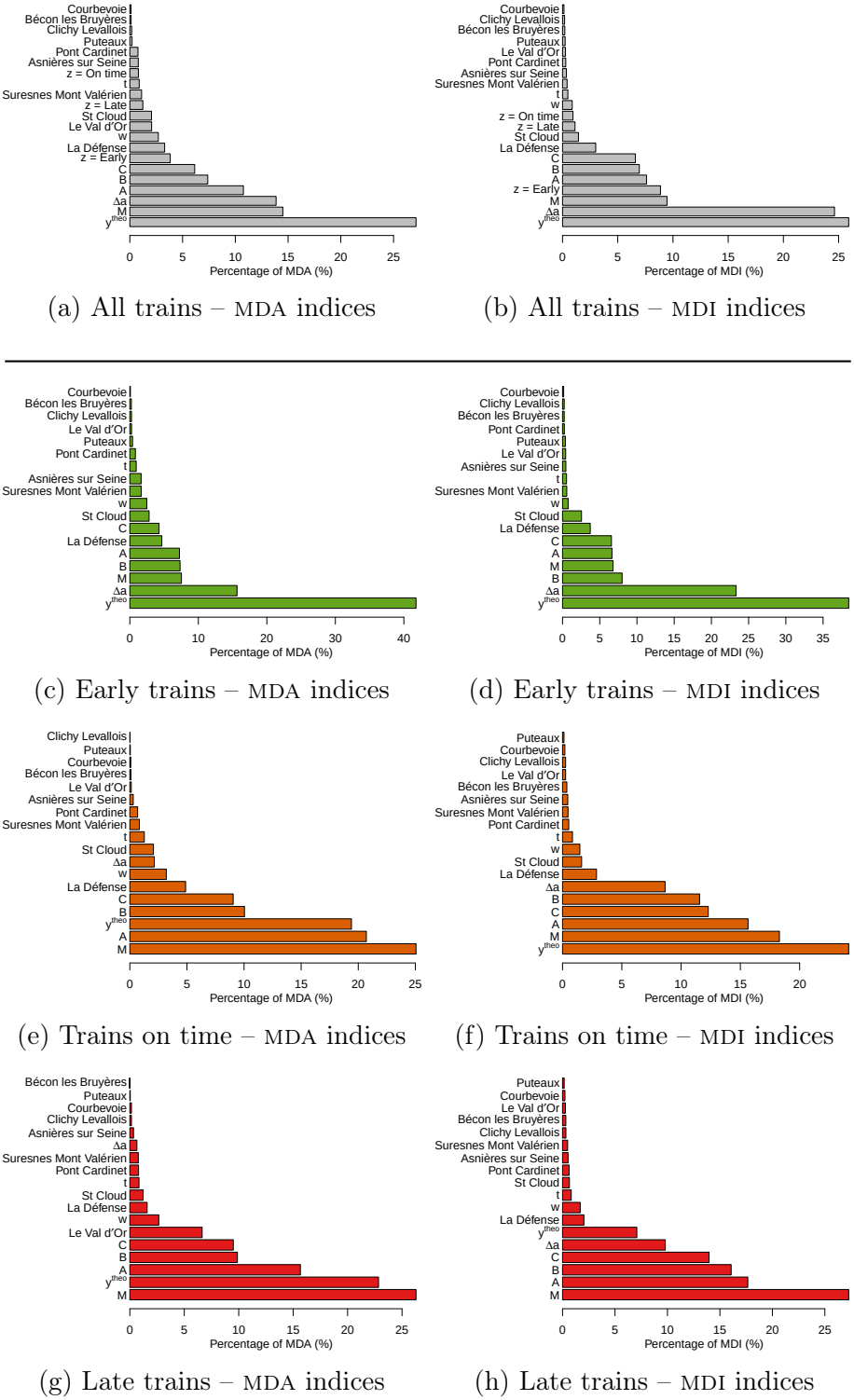
Results : trains on time and late trains. For trains on time and late trains, the MDA procedure rather points to the critical passenger affluence M as the main driving factor, with alighting number A and scheduled dwell time y^{theo} as the next most important variables. As mentioned above, this may seem surprising given the numerical results, where modest overall improvements of about 0.1 s to 0.2 s are achieved with the addition of the M variable. These modest overall improvements however hide (again) significant local improvements in critical situations, most of them related to trains on time and late trains : we noted, when producing the various graphs of Sections 3.4.2 and 3.4.3, that they reported fewer significant improvements in terms of shares of data points concerned when the variable M was omitted, i.e., when RF–[RO+PF], instead of RF–All, was compared to RF–RO. We do not provide further details for the sake of conciseness but wanted to mention this fact, as it explains the "qualitative" importance of M, which the MDA procedure confirms.

Results : stations. In all cases, stations are among the least influential variables, except maybe for La Défense and Saint-Cloud.

3.5 Conclusions and research perspectives

Conclusions. The main findings of our study are the following ones ; they hold for the considered data set of line L and are globally robust with respect to variations on the methodology or of the considered railway line (see Appendix 3.B).

1. Railway operations variables are key for low modeling errors. This being said, on average and at a global level, the consideration of passenger flows variables on top of railway operations variables (only) decreases by about 0.5 s the



(a) All trains – MDA indices (b) All trains – MDI indices

(c) Early trains – MDA indices (d) Early trains – MDI indices

(e) Trains on time – MDA indices (f) Trains on time – MDI indices

(g) Late trains – MDA indices (h) Late trains – MDI indices

FIGURE 3.16 – The most influential variables using normalized MDA indices (left) and normalized MDI indices (right) for RF–All on the entire data set (first line) or RF–All run on sub-data sets corresponding to the regimes of punctuality (last three lines).

modeling error on the observed dwell time based on mere railway operations variables.

2. However, the consideration of these passenger flows variables locally improves this modeling error in critical situations (while never deteriorating performance in non-critical situations) by sometimes up to 5 – 10 s on average : most notably, for late arrivals or for dense situations (when passenger affluence is large).
3. More generally, on this data set, railway operations variables are the most influential variables for early trains (which are constrained by the scheduled departure time and must wait possibly for an extended amount of time) while passenger flows variables are the most influential variables for late trains (which leave the station right after the passenger exchange time), and also, for trains on time. These phenomena were expected, of course, but are confirmed on data.
4. Method-wise, we discussed fully automated model-building techniques (in particular, thanks to setting their hyperparameters on data). Among them, we favored random forests but note that a closed-form linear regression model with multiplicative effects (by stations, ways and regimes of punctuality—trains that are early, on time, or late), that is also fully data driven, obtains a global modeling performance that is only slightly worse.

Discussion : alternative summaries of passenger flows variables. While we could prove the existence of an added value for passenger flows variables, there is still some room for study to determine the most efficient (effective and concise) formulation for these variables. Our rich data set includes the door-by-door numbers A^i and B^i of passengers alighting and boarding, where i ranges between 1 and I. We chose not to use all these $2I$ variables but summarized them into the total numbers A and B of passengers alighting and boarding the train (i.e., we summed up the A^i and B^i over the doors i) and also considered the maximum of their sums, $M = \max\{A^i + B^i : i = 1, \ldots, I\}$. That is, we summarized the $2I$ original variables into three variables A, B, and M only. We did so for the sake of interpretability of the models built. However, the choices made to rely on three variables only were somewhat arbitrary : to the least, the impact of these choices should be explored. The main concern (mentioned by an anonymous reviewer) is how the critical door is taken into account. It seems intuitive that the impact of passenger flows is determined by the passenger exchanges at the critical door. Now, the survey by Kuipers et al. [2021a] points out that, for a given door i, identical values of the sum $A^i + B^i$ will lead to different passenger exchange times. Typically, an entirely one-directional passenger flow is fastest ; then come equally balanced flows, while uneven flows tends to be more turbulent and thus lead to longer exchange times. It is therefore even unclear how to define the critical door $i^\star$, and when this is achieved, it would probably be wiser not to only consider the sum $A^{i^\star} + B^{i^\star}$ but the individual variables $A^{i^\star}$ and $B^{i^\star}$ instead, hoping that the machine-learning methods would combine $A^{i^\star}$ and $B^{i^\star}$ in some nonlinear fashion if this is relevant. We leave this issue for follow-up studies.

Other research perspectives. A main research perspective is to now provide forecasts of the dwell time. The models studied in this chapter rely on information (passenger exchange numbers, deviation to scheduled arrival time) that are unknown in advance but could be predicted, possibly in simple ways : the deviation to the scheduled arrival time expected at future stations equals the deviation to the scheduled departure time suffered at the present station, for instance, while passenger exchange numbers could be predicted by some average values. Doing so, and using one of the models built on historical data, we would obtain real-time predictions of the dwell time at future stations, that would get updated each time the considered train leaves a station.

Other research perspectives lies in drawing conclusions on the models built in terms of designs : design of the timetables or design of the platforms of the stations. More precisely, the model could be fed with observed (joint) distributions of deviations to scheduled arrival time and passenger flows to simulate distributions of dwell times and better design the timetables through setting a careful but possibly shorter buffer time (see Figure 3.1), or studying the effect of adding trains during peak hours. Also, the role and importance of the critical door on dwell time could be better understood, so as to draw conclusions in terms of physical design of the platforms, if needed.

All in all, we provided methods to output modelings of the dwell time, but these models should be extended to predictive models, or should be used for simulation and design purposes.

3.A Details on hyperparameters (a.k.a. tuning parameters) of machine-learning methods

Section 3.3.5 briefly explains that machine-learning methods need to pick some hyperparameters on the train data set—two per method, which we indicated in Table 3.6—and fit a model on the same train data set based on these hyperparameters.

In this appendix, we first illustrate through a sensitivity analysis (Section 3.A.1) that the selection of these hyperparameters is not crucial ; we then nonetheless describe in detail the cross-validation methodology we used to perform this selection (Section 3.A.2).

3.A.1 Sensitivity analysis

We illustrate in Figures 3.17–3.19 how tuning parameters affect performance, both in RMSE and MAE, when taking all variables (RO, PF and M ones) into account ; similar conclusions are reached for subsets of variables. On this figures, we report the performance obtained on the test data set by fitting models on the train data set based on each pair of tuning parameters of Table 3.6. The overall conclusion is that many pairs of tuning parameters lead to an approximately equal performance, and that these pairs consist of large enough parameters, while some parameters that are too small may lead to suboptimal performance. We conclude that the selection of tuning parameters is not a crucial issue.

More precisely, the sensitivity of random forests is illustrated in Figure 3.17 ; for clarity, we represent only a part of all possible pairs (m, T) of tuning parameters. Pairs with numbers of variables $m \geqslant 6$ and numbers of trees $T \geqslant 50$ obtain basically the same performance. The stability of performance is even more remarkable for feed-forward neural networks, as can be seen on Figure 3.19 : all pairs exhibit a performance that lies in a range of radius of order $\pm 0.5s$. There is slightly more instability for gradient boosting with regression trees, even though taking a large number of trees (several hundreds) eventually equalizes all performance ; see Figure 3.18.

3.A.2 Automatic selection of tuning parameters through 5-fold cross-validation

Even if the performance is not (much) sensitive to the choice of tuning parameters, we alleviate the burden of users by providing a fully automated procedure to select these parameters. This procedure uses two passes on the train data set : in the first pass (Steps 1 and 2 of Figure 3.20), it selects the tuning parameters through a 5-fold cross-validation estimation of performance in generalization (more detail are

provided below). In the second pass (Step 3 of Figure 3.20), it fits the model on the entire train data set based on the selected tuning parameters. We do so to avoid over-fitting issues on the train data set : selecting the best tuning parameters on the train data set by comparing the performance of models fit on the entire train data set is prone to biases ; indeed, with such a procedure, we would be comparing some in-sample errors rather than out-of-sample errors, which is what we need. Put differently, this simpler procedure would evaluate the respective performance of tuning parameters in too optimistic a way.

The first pass is a 5-fold cross-validation estimation of performance which consists of separating the train data set in a random partition with 5 folds (i.e., in 5 random non-overlapping subsets), fitting the model on 4 of them and evaluating the obtained performance on the 5th fold. This 5th fold varies, and we average out the five measures of performance obtained to determine the best tuning parameters.

The tuning parameters selected by this two-pass procedure were provided in Table 3.7. There is an important variability in the specific values selected for the pairs of tuning parameters by the subsets of variables considered. However, the performance of a pair of a given cell of Table 3.7 (i.e., for a given pair of a method plus subset of variables) is not even 0.1 s apart from the performance obtained by the pair of another cell in the same line of the table (i.e., for the same method but for a different subset of variables). The seemingly instability of the values of the tuning parameters hides a remarkable stability in the underlying performance, which was already exhibited in the sensitivity analysis of Section 3.A.1.

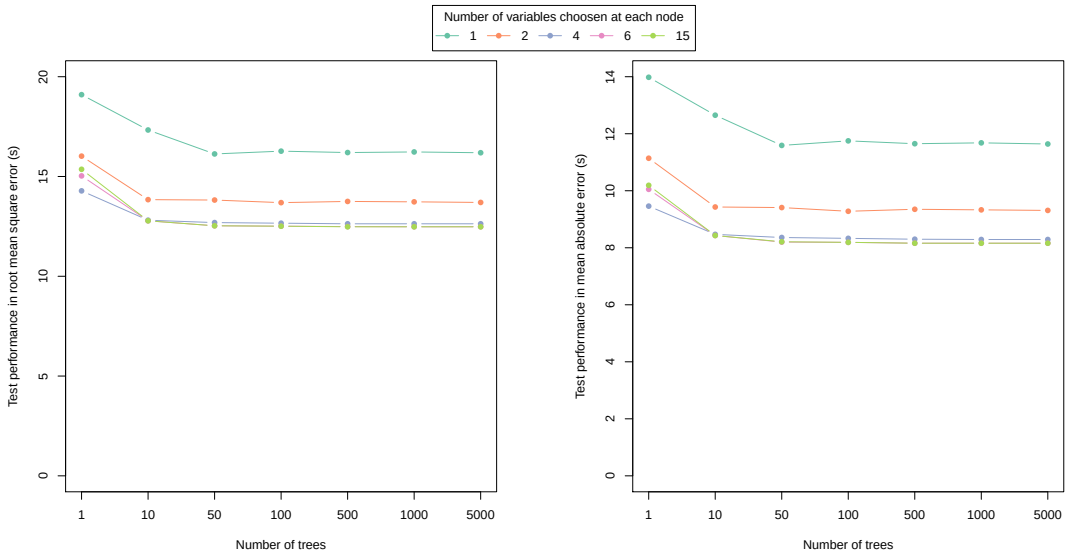


FIGURE 3.17 – Performance of random forests on the test data set for various values of the pair (m, T) of tuning parameters, where m is the number of variables chosen at each split and T is the number of trees in the forest. The left picture measures performance in RMSE and the right picture does so in MAE.

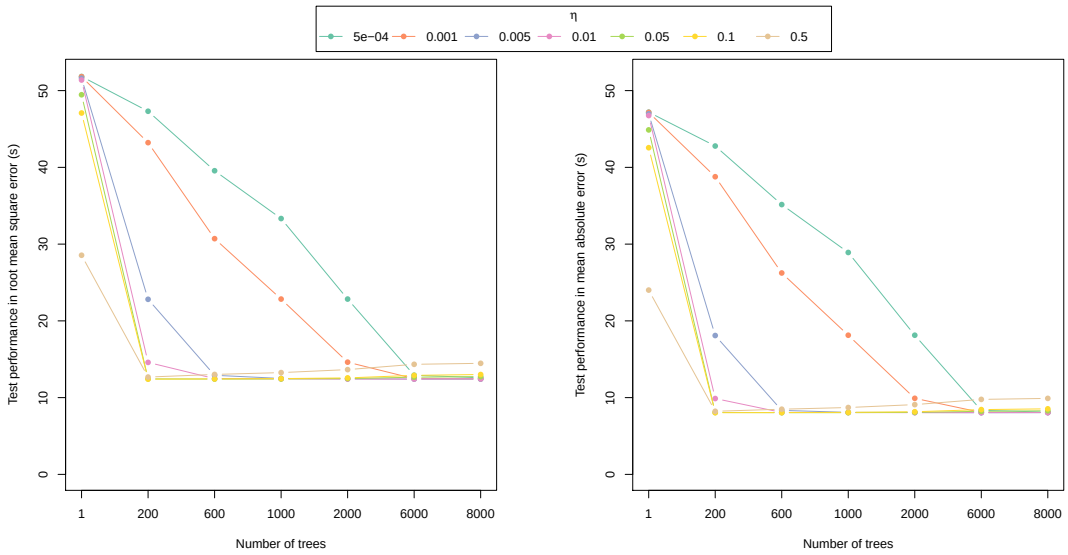


FIGURE 3.18 – Performance of gradient boosting with regression trees on the test data set for various values of the pair (η, T) of tuning parameters, where η is the shrinkage parameter and T is the number of trees. The left picture measures performance in RMSE and the right picture does so in MAE.

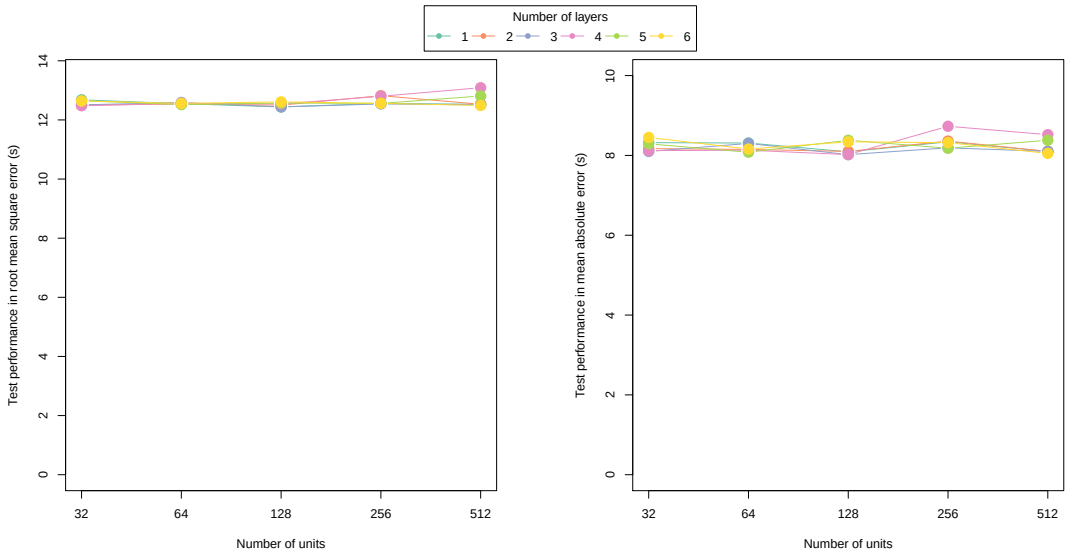


FIGURE 3.19 – Performance of feed-forward neural networks on the test data set for various values of the pair (H, N) of tuning parameters, where H is the number of hidden layers and N is the number of nodes. The left picture measures performance in RMSE and the right picture does so in MAE.

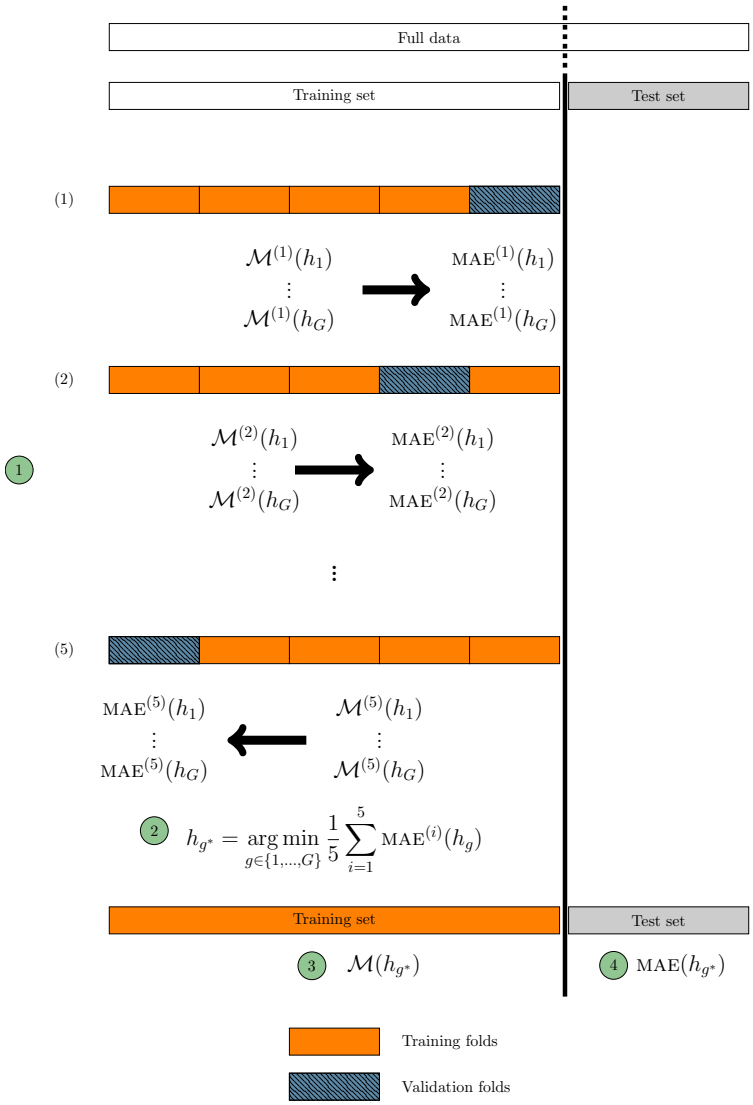


FIGURE 3.20 – Principle of the automated selection procedure proposed, for a given machine-learning method. The full dataset is split into a training set (in orange) and a test set (in grey). The training set is itself split into a random partition consisting of 5 subsets called folds.
(1) For each pair of hyperparameters h_g, the model $\mathcal{M}^{(i)}(h_g)$ is fit on all train data but the one of fold i and the performance $\text{MAE}^{(i)}(h_g)$ is computed for this model on fold i.
(2) The performance in generalization of the method for hyperparameters h_g is estimated by averaging the five errors $\text{MAE}^{(i)}(h_g)$, for $i \in \{1, \dots, 5\}$. We then select the hyperparameters h_{g^*} minimizing $\text{MAE}(h_g)$.
(3) Model $\mathcal{M}(h_{g^\star})$ is fit on the entire train data set.
(4) We compute and report the performance of $\text{MAE}(h_{g^\star})$ of the model $\mathcal{M}(h_{g^\star})$ on the test data set.

3.B Robustness checks

In this section, we discuss the robustness of our results : when the differences $y^{\text{obs}} - y^{\text{theo}}$ to the scheduled dwell time are modeled in lieu of the observed dwell times y^{obs}, still on line L (Section 3.B.1) ; and what happens for the other line considered, line H (Section 3.B.2).

3.B.1 Modeling the differences $y^{\text{obs}} - y^{\text{theo}}$ to the scheduled dwell time

So far, we modeled directly the observed dwell time y^{obs}. We now look instead into the modeling of the deviations $\Delta y = y^{\text{obs}} - y^{\text{theo}}$ to the scheduled dwell time y^{theo}, i.e., we run the methods discussed in Section 3.3 to model Δy (with newly optimized hyperparameters determined by following the methodology described in Section 3.3.5 and Appendix 3.A.2). We obtain modelings $\widehat{\Delta y}$ which we turn into modelings $\hat{y}^{\text{obs}}$ of the observed dwell time by adding y^{theo}.

We observe in Table 3.11 (absolute performance of this modeling of the deviations) and in Table 3.12 (difference in modeling performance between direct modeling of y^{obs} and modeling of the deviations Δy) that the two approaches yield extremely similar results, except when run on the PF variables in isolation, where the modeling of deviations is significantly more efficient. About 1 s of reduction in average modeling errors is gained. This was expected as modeling Δy amounts to using the RO variable y^{theo}, which we identified as a key determinant of the observed dwell time (see, e.g., Section 3.4.4).

In the main body of the chapter, we rather performed a direct modeling of y^{obs} to be able to report some "pure" performance for the PF variables.

3.B.2 Results for line H—in brief

The main body of this chapter considered a specific sub-branch of line L (see Figure 3.2), and we now study what happens for the sub-branch of line H for which data is also available (see also Figure 3.2). This sub-branch features 11 stations on top of the origin and terminus stations. The same time period is considered as for line L and 145,609 triplets (k, s, d) are available. We may define similarly regimes of punctuality and regimes of passenger affluence (based on thresholds $\leqslant 53$ and $\geqslant 54$ passengers), and obtain the breakdown summarized in Table 3.13.

We (re-)optimized the hyperparameters of random forests on this new data set, following the methodology described in Section 3.3.5 and Appendix 3.A.2, and provide in Table 3.14 (the counterpart of Table 3.10) the modeling performance of the dwell time for random forests, globally or by the regimes of punctuality or

passenger affluence.

Table 3.14 confirms that railway operations [RO] variables are more useful for the modeling than passenger flow [PF] variables, with a difference in performance of about 2 s. This was expected. The true confirmation expected was on the improvement of RO and PF variables considered simultaneously (possibly together with the passenger affluence M at the critical door) over RO variables used in isolation : we get a mixed picture of virtually no improvement in most situations, except for late trains and in case of a high affluence, where improvements of about 0.3 s are obtained. The patterns observed are therefore similar to the ones of Table 3.10, but take place with a lower intensity.

Figures 3.22 and 3.23 further illustrate these (more moderate) improvements in the modeling performance : no significant improvements are observed when observations are broken down by passenger affluence regimes (Figure 3.23) while significant improvements are observed only in the case of late trains (Figure 3.22), with a significant worsening for a tiny fraction of the observations is simultaneously observed for early trains. All in all, the existence of improvements only for the most challenging situation of late trains and their smaller intensity may be caused by the absence of peaks of passenger affluence on line H (see Figure 3.21), while such peaks exist for line L.

TABLE 3.11 – Modeling performance for the observed dwell time y^{obs} based on a modeling of the deviation $\Delta y = y^{\text{obs}} - y^{\text{theo}}$: results are formatted as in Tables 3.8 and 3.10, with standard errors still smaller than 0.03 seconds.

Methods	MAE				RMSE			
	PF	RO	RO PF	RO PF+M	PF	RO	RO PF	RO PF+M
1. Linear regression with additive effects	12.4	10.5	10.2	10.1	16.9	14.8	14.5	14.3
2. Linear regression with a multiplicative effect of Δa by z	12.4	9.1	8.9	8.8	16.9	13.6	13.2	13.1
3. Linear regression with multiplicative effects by triplets	12.2	8.8	8.3	8.3	16.7	13.2	12.6	12.5
4. Random forests	12.5	8.5	8.1	8.0	17.1	13.1	12.4	12.3
5. Gradient boosting with regression trees	12.0	8.5	8.0	7.9	16.5	13.0	12.4	12.2
6. Feed-forward neural networks	11.9	8.5	7.9	7.9	16.4	13.0	12.4	12.3
Random forests								
Early trains	15.4	7.9	8.0	7.9	21.0	12.5	12.6	12.4
Trains on time	11.8	8.3	7.8	7.8	16.2	12.4	11.9	11.7
Late trains	12.7	9.7	8.5	8.5	17.2	14.2	12.9	12.7
Random forests								
Low passenger affluence	12.4	7.6	7.5	7.5	16.9	11.4	11.3	11.4
High passenger affluence	15.0	9.2	8.7	8.5	20.5	14.3	13.5	13.1

TABLE 3.12 – Differences in modeling performance between Table 3.11 (based on a modeling of the deviation $\Delta y = y^{\text{obs}} - y^{\text{theo}}$) and Tables 3.8 and 3.10 (based on a direct modeling of y^{obs}). Negative numbers indicate a more accurate modeling in Table 3.11.

Methods	MAE				RMSE			
	PF	RO	RO PF	RO PF+M	PF	RO	RO PF	RO PF+M
1. Linear regression with additive effects	−1.3	0.0	0.0	0.0	−1.5	0.0	0.0	0.0
2. Linear regression with a multiplicative effect of Δa by z	−1.3	0.0	0.0	0.0	−1.5	0.0	0.0	0.0
3. Linear regression with multiplicative effects by triplets	−1.1	0.0	0.0	0.0	−1.3	0.0	0.0	0.0
4. Random forests	−1.2	0.1	0.0	0.0	−1.7	0.2	−0.1	0.0
5. Gradient boosting with regression trees	−0.9	0.0	0.0	0.0	−1.4	0.0	0.0	0.0
6. Feed-forward neural networks	−0.8	0.1	−0.1	−0.1	−1.0	0.0	0.0	0.1
Random forests								
Early trains	−1.8	0.0	−0.1	0.0	−2.3	0.2	−0.1	0.0
Trains on time	−0.8	0.0	0.0	0.0	−1.0	0.1	0.0	0.0
Late trains	−0.7	0.2	0.1	0.0	−0.9	0.3	0.1	0.0
Random forests								
Low passenger affluence	−1.1	0.1	−0.1	0.0	−1.5	0.2	0.0	0.0
High passenger affluence	−1.4	0.1	0.0	0.0	−1.8	0.2	0.0	0.0

TABLE 3.13 – Breakdown of the data set for line H by regimes of punctuality or passenger affluence.

Punctuality	Early : 33,448	On time : 58,755	Late : 53,406
Affluence	Low : 72,795	High : 72,814	

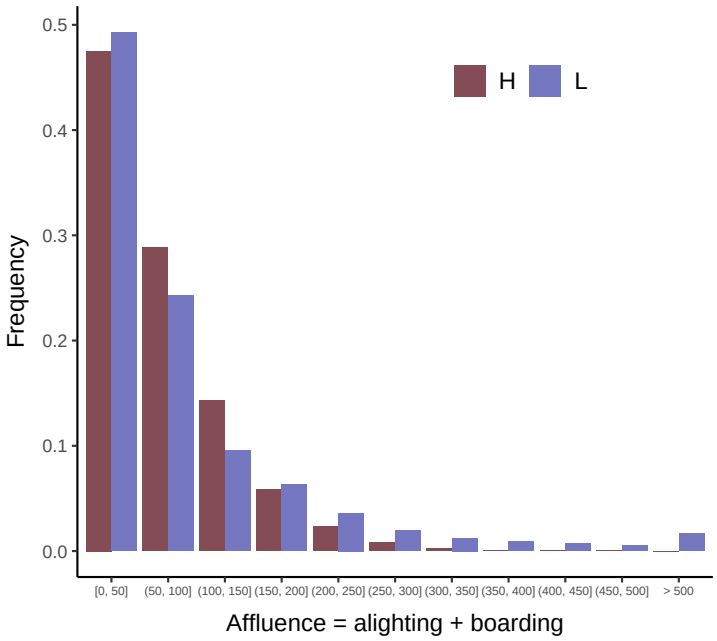


FIGURE 3.21 – Histograms of passenger affluence (numbers of passengers alighting and boarding, for all stations and all trains considered) for lines H and L.

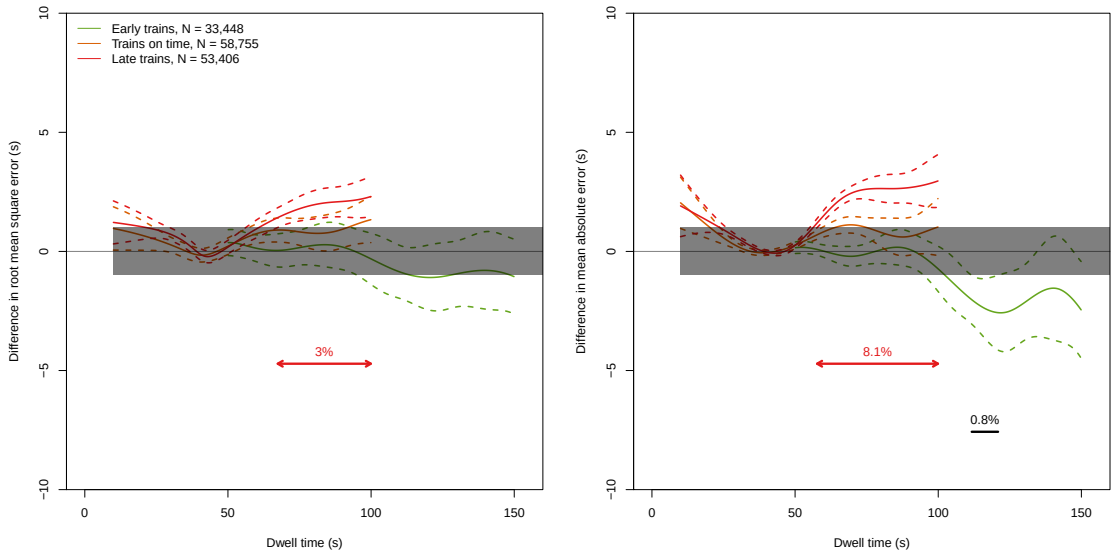


FIGURE 3.22 – Differences in mean square errors (left graph) and absolute errors (right graph) between RF–RO and RF–All for the modeling of dwell time (y–axis) by observed dwell time (x–axis), factored by regimes of punctuality. Positive numbers correspond to the superiority of RF–All over RF–RO. The horizontal arrows indicate the data ranges where significant improvements are achieved on average by RF–All over RF–RO ; the percentages below the arrows are the corresponding data shares.

TABLE 3.14 – Modeling performance of the dwell time for random forests by regimes of punctuality or regimes of passenger affluence (lines) and for each subset of variables (rows) ; see the legend of Table 3.8), in MAE (*left part of the table*) and RMSE (*right part of the table*). Standard errors are smaller than 0.03 seconds.

Random forests	MAE				RMSE			
	PF	RO	RO PF	RO PF+M	PF	RO	RO PF	RO PF+M
All trains	10.6	8.0	7.9	7.8	15.1	11.9	11.7	11.6
Early trains	13.2	7.8	7.9	7.9	19.2	11.9	11.9	11.9
Trains on time	9.6	7.8	7.7	7.6	13.4	11.6	11.5	11.4
Late trains	10.1	8.4	8.1	8.0	13.8	12.3	12.0	11.8
Low passenger affluence	9.9	7.5	7.4	7.4	13.7	10.9	10.8	10.8
High passenger affluence	11.3	8.5	8.4	8.2	16.4	13.0	12.6	12.5

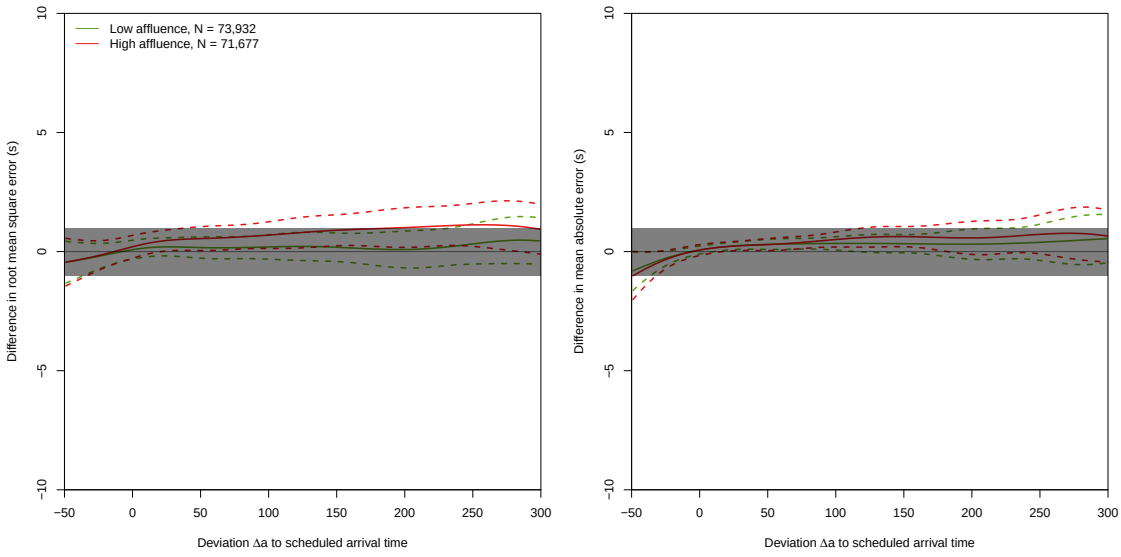


FIGURE 3.23 – Differences in mean square errors (left graph) and absolute errors (right graph) between RF–RO and RF–All for the modeling of dwell time (y–axis) by deviation Δa to scheduled arrival time (x–axis), factored by regime of passenger affluence. Positive numbers correspond to the superiority of RF–All over RF–RO.

4

Modèles de prévision à court terme

Nous précisons en quoi la prévision à court terme des variables d'exploitation ferroviaire et de flux de voyageurs est un enjeu en science des transports. Nous formalisons une méthode de prévision unique pour 5 variables à partir de la grille horaire. Ce modèle bi-directionnel exploite pour chaque arrêt (k, s) le passé des précédents arrêts d'un train k et celui des trains précédents à une gare s. Nous proposons deux simplifications, l'une à partir de la notion de motifs répétés dans la grille horaire, l'autre à l'aide d'une stratégie de sélection automatique d'un voisinage optimal.

Contents

4.1 Introduction

Le modèle des temps de stationnement du Chapitre 3 est adapté à tous les régimes de ponctualité des trains : en avance, à l'heure ou en retard. Cependant, la modélisation des temps de stationnement utilise des variables explicatives, mesurées en temps réel, mais inconnues pour les prochains arrêts. Ainsi, un tel modèle, pour pouvoir être utilisé en opérationnel, doit être alimenté par des prévisions. En parallèle, les canaux de distribution de l'information aux voyageurs ont aussi besoin de prévision notamment des retards ou de l'affluence à bord aux prochains arrêts pour les communiquer aux voyageurs. Ces deux objectifs : gérer en opérationnel les temps de stationnement et informer les voyageurs, justifient le développement dans ce chapitre d'une stratégie unifiée de prévision à court terme (à l'horizon d'une gare) des variables de flux de voyageurs et de retard sur un réseau de transport. Le modèle de ce chapitre est générique grâce à l'utilisation de la stabilité du plan de transport (grille horaire, respect de l'horaire, etc.). Le modèle n'utilise que les réalisations d'un passé proche afin de ne pas tomber dans le même travers que le modèle du Chapitre 3. Ce modèle est suffisamment générique pour pouvoir être testé sur la prévision des temps de stationnement directement.

L'introduction de ce chapitre débute par une présentation des notations utiles à la prévision à court terme, s'en suit une mise en évidence de nos contributions vis à vis de l'état de l'art. Nous concluons par une rapide présentation du plan du chapitre.

4.1.1 Notations

Nous notons X une variable aléatoire générique représentant les variables de flux de voyageurs [PF] (nombre B de montées, nombre A de descentes, charge à bord L) ou d'exploitation ferroviaire [RO] (retard à l'arrivée[1] ΔA, temps de stationnement T)[2]. Ces variables aléatoires sont observées à chaque arrêt d'un train k à une gare s. Nous faisons l'hypothèse que les réalisations sont indépendantes et identiquement distribuées suivant les jours d. Pour résolument dissocier les indices du modèle de ceux des répétitions, nous notons $X^d_{k,s}$: la réalisation de $X_{k,s}$ le jour d. Cette indexation est différente de celle du Chapitre 3 : $X_{k,s,d}$. Un modèle de prévision à court terme repose sur l'exploitation d'un passé proche. Pour chaque arrêt (k, s), le passé proche provient soit des P trains précédents desservant la même gare s, soit des Q arrêts précédents pour le même train k. Le premier est appelé P-voisinage en gare, noté $X_{(k-P):(k-1),s} = \big(X_{(k-P),s}, \ldots, X_{k-1,s}\big)$. Le second est appelé Q-voisinage en train, noté $X_{k,(s-Q):(s-1)} = \big(X_{k,(s-Q)}, \ldots, X_{k,s-1}\big)$. Un voisinage en forme de L d'une profondeur P, Q désigne l'ensemble des réalisations pour les P trains précédents à la gare s et les Q arrêts précédents du train k. Les arrêts non desservis ne sont pas pris en compte dans le calcul du voisinage. Les notations présentées ci-dessus, et celles introduites par la suite sont résumées dans les Tables 4.1–4.4.

1. Appelé « écart à l'heure d'arrivée théorique »dans le Chapitre 3.
2. Le temps de stationnement est noté T à la place de y^{obs} dans ce chapitre.

TABLE 4.1 – liste et ensemble des indices

k	train
s	gare
d	jour
$\{1, \ldots, S\}$	ensemble des gares
$\{1, \ldots, K\}$	ensemble des trains
$\{1, \ldots, D\}$	ensemble des jours
$\mathcal{N}^d$	ensemble des arrêts (k, s) pour un jour d, défini en Section 4.4.1

TABLE 4.2 – Horaires

Théoriques	
$a_{k,s}^{\text{theo}}$	heure d'arrivée
$d_{k,s}^{\text{theo}}$	heure de départ
Observées	
$a_{k,s}^{\text{obs}}$	heure d'arrivée
$d_{k,s}^{\text{obs}}$	heure de départ

TABLE 4.3 – Variables à prédire pour un train k à la gare s

Flux de voyageurs [PF]	
$L_{k,s}$	charge à bord en sortie de gare
$A_{k,s}$	nombre de descentes
$B_{k,s}$	nombre de montées
Exploitation ferroviaire [RO]	
$\Delta A_{k,s} = a_{k,s}^{\text{theo}} - a_{k,s}^{\text{obs}}$	retard à l'arrivée
$T_{k,s} = d_{k,s}^{\text{obs}} - a_{k,s}^{\text{obs}}$	temps de stationnement

TABLE 4.4 – Voisinage pour une variable générique X

$X_{(k-P):(k-1),s} = (X_{(k-P),s}, \ldots, X_{(k-1),s})$	voisinage en gare de profondeur P
$X_{k,(s-Q):(s-1)} = (X_{k,(s-Q)}, \ldots, X_{k,(s-1)})$	voisinage en train de profondeur Q
$\mathcal{L}_{k,s}^{P,Q} = \left(X_{(k-P):(k-1),s}, X_{k,(s-Q):(s-1)}\right)$	voisinage en forme de L de profondeur P et Q

TABLE 4.5 – Motifs et voisinage pour une variable générique X

M	taille d'un motif
$k[M]$	projection d'un train k sur un motif de taille M, défini en Section 4.3.3
P	profondeur du voisinage en gare
Q	profondeur du voisinage en train

4.1.2 Contributions et état de l'art

TABLE 4.6 – Synthèse des travaux sur la prévision à court terme des variables de flux de voyageurs [PF] et d'exploitation ferroviaire [RO]. $\emptyset$ indique que l'information est inconnue pour l'article en question.

Références	Modes	Variables	Modèles	Horizons	Voisinage
Zhang and Teng [2013]	Bus	A, B, T	Régression linéaire	1 gare	Aucun
Kecman et al. [2015]	Train	ΔA	Chaîne de Markov	20-120 min	Train
Li et al. [2016]	Train	T	Régression linéaire	1 gare	Gare et train
Corman and Kecman [2018]	Train	ΔA	Réseau bayésien	60 min	Gare et train
Jenelius [2019]	Métro	L	Apprentissage automatique	toute la course	Train
Pasini et al. [2019]	Train	L	Apprentissage profond	6 gares	Gare
Noursalehi et al. [2021]	$\emptyset$	A, B, T, L	Choix d'itinéraire	15-30 min	$\emptyset$
Bapaume et al. [2021]	Métro	L	Apprentissage profond	4 gares	$\emptyset$

Notre première contribution concerne la généralisation d'une seule méthode de prévision à court terme pour cinq variables différentes. Dans la littérature en science des transports, les méthodes de prévision à court terme sont souvent développées pour une seule variable. Par exemple, le temps de stationnement pour Li et al. [2016], le retard pour Kecman et al. [2015], Corman and Kecman [2018] ou la charge à bord pour Jenelius [2019], Pasini et al. [2019], Bapaume et al. [2021]. Les seuls auteurs à notre connaissance qui prédisent plusieurs de nos variables sont Zhang and Teng [2013], Noursalehi et al. [2021]. Ils le font en injectant la prévision des montées et des descentes dans le calcul de la charge à bord ou du temps de stationnement. Nous avons décidé d'évaluer les performances de prévision à l'horizon d'une gare, à l'instar de Zhang and Teng [2013], Li et al. [2016], mais cela est seulement un premier pas vers une solution plus opérationnelle, qui nécessiterait des prévisions à plusieurs arrêts, comme le soulignent Bapaume et al. [2021].

Notre seconde contribution est de proposer une représentation originale des données à l'aide d'un graphe gares-trains pour une ligne de transports en commun avec deux branches et des missions non omnibus. La projection d'un graphe espace-temps vers un graphe gares-trains a déjà été proposé par Jenelius [2019], Bapaume et al. [2021], mais pour une ligne de métro omnibus. Corman and Kecman [2018] ont également proposé une approche adaptée à une ligne avec plusieurs branches, cependant leur modèle repose sur un grand nombre de paramètres que nous discuterons par la suite.

Notre troisième contribution est un équilibre entre performance et parcimonie. Certains modèles de la littérature sont parfois trop parcimonieux, ils perdent de l'information en ne prenant pas en compte la gare s ou le train k dans la modélisation [Li et al., 2016]. D'autres, à l'inverse reposent sur un très grand nombre de paramètres, comme Corman and Kecman [2018] qui proposent une modélisation plus complexe que le *modèle non stationnaire*, défini en Section 4.3.2,

où chaque arrêt a un ensemble de paramètres dédié. Grâce à l'identification de motifs répétés dans la grille horaire, nous proposons une simplification de ce modèle, où les coefficients dépendent de la localisation d'un train sur un motif de taille M ($k[M]$, s). L'approche proposée, comme celle de Corman and Kecman [2018], est mal adaptée aux déviations fortes du plan de transport ; la méthode est donc moins flexible que celle de Li et al. [2016], qui cependant négligent une grande partie de l'information disponible.

Notre quatrième contribution est de proposer une méthode pour sélectionner le meilleur voisinage par arrêt pour chaque variable. À notre connaissance, une telle étude systématique du meilleur voisinage n'a pas encore été effectuée pour aucune de nos variables. Seul Jenelius [2019] propose un modèle de régression pénalisée Lasso permettant de sélectionner le meilleur voisinage en train pour la prévision de la charge à bord.

Il convient de relever que certains modèles plus complexes *ad hoc* comme les modèles d'apprentissage profond de Pasini et al. [2019] et de Bapaume et al. [2021], sont aussi une solution pour la prévision à court terme. Cependant, ces solutions ne sont pas en adéquation avec notre objectif d'un modèle parcimonieux. Dans la Table 4.6, nous résumons les principaux articles qui ont inspiré ce chapitre. Notons indépendamment de cette table que la prévision de retard à court terme a une littérature plus riche que la prévision à court terme des flux de voyageurs.

4.1.3 Plan du chapitre

Nous présentons d'abord dans la Section 4.2 la structure et la qualité des données Transilien employées. Nous développons ensuite dans la Section 4.3 les principales méthodes de prévision à court terme utilisées dans le chapitre. Nous présentons les résultats quantitatifs et qualitatifs de ces méthodes dans la Section 4.4. Nous concluons le chapitre et proposons plusieurs perspectives de recherche intéressantes pour les prochaines années dans la Section 4.5.

4.2 Structure et qualité des données Transilien

Nous étudions des données provenant des trains de la ligne H de Transilien, qui circulent depuis les gares de Pontoise et Montsoult-Maffliers jusqu'à la gare terminus parisienne de Gare du Nord. Le schéma de la portion de ligne étudiée est présenté sur la Figure 4.1, les différentes branches de la ligne ont des couleurs distinctes. Nous commençons par montrer en Section 4.2.1 que l'ordre de la grille horaire théorique définit une structure de graphe. À partir de cette structure, nous définissons pour chaque arrêt un voisinage utile pour la prévision à court terme. Nous étudions dans la Section 4.2.2, la cohérence des voisinages observés par rapport à ceux théoriques de la grille horaire. Enfin, dans la Section 4.2.3, nous calculons le taux de couverture associé à chacune des variables.

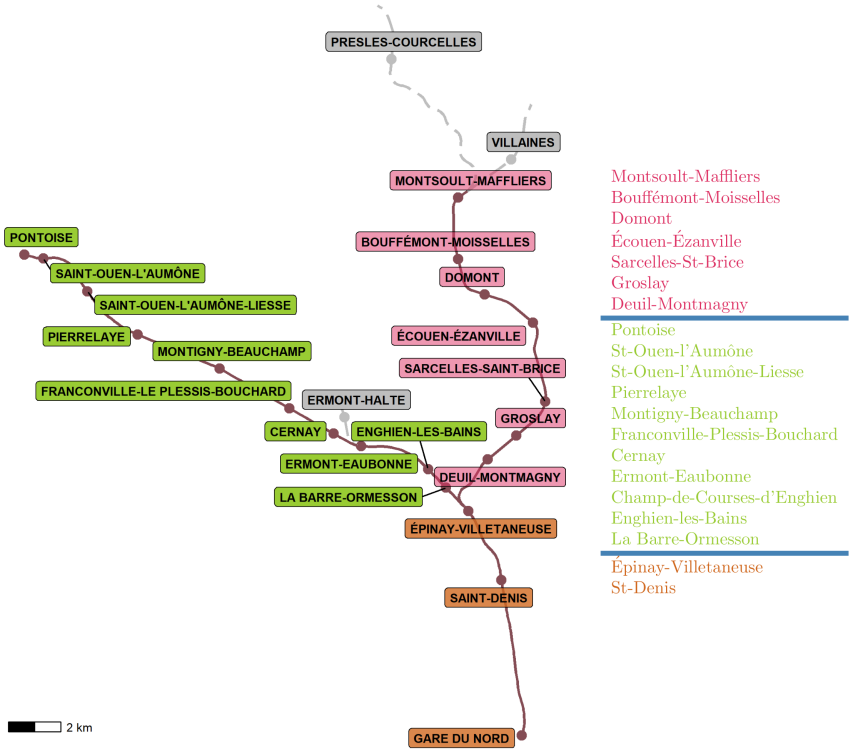


FIGURE 4.1 – Schéma du réseau de la ligne H composée de deux branches (verte au départ de Pontoise, rose au départ de Montsoult-Maffliers) qui se rejoignent à Épinay-Villetaneuse pour former une seule branche (marron) au niveau des trois dernières gares. Les gares non étudiées sont en gris. La colonne de droite indique la liste des gares (sans la gare terminus Gare du Nord car le temps de stationnement, le nombre de montées et la charge à bord n'existent pas pour cette gare).

Nom du train Notes à consulter	APOR	APOR	APOR	APOR	ADDO LV	APOR SD	ADDO LV	ADDO LV	APOR SD	ADDO LV	ADDO LV	APOR SD	ADDO LV
Pontoise	04:36	05:06	05:36	06:06	06:22	06:36	06:37	06:52	07:06	07:07	07:22	07:36	07:37
Saint-Ouen l'Aumône	04:39	05:09	05:39	06:09	06:24	06:39	06:39	06:54	07:09	07:09	07:24	07:39	07:39
Saint-Ouen l'Aumône Liesse	04:42	05:12	05:42	06:12	06:28	06:42	06:43	06:58	07:12	07:13	07:28	07:42	07:43
Pierrelaye	04:45	05:15	05:45	06:15	06:31	06:45	06:46	07:01	07:15	07:16	07:31	07:45	07:46
Montigny Beauchamp	04:49	05:19	05:49	06:19	06:35	06:49	06:50	07:05	07:19	07:20	07:35	07:49	07:50
Franconville Plessis Bouchard	04:52	05:22	05:52	06:22	06:39	06:52	06:54	07:09	07:22	07:24	07:39	07:52	07:54
Cernay (Val d'Oise)	04:55	05:25	05:55	06:25	06:42	06:55	06:57	07:12	07:25	07:26	07:41	07:55	07:56
Ermont-Eaubonne	04:58	05:28	05:58	06:28	06:45	06:58	07:00	07:15	07:28	07:30	07:45	07:58	08:00
Champ de Courses d'Enghien	05:00	05:30	06:00	06:30	|	07:00	|	|	07:30	|	|	08:00	|
Enghien les Bains	05:03	05:33	06:03	06:33	06:49	07:03	07:04	07:19	07:33	07:34	07:49	08:03	08:04
La Barre Ormesson	05:05	05:35	06:05	06:35	|	07:05	|	|	07:35	|	|	08:05	|
Epinay Villetaneuse	05:07	05:37	06:07	06:37	06:53	07:07	07:07	07:22	07:37	07:37	07:52	08:07	08:07
Saint-Denis	05:11	05:41	06:11	06:41	06:57	07:11	07:11	07:26	07:41	07:41	07:56	08:11	08:11
Gare du Nord (Surface)	05:17	05:47	06:17	06:47	07:03	07:17	07:18	07:33	07:47	07:48	08:03	08:17	08:18

Nom du train Notes à consulter	ADDO LV	APOR SD	ADDO LV	ADDO LV	APOR SD	ADDO LV	ADDO LV	APOR	APOR	APOR	APOR	APOR	APOR
Pontoise	07:52	08:06	08:07	08:22	08:36	08:37	08:52	09:06	09:36	10:06	10:36	11:06	11:36
Saint-Ouen l'Aumône	07:54	08:09	08:09	08:24	08:39	08:39	08:54	09:08	09:38	10:08	10:38	11:08	11:38
Saint-Ouen l'Aumône Liesse	07:58	08:12	08:13	08:28	08:42	08:43	08:58	09:12	09:42	10:12	10:42	11:12	11:42
Pierrelaye	08:01	08:15	08:16	08:31	08:45	08:46	09:01	09:15	09:45	10:15	10:45	11:15	11:45
Montigny Beauchamp	08:05	08:19	08:20	08:35	08:49	08:50	09:05	09:18	09:48	10:18	10:48	11:18	11:48
Franconville Plessis Bouchard	08:09	08:22	08:24	08:39	08:52	08:54	09:09	09:22	09:52	10:22	10:52	11:22	11:52
Cernay (Val d'Oise)	08:11	08:25	08:26	08:41	08:55	08:56	09:11	09:25	09:55	10:25	10:55	11:25	11:55
Ermont-Eaubonne	08:15	08:28	08:30	08:45	08:58	09:00	09:15	09:27	09:57	10:27	10:57	11:27	11:57
Champ de Courses d'Enghien	|	08:30	|	|	09:00	|	|	09:30	10:00	10:30	11:00	11:30	12:00
Enghien les Bains	08:19	08:33	08:34	08:49	09:03	09:04	09:19	09:32	10:02	10:32	11:02	11:32	12:02
La Barre Ormesson	|	08:35	|	|	09:05	|	|	09:34	10:04	10:34	11:04	11:34	12:04
Epinay Villetaneuse	08:22	08:37	08:37	08:52	09:07	09:07	09:22	09:37	10:07	10:37	11:07	11:37	12:07
Saint-Denis	08:26	08:41	08:41	08:56	09:11	09:11	09:26	09:40	10:10	10:40	11:10	11:40	12:10
Gare du Nord (Surface)	08:33	08:47	08:48	09:03	09:17	09:18	09:33	09:47	10:17	10:47	11:17	11:47	12:17

FIGURE 4.2 – Grille horaire de l'heure de pointe du matin des trains de Pontoise à Paris pour l'année 2021. Chaque cellule du tableau indique l'heure de départ d'un train, avec un type de mission (indiqué en haut), à une gare donnée (à gauche). L'indication LV ou SD précise si la mission circule du lundi au vendredi ou seulement le samedi et dimanche.

4.2.1 Grille horaire comme structure de graphe

La grille horaire définit les horaires théoriques de passage en gare des trains. La Figure 4.2 est un exemple de grille horaire pour les trains de banlieue de la ligne H. Nous remarquons que les grilles horaires, des jours de semaine (LV) et des jours de week-end (SD) sont différentes. Nous nous concentrons sur une des zones critiques, en l'occurrence, celle des trains vers Paris aux heures de pointe du matin pendant les jours de semaine.

Projection sur un graphe gares-trains

La grille horaire de la Figure 4.2 concerne seulement une branche de la ligne H, celle reliant Pontoise à Paris. Les trains y sont ordonnés par heure de départ à Pontoise. Une grille similaire est disponible pour l'autre branche. Dès lors, puisqu'il n'y a pas de dépassement possible entre les trains sur cette ligne, nous fusionnons les deux grilles. Sur la Figure 4.3, l'absence de dépassement sur la branche commune permet de ré-indexer le graphe gares-temps (à gauche) en un graphe gares-trains (à droite). Cette opération, appelée *projection*, consiste à transformer le graphes gares-temps des deux branches en un unique graphe gares-train. Nous ordonnons les trains, venant des deux branches, suivant leur heure de passage à la gare centrale d'Épinay-Villetaneuse. Dans la représentation tabulaire du graphe gares-trains, chaque cellule non grisée du graphe est un arrêt (k, s) d'un train k à la gare s. Ces arrêts forment les nœuds du graphe de la grille horaire. Nous associons à chaque nœud du graphe une variable aléatoire $X_{k,s}$, où la variable X représente génériquement l'une de nos variables. Par exemple, sur le graphe à droite de la Figure 4.3, les couleurs des cellules représentent la valeur moyenne du nombre $b_{k,s}$ de montées par arrêt. La Gare du Nord, terminus de la ligne, n'est pas représentée sur la Figure 4.3 car la plupart des variables, en l'occurrence le temps de stationnement, le nombre de montées et la charge à bord, n'y sont pas définies.

Voisinage sur un graphe gares-trains

Le graphe gares-trains permet non seulement de définir des nœuds mais il permet également de définir des voisins pour chaque chaque nœud. Ainsi, pour chaque nœud, nous définissons son passé, à la fois en gare (horizontalement, à gare fixée) et en train (verticalement, à train fixé). Le passé considéré a une plus ou moins grande profondeur horizontale, notée P, ou verticale, notée Q. La Figure 4.4 représente le passé le plus profond en train (en rose) et en gare (en bleu) du nœud violet, un passé moins profond étant indiqué avec ces mêmes couleurs, mais foncées. Le passé proche définit le voisinage. La grille horaire a une périodicité en train, au sens où un même motif de missions se répète tous les cinq trains. Par la suite, nous appellerons voisinage standard, un voisinage en train et en gare de profondeur trois arrêts.

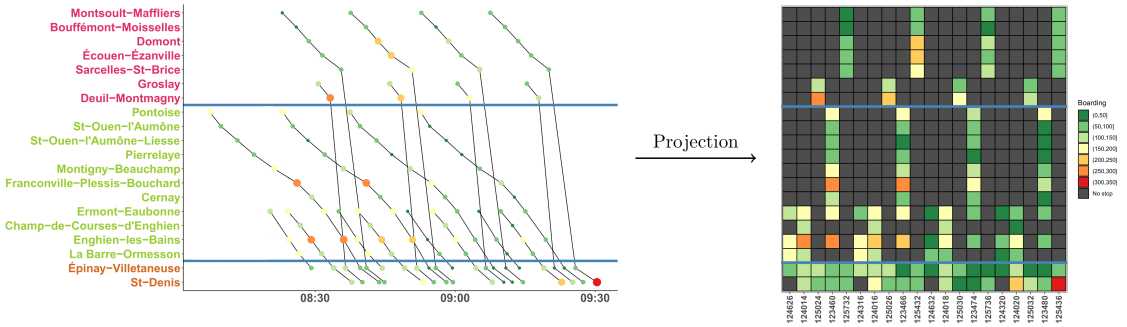


FIGURE 4.3 – Projection d'un graphe gares-temps (graphique de gauche) dans une grille horaire pour former un graphe gares-trains (graphique de droite). Les couleurs et la taille des cercles sur le graphique de gauche et la couleur du graphique de droite représentent le nombre de montées. Un arrêt d'un train qui ne circule pas sur la branche ou qui ne s'arrête pas à une gare est représenté en gris sur le graphique de droite.

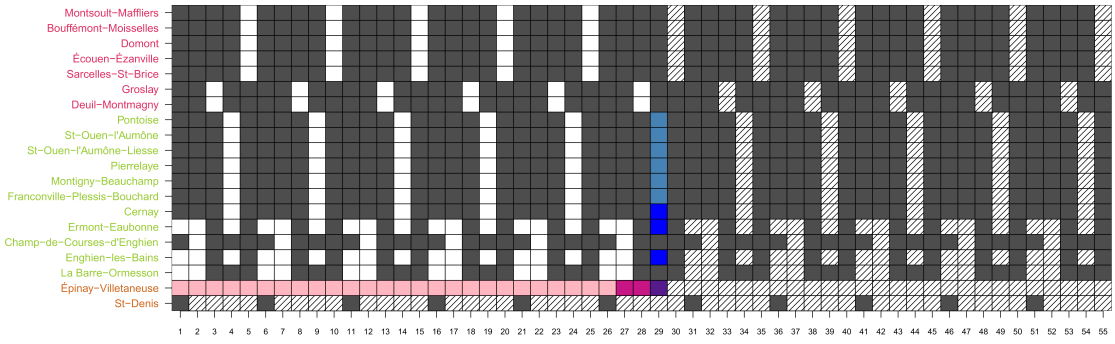


FIGURE 4.4 – Identification de la structure de voisinage pour le graphe gares-trains de l'heure de pointe du matin. Les cases gris foncé indiquent qu'il n'y a pas d'arrêt. Les informations disponibles des gares et des trains précédents, sont respectivement en bleu et en rose. Le bleu ou rose foncés représente les informations les plus récentes. Les cases hachurées représentent des informations futures ou non utilisées pour la prévision.

Graphe gares-trains dans la littérature

Associer chaque variable au graphe gares-trains que nous avons défini à partir de la grille horaire est une proposition originale ; cette construction ne suit ni l'approche usuelle pour prédire les flux de voyageurs, ni celle habituellement utilisée pour prédire les retards. En effet, la prévision des flux de voyageurs est, en général, effectuée en agrégeant les variables par pas de temps de 1 à 15 minutes. Par exemple, Baro and Khouadjia [2021] agrègent par pas de 15 minutes, les données de comptage automatique à bord de la ligne H, Zhong et al. [2016] agrègent les données de validation de Londres, Singapour et Pékin en testant différents pas de temps. Dans notre cas, cette stratégie n'est pas satisfaisante, en effet elle nécessiterait deux étapes [Noursalehi et al., 2021] : une étape de prévision des quantités agrégées, puis une étape d'allocation aux trains des variables prédites. Concernant la prévision des retards, les variables ne sont en général pas agrégées, excepté dans le travail de Ulak et al. [2020]. Cependant, ces variables sont généralement représentées par un graphe temps-événement où chaque noeud est un événement horodaté représentant l'heure d'arrivée, de départ ou de passage d'un train à une gare [Büker and Seybold, 2012, Kecman et al., 2015, Corman and Kecman, 2018]. Ce graphe n'est pas adapté à la prévision des montées ou des descentes qui sont observées sur une arrête du graphe. Nous construirons pour notre part une même structure de graphe pour cinq variables aléatoires où les nœuds sont les arrêts (k, s) et les liens entre les arrêts sont définis par la grille horaire. Cette modélisation est proche de ce que proposent Berger et al. [2011] et Lessan et al. [2019].

4.2.2 Cohérence du voisinage et stabilité de la grille horaire

Afin d'inférer les dépendances entre les nœuds voisins sur la structure que nous avons définie, il est nécessaire que le voisinage pour chaque arrêt soit stable de jour en jour, cependant l'exploitation des transports en commun est sujette à de nombreux aléas. S'ils sont souvent mineurs et ne perturbent pas la structure globale de la grille horaire, certains peuvent être critiques : par exemple, un arrêt ou un dépassement de train non prévus. Nous définissons la notion de cohérence d'un voisinage par rapport à la grille horaire. Un voisinage est cohérent si ses voisins sont ordonnés exactement de la même façon que ceux de la grille horaire. Plus la profondeur du voisinage est grande, moins le voisinage a de chance d'être cohérent. Pour la profondeur maximale, il est très rare que tous les trains de la journée respectent l'ordre de la grille horaire. Par exemple, il n'y a eu que 3 jours cohérents avec une telle profondeur parmi les 124 jours de notre jeu de données. L'objectif est donc de garantir la cohérence d'un voisinage standard pour chaque arrêt. On rappelle qu'un voisinage standard est de profondeur de trois arrêts ($P = Q = 3$). Cette profondeur est intéressante pour la modélisation tout en limitant les pertes d'observations dues à des incohérences. En effet, nous excluons de l'étape d'estimation et de prévision les observations dont le voisinage est incohérent. Li et al. [2016], Pasini et al. [2019] ne vérifient pas de telles contraintes, car ils définissent un unique modèle, quelque soit l'arrêt, en utilisant

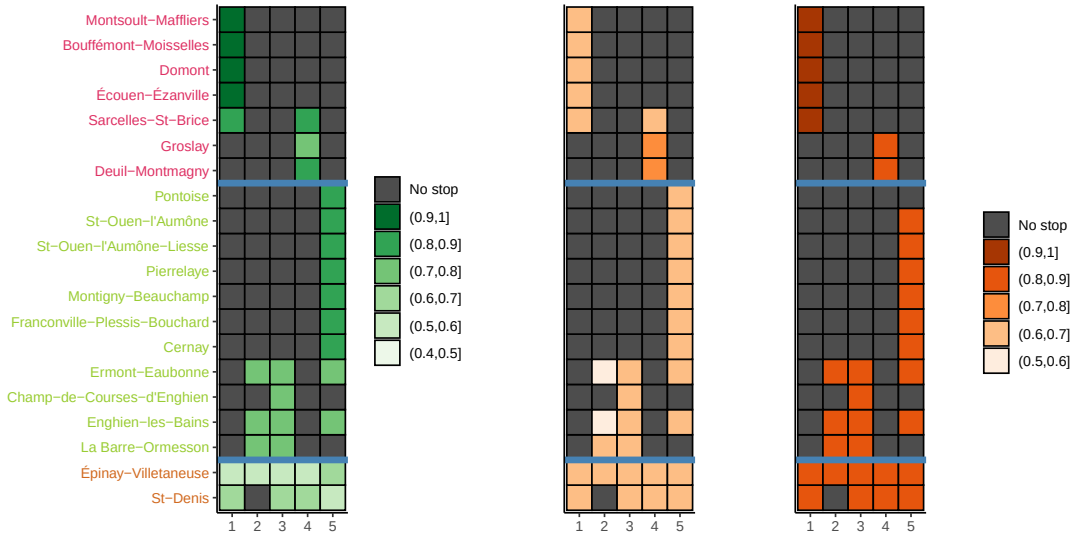


FIGURE 4.5 – Proportions moyennes des situations où le voisinage observé est cohérent avec celui de la grille horaire (en vert, à gauche). Taux de couverture moyens (en orange) pour les variables de flux voyageurs (au centre) et pour les variables d'exploitation ferroviaire (à droite).

le voisinage qu'il soit cohérent ou non. Par la suite, nous posons l'hypothèse que le voisinage local d'un arrêt est connu et stable de jour en jour. Si Corman and Kecman [2018] supposent connus les arrêts et leurs voisinages pour l'heure à venir, nous considérons de notre côté, que l'ordre de passage des trains est fixe. Dans la Figure 4.5 à gauche, nous calculons le pourcentage de situations où le voisinage est cohérent, pour la grille horaire de la Figure 4.4. Ce graphe montre que l'ordre est très souvent cohérent lorsque les circulations sont peu denses (en moyenne dans 80 % des situations pour les gares loin de Paris), mais plus rarement quand les circulations sont denses (en moyenne dans 40 à 60 % des situations pour les gares proches de Paris). Toutefois dans les zones denses, le volume de trains et donc d'arrêts est suffisamment grand pour compenser la perte d'observations liée à l'incohérence des voisinages standards. À noter que la cohérence du voisinage ne dépend pas de la variable considérée mais uniquement de l'ordre des trains.

4.2.3 Taux de couverture

Le taux de couverture d'une variable générique x pour chaque arrêt (k, s) est le ratio exprimé en pourcentage, entre le nombre d'observations sans données manquantes, dans le voisinage standard, et le nombre de jours total. En théorie, il y a une observation par jour. Le taux de couverture est de 100 %, s'il n'y a aucune donnée manquante dans le voisinage standard associé à l'arrêt. Le taux de couverture est spécifique à chaque source de données, ainsi les variables de flux de voyageurs [PF] et d'exploitation ferroviaire [RO] ont des taux de couverture différent. On constate sur le graphe central de la Figure 4.5 que le taux de couverture pour les variables de flux de voyageurs est de 60 à 80 %. Les variables

d'exploitation ferroviaire ont quant à elles un taux de couverture sensiblement plus élevée, autour de 80 %, visible sur le graphe de droite de la Figure 4.5.

Résumé de la Section 4.2. La grille horaire permet de définir une structure de graphe gares-trains pertinente pour toutes les variables. Cette structure de graphe permet également de définir un voisinage d'une profondeur donnée. Ce voisinage est suffisamment stable d'un jour à l'autre pour pouvoir être utilisé pour la prévision. Nous excluons de la phase d'estimation et de prévision les observations dont le voisinage est incohérent *i.e.* dont l'ordre des voisins n'est pas cohérent avec l'ordre des voisins de la grille horaire. Les taux de couverture des variables de flux de voyageurs et d'exploitation ferroviaire sont jugés satisfaisants pour estimer les paramètres des modèles, même si le taux de couverture des variables de flux de voyageurs est plus faible.

4.3 Modèles de prévision à court terme

Nous présentons les modèles de prévision à court terme des variables de flux de passagers (le nombre B de montées, le nombre A de descentes, la charge à bord L) et d'exploitation ferroviaire (le retard à l'arrivée ΔA, le temps de stationnement T). On rappelle que la prévision à court terme est définie à l'horizon d'une gare $s+1$ pour un train k à la gare s. On construit un modèle grâce à la stabilité des relations entre variables de jour en jour. Les jours sont donc des répétitions indépendantes d'un même modèle. Les variables explicatives du modèle sont les réalisations passées proches de l'arrêt. Le voisinage en gare d'une profondeur P : $\boldsymbol{X}_{(k-P):(k-1),s}$, et le voisinage en train d'une profondeur de Q : $\boldsymbol{X}_{k,(s-Q):(s-1)}$. Afin d'alléger les notations, nous ne faisons référence au jour d que si nécessaire. Nous définissons le modèle suivant pour chaque arrêt (k,s) de la grille horaire :

$$X_{k,s} = f_{k,s}\big(\boldsymbol{X}_{k,(s-Q):(s-1)}, \boldsymbol{X}_{(k-P):(k-1),s}\big) + \varepsilon_{k,s} \tag{4.1}$$

où $f_{k,s}$ est une fonction quelconque et $\varepsilon_{k,s}$ une variable aléatoire modélisant le bruit. Les paramètres de la fonction et le voisinage varient pour chaque arrêt, donc les modèles sont différents, tout en étant issus d'une stratégie de modélisation commune. Dans cette section, nous présentons d'abord le modèle de Transilien pour la prévision de la charge à bord et du retard à court terme, que nous adaptons aux autres variables. Nous développons ensuite un modèle de prévision inspiré des travaux de Corman and Kecman [2018]. Nous concluons en présentant nos contributions avec un modèle aussi performant que la littérature mais plus parcimonieux.

4.3.1 Stratégies de prévision de Transilien

Nous présentons en détails les stratégies de prévision de la charge à bord et du retard actuellement utilisées par Transilien. Nous les illustrons au travers de la Figure 4.6.

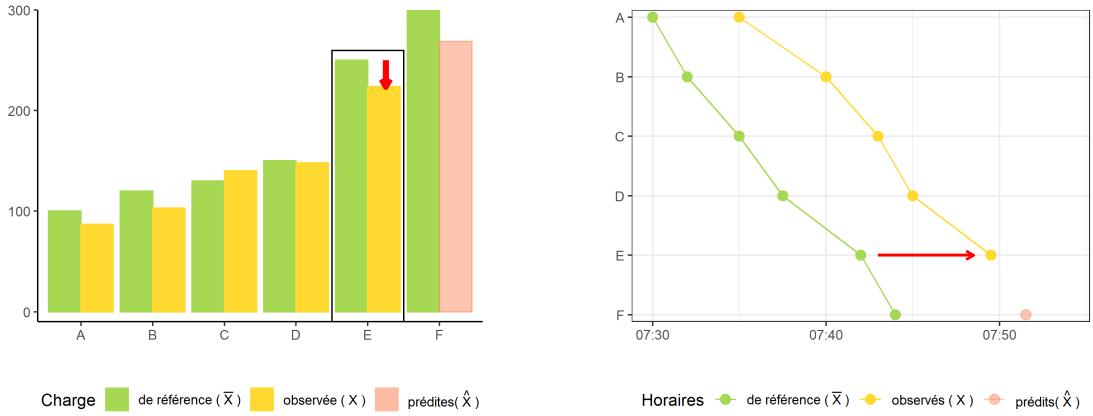


FIGURE 4.6 – Exemple fictif de 6 gares illustrant les stratégies de prévision de Transilien pour prédire la charge à bord (à gauche) et le retard (à droite).

Ces stratégies sont simples et donnent une référence à dépasser. Elles reposent sur l'actualisation de valeurs de référence à partir de l'état actuel du trafic suivant trois étapes : la définition de valeurs de référence, le calcul d'un coefficient d'actualisation et la prévision à partir des valeurs de référence futures.

Étape 1 : définition des valeurs de référence. À Transilien, les charges à bord de référence sont calculées, pour chaque arrêt (k, s) et type de jour (jour de semaine, samedi, dimanche), à partir de la moyenne des charges à bord observées l'année précédente pour cet arrêt et ce type de jour. Dans ce chapitre, nous utilisons un seul type de jour, il n'y a donc qu'une valeur de référence par arrêt (k, s), notée $\overline{L}_{k,s}$. La moyenne des charges à bord est calculée par Transilien à partir des trains qui ont circulé (l'année précédente) hors vacances scolaire et qui sont conformes à la grille horaire *i.e.* qui ont desservi les arrêts définis dans l'offre de référence, voir la Section 2.1.1 pour des détails sur l'offre de référence. Les valeurs de référence pour la prévision de retard sont normalement les horaires théoriques. La prévision de retard à Transilien consiste à translater la déviation de l'horaire réel par rapport à l'horaire théorique aux autres gares *i.e.* aux horaires d'arrivés théoriques futurs, comme illustré sur la Figure 4.6 de droite. La notion de retard de référence n'est pas naturelle. Cependant, pour présenter les deux méthodes de prévision à court terme de Transilien de la même manière, nous introduisons la notion de retard de référence $\overline{\Delta a}$, que nous fixons artificiellement à 1, quelque soit l'arrêt. À l'aide de cet artifice le retard de référence est neutre lors de la prévision.

Étape 2 : calcul des taux d'actualisation. Pour la prévision de la charge à bord, le taux d'actualisation est un taux d'accroissement relatif défini par l'augmentation de la charge à bord observée par rapport à la charge à bord de référence, c'est-à-dire :

$$\rho^d_{k,s} = \frac{\ell^d_{k,s} - \overline{L}_{k,s}}{\overline{L}_{k,s}}.$$

Afin d'unifier la présentation des deux stratégies de prévision, nous présentons également un taux d'actualisation pour les retards :

$$\rho^d_{k,s} = \frac{\Delta a^d_{k,s} - \overline{\Delta A}_{k,s}}{\overline{\Delta A}_{k,s}} = \Delta a^d_{k,s} - 1.$$

Étape 3 : prévision. La prévision de la charge à bord à la gare suivante, notée $\hat{\ell}^d_{k,s+1}$, est simplement définie par l'application du taux d'accroissement relatif $\rho^d_{k,s}$ à la charge de référence à la gare $s+1$:

$$\widehat{\ell}^d_{k,s+1} = (1 + \rho^d_{k,s})\overline{L}_{k,s+1}.$$

La mise à jour du retard de référence revient simplement à translater l'horaire théorique d'arrivée à la gare suivante du retard observé à la gare actuelle :

$$\widehat{\Delta a}^d_{k,s+1} = (1 + \rho^d_{k,s})\overline{\Delta A}_{k,s+1} = \Delta a^d_{k,s}.$$

Généralisation. Partant de la stratégie de Transilien, nous proposons une méthode alternative pour calculer les valeurs de référence, généralisable à toutes les variables. Le principe est de calculer une valeur moyenne pour chaque arrêt (gare, train), notée $\overline{x}_{k,s}$. Cette valeur moyenne est calculée sur une partie de notre jeu de données allant de janvier à septembre 2019. Nous notons $D_{\mathcal{T}\text{rain}}$ l'ensemble des jours de notre jeu de données d'entrainement. Ainsi, la stratégie de prévision applicable à toutes les variables, inspirée de la stratégie de Transilien suit les trois étapes suivantes :

1. Calcul d'une valeur de référence :
$$\forall k, s \quad \overline{x}_{k,s} = \frac{1}{D_{\mathcal{T}\text{rain}}} \sum_{d=1}^{D_{\mathcal{T}\text{rain}}} x^d_{k,s},$$
2. Calcul d'un taux d'actualisation :
$$\rho^d_{k,s} = \frac{x^d_{k,s} - \overline{x}_{k,s}}{\overline{x}_{k,s}},$$
3. Prévision :
$$\widehat{x}^d_{k,s+1} = (1 + \rho^d_{k,s})\overline{x}_{k,s+1}.$$

Limites. Une telle stratégie de prévision, bien qu'intéressante par sa simplicité et sa capacité à s'ajuster à des situations particulières, présente plusieurs limites. D'une part, elle nécessite de calculer un taux d'actualisation pour chaque arrêt et chaque jour, ce qui revient à calculer autant de « paramètres » qu'il n'y a d'arrêts. D'autre part, elle ne s'intéresse qu'à la déviation linéaire par rapport à une valeur de référence qui est ensuite extrapolée. L'objectif de ce chapitre est de développer un modèle capable d'exploiter la propagation locale d'une variable (le nombre B de montées, le nombre A de descentes, etc.). Pour cela, nous simplifions le modèle proposé par Corman and Kecman [2018] en le rendant plus parcimonieux tout en conservant son niveau de performance.

4.3.2 Modèle bi-directionnel sur un graphe

Le modèle bi-directionnel de ce chapitre est inspiré de la littérature sur les modèles de chaîne de Markov uni-directionnelle sur un graphe d'événements [Kecman and Goverde, 2015] et celles des modèles de réseaux bayésiens [Corman and Kecman, 2018]. Kecman and Goverde [2015] ont proposé un modèle de chaîne de Markov uni-directionnel sur un graphe d'événements (arrivée ou départ) qui permet de prédire le retard à partir du retard précédent pour le même train, autrement dit à l'aide d'un voisinage en train d'une profondeur de 1. Corman and Kecman [2018] ont généralisé cette idée grâce aux réseaux bayésiens, à partir d'un voisinage en train et du retard pour le train précédent à la même gare (autrement dit un voisinage en gare), tous les deux d'une profondeur d'un arrêt. Ils supposent que les lois de probabilités conditionnelles sont gaussiennes. De notre côté, nous utilisons le graphe défini en Section 4.2 et la structure périodique du plan de transport pour définir un voisinage en train et en gare, dont la profondeur est un hyperparamètre. Dans la suite, la fonction $f_{k,s}$ de l'équation (4.1) est un modèle de régression linéaire quel que soit la variable d'intérêt et $\varepsilon_{k,s}$ est gaussien.

Voisinage en forme de L

Nous avons défini dans la Section 4.2, la notion de voisinage en train (gares précédentes desservies par le train) et en gare (trains précédents desservant la gare). Nous commentons ces notions avec pour support la Figure 4.4. Pour prédire une variable d'intérêt du train 29 à la gare d'Épinay-Villetaneuse (la cellule violette), on peut utiliser les arrêts passés venant des deux directions : des trains précédents à la gare d'intérêt (les cellules roses) ; du train d'intérêt aux gares précédentes (les cellules blues). Nous nous restreignons à un passé proche ($P = 2$ derniers trains en rose foncé et $Q = 3$ dernières gares en bleu foncé). L'information utilisée pour prédire une variable d'intérêt à chaque arrêt forme un L inversé que nous appelons voisinage en forme de L de profondeur P et Q, et que nous notons $\mathcal{L}_{k,s}^{P,Q}$. Ce voisinage est bi-directionnel au sens où les variables explicatives d'un arrêt (k, s), proviennent de deux sources différentes. La première source, représentée horizontalement sur la Figure 4.4, vient d'un observateur fixe à une gare s qui enregistre toutes les réalisations des trains qui passent. La seconde

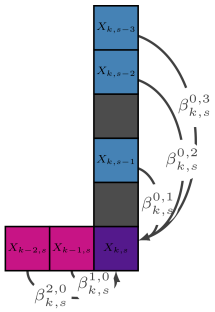


FIGURE 4.7 – Exemple de régression linéaire en forme de L pour un arrêt (k, s).

source, représentée verticalement sur la Figure 4.4, vient d'un observateur mobile dans un train k qui enregistre toutes les réalisations des arrêts de ce train. L'arrêt (k, s) est le point de rencontre de ces deux observateurs. Ainsi, les arrêts observés avant leur rencontre forme le voisinage en forment de L. Pour rappel et comme illustré sur la Figure 4.7, les gares non desservies n'appartiennent pas au voisinage en forme de L.

Modèle de régression linéaire en forme de L

Dans un premier temps, nous considérons un modèle par arrêt (k, s). Nous définissons un modèle de régression linaire par variable $X_{k,s}$ (temps de stationnement T, nombre B de montées, etc.) conditionnellement à son voisinage en forme de L, de la façon suivante :

$$X_{k,s} = \beta_{k,s}^{0,0} + \sum_{p=1}^{P} \beta_{k,s}^{p,0} x_{k-p,s} + \sum_{q=1}^{Q} \beta_{k,s}^{0,q} x_{k,s-q} + \varepsilon_{k,s}. \tag{4.2}$$

où l'ordonnée à l'origine $\beta_{k,s}^{0,0}$ dépend du noeud (k, s), de même que les coefficients $\beta_{k,s}^{1:P,0}$ et $\beta_{k,s}^{0,1:Q}$ modélisant les liens avec le passé. $\varepsilon_{k,s}$ est un bruit gaussien. Ce modèle, représenté pour un arrêt (k, s) sur la Figure 4.8 avec $P = 2$ et $Q = 3$, est appelé *modèle non stationnaire*. Un modèle non stationnaire avec $P = Q = 1$ est proche du modèle de Corman and Kecman [2018]. Les modèles non stationnaires avec $P = Q = 0$ et $P = Q = 1$ servent, avec le modèle de Transilien, de référence. L'estimation des paramètres de chaque modèle se fait classiquement par minimisation des moindres carrés.

Limites. Le modèle non stationnaire permet de modéliser chaque arrêt à l'aide d'un modèle de régression linéaire prenant en compte la propagation locale des variables avec entre 1 et 7 paramètres par arrêt (k, s). On constate sur la Figure 4.4, que la grille horaire est composée de motifs répétés. L'objectif est donc d'exploiter ces motifs pour réduire le nombre de paramètres du modèle.

4.3.3 Modèle bi-directionnel avec motifs et voisinages optimaux

Nos contributions sont de deux ordres. D'une part, nous proposons un modèle plus parcimonieux en réduisant significativement le nombre de paramètres nécessaires à la modélisation à l'aide des motifs. D'autre part, nous opérons la sélection d'un voisinage optimal pour chaque arrêt.

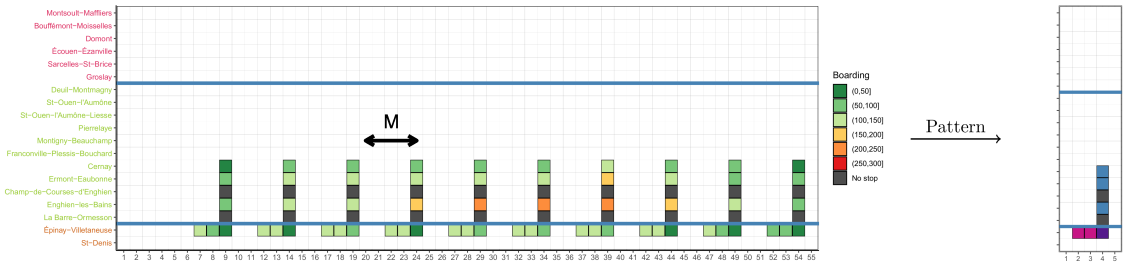


FIGURE 4.8 – Illustration de la répétition d'un voisinage en L sur le graphique gares-trains de l'heure de pointe du matin.

Idée de motifs

Sur la partie droite de la Figure 4.3, nous remarquons quatre répétitions d'une même succession de cinq missions différentes. Ces répétitions ne sont pas étonnantes car la grille horaire est construite à partir de motifs, garantissant un espacement fixe entre les différents missions. Nous représentons sur la Figure 4.8, les répétitions d'un même voisinage en forme de L déjà introduit dans la Figure 4.4. Nous notons M l'espacement inter-motif. Ici, le même motif se répète tous les $M = 5$ trains. Cette répétition à l'intérieur d'une même plage horaire vient s'ajouter aux répétitions entre les jours. Chaque index de train k correspond à un index de mission du motif, obtenu en calculant la valeur k modulo M, notée $k[M]$. Le nombre d'arrêts (k, s) à modéliser, et donc le nombre de paramètres, diminue d'autant qu'il y a de répétitions du motif contenant k. L'arrêt $(k[M], s)$ est l'arrêt projeté issu de (k, s). Tous les arrêts (k, s) projetés dans $(k[M], s)$ ont les mêmes voisinage et paramètres. L'équation associée à chaque arrêt est :

$$X_{k,s} = \beta^{0,0}_{k[M],s} + \sum_{p=1}^{P} \beta^{p,0}_{k[M],s} x_{k-p,s} + \sum_{q=1}^{Q} \beta^{0,q}_{k[M],s} x_{k,s-q} + \varepsilon_{k,s}. \tag{4.3}$$

L'idée de projeter chaque train sur un ensemble de missions qui se répète tous les cinq trains a deux sources d'inspiration. La première vient des séries temporelles, pour ces dernières ; il est courant de modéliser la saisonnalité en plus du processus stochastique stationnaire d'intérêt. La seconde vient des réseaux bayésiens dynamiques qui permettent d'utiliser une même structure de graphe pour un processus stochastique variant dans le temps [Roos et al., 2017]. Ici, nous généralisons cette idée de répétition à un phénomène spatio-temporelle, en appelant *modèle stationnaire* un modèle en forme de L stable d'un motif à l'autre. Les différences entre le modèle non stationnaire de l'équation (4.2) et le modèle stationnaire de l'équation (4.3) sont représentées en orange dans cette dernière. Nous définissons également un *modèle semi-stationnaire*, à mi-chemin entre le modèle stationnaire et le modèle non stationnaire. Il autorise des variations moyenne du phénomène par arrêt de la grille horaire, tout en conservant, l'idée de stabilité des dépendances entre les arrêts par motif, ce qui donne l'équation suivante :

$$X_{k,s} = \beta^{0,0}_{k,s} + \sum_{p=1}^{P} \beta^{p,0}_{k[M],s} x_{k-p,s} + \sum_{q=1}^{Q} \beta^{0,q}_{k[M],s} x_{k,s-q} + \varepsilon_{k,s}. \tag{4.4}$$

Nous colorons en rouge dans l'équation 4.4, la différence entre le modèle stationnaire de l'équation 4.3 et le modèle semi-stationnaire de l'équation 4.4.

Sélection d'un voisinage optimal

L'introduction des motifs permet d'élargir la profondeur des voisinages en forme de L testés sans augmenter le nombre de paramètres. Nous proposons deux stratégies de sélection d'un voisinage. La première consiste à tester uniquement certaines combinaisons particulières définies par un voisinage en forme de L symétrique : $P = Q = 1$, $P = Q = 2$, $P = Q = 3$. La seconde consiste à sélectionner un voisinage optimal pour chaque arrêt projeté à l'aide d'un critère de choix de modèle. Cette sélection du voisinage est contrainte. Les voisins ne peuvent pas être disjoints de l'arrêt *i.e.* les sous-graphes du voisinage en train et en gare sont connexes. De plus, la profondeur des voisinages en gare P ou en train Q est comprise entre 0 et 3. Nous choisissons une profondeur de 3 car une plus grande profondeur écarterait trop de situations dues aux aléas affectant l'ordre de la grille horaire, voir la Figure 4.5 de la Section 4.2. Nous n'imposons pas ici que les voisinages soient symétriques. Dans cet espace contraint, l'ensemble des voisinages en forme de L pour chaque arrêt, est égal à l'ensemble des couples (i, j) avec $i, j \in \{0, 1, 2, 3\}^2$, il y a donc $4^2 = 16$ combinaisons possibles. Ce faible nombre de combinaisons permet une recherche exhaustive pour chaque modèle. Le critère de choix de modèle est l'erreur moyenne absolue par validation croisée à 10 sous-ensembles pour chaque couple (P, Q). Le principe de la validation croisée est détaillé dans la Section 3.A.2 mais nous en donnons les grandes idées ici. Nous partitionnons les jours du jeu de données d'entraînement en 10 sous-ensembles disjoints. Pour chaque couple, nous utilisons successivement neuf des dix sous-ensembles pour entrainer le modèle et le dixième pour calculer les erreurs moyennes absolue de prévision sur chaque sous-ensemble. Enfin, nous sélectionnons pour chaque arrêt le couple qui minimise l'erreur absolue moyenne sur les 10 sous-ensembles.

4.4 Résultats

Nous détaillons dans la Section 4.4.1 les données utilisées pour obtenir les résultats. Nous donnons les performances de prévision globales dans la Section 4.4.2. Nous confirmons ces résultats par une analyse plus locale des erreurs dans la Section 4.4.3. Dans la Section 4.4.4, nous présentons de façon synthétique, à l'aide d'*un graphe des vents*, les voisinages optimaux sélectionnés.

4.4.1 Périmètre d'étude

Dans ce chapitre, nous nous restreignons aux jours de semaine, hors vacances et jours fériés, pour définir un contexte relatif à une grille stable et une période critique. Nous nous focalisons sur l'heure de pointe du matin (55 trains entre 6h33 et 9h28) sur une période de 106 jours entre le 7 janvier et le 5 juillet 2019. Nous considérons les 20 stations représentées sur la Figure 4.1. Au total, le jeu de données contient un peu plus de 34 000 arrêts. Les variables de flux voyageurs (le nombre A de descentes et le nombre B de montées, la charge à bord L) sont mesurées à partir du comptage automatique voyageurs (APC). Les variables d'exploitation ferroviaire (le retard à l'arrivée Δa et le temps de stationnement T) sont mesurées à partir de la vitesse des trains (AVL). Nous découpons le jeu de données en deux sous-ensembles : un jeu de données d'entraînement (du 7 janvier au 20 mai) et un jeu de données test (du 21 mai au 5 juillet), ce qui représente un découpage en 70 % et 30 % des observations. Nous estimons les paramètres des modèles de régression en forme de L, sur le jeu de données d'entraînement, par minimisation des moindres carrés, puis nous calculons les performances de prévision des modèles associés sur le jeu de données test. Le nombre d'arrêts, dans le jeu de données test, est égal à $N_{\mathcal{T}\text{est}} = \sum_{d=1}^{D_{\mathcal{T}\text{est}}} \mathcal{N}^d$ où $D_{\mathcal{T}\text{est}}$ est l'ensemble des jours du jeu de données test et $\mathcal{N}^d$ l'ensemble des arrêts du jour d. L'indicateur de performance utilisé est l'erreur moyenne absolue défini par :

$$\text{MAE} = \frac{1}{N_{\mathcal{T}\text{est}}} \sum_{d=1}^{D_{\mathcal{T}\text{est}}} \sum_{(k,s)\in\mathcal{N}^d} \left| x_{k,s}^d - \widehat{x}_{k,s}^d \right|,$$

où x est une des variables aléatoires.

4.4.2 Résultats globaux

Nous présentons dans la Table 4.7 les performances en MAE de quatre familles de modèles : le modèle de Transilien, les modèles non stationnaire, semi-stationnaires et stationnaires. Le modèle non stationnaire avec $P = Q = 0$ revient à prédire pour chaque arrêt (k, s) la moyenne calculée sur le jeu de données d'entraînement. Les « vrais »modèles de prévision qui exploitent un voisinage pour prédire utilisent P ou $Q \geq 1$. Une des méthodes de référence, au-delà du modèle Transilien, correspond à la régression linéaire en forme de L non stationnaire avec $P = Q = 1$, voir Section 4.3. On constate dans la Table 4.7 qu'au global les différences de performances entre les familles de modèle sont particulièrement prononcées pour la prévision de retard et de charge à bord. Pour ces variables, le modèle non stationnaire avec $P = Q = 0$ a de très mauvaises performances, ce qui révèle que le retard et la charge à bord se propage de gare en gare. À l'inverse, l'écart entre le modèle non stationnaire avec $P = Q = 0$ et les autres familles de modèles est beaucoup plus faible pour la prévision des temps de stationnement. Ce constat révèle que prédire les temps de stationnement seulement avec son passé est peu pertinent. Concernant le nombre de paramètres, les modèles non stationnaires ont huit fois plus de paramètres, à voisinage équivalent, que les modèles stationnaires et deux fois plus que les modèles semi-stationnaires.

TABLE 4.7 – MAE (erreur moyenne absolue) des méthodes de prévision considérées (les deux premières colonnes). Pour chaque méthode, on donne le nombre de paramètres utilisés pour les variables T, ΔA, A (colonne 3) et pour les variables B, L (colonne 4). Ces deux quantités diffèrent car pour les gares origines : T, ΔA, A n'existent pas (cf. Figure 4.11). On donne également l'erreur moyenne absolue globale des cinq variables à prédire (colonnes 5-8) : le temps de stationnement T, le retard à l'arrivée ΔA, le nombre A de descentes, le nombre B de montées et la charge à bord L. La méthode de référence est en violet. Les méthodes sélectionnées sont en vert.

Modèles				Exploitation ferroviaire [RO]		Flux de voyageurs [PF]		
Nom	L-forme	Nombre de paramètres T, ΔA et A	Nombre de paramètres B et L	T [s]	ΔA [s]	A [voy]	B [voy]	L [voy]
Modèle de Transilien				11.7	23.3	10	22	24
Non-stationnaire	P = Q = 0	317	337	9.7	35.8	10	21	69
	P = Q = 1	885	956	9.5	16.1	9	18	20
Semi-stationnaire	P = Q = 1	391	417	9.3	18.6	10	19	23
	P = Q = 2	427	455	9.2	18.1	9	19	23
	P = Q = 3	452	482	9.2	18.1	9	18	23
	Automatique	Voir *Table 4.8*		9.2	18.3	9	19	23
Stationnaire	P = Q = 1	74	80	9.3	16.2	10	21	27
	P = Q = 2	110	118	9.2	15.8	8	20	27
	P = Q = 3	135	145	9.2	15.9	8	20	27
	Automatique	Voir *Table 4.8*		9.2	15.9	8	20	26

TABLE 4.8 – Dimension pour chaque variable des voisinages optimaux sélectionnés par validation croisée pour les modèles de régression linéaire en forme de L.

Variables		Nombre de paramètres
Exploitation ferroviaire [RO]	T	108
	ΔA	99
Flux de voyageurs [PF]	A	452
	B	443
	L	440

Le modèle de référence (cellules en violet) dans la Table 4.7 est le modèle non stationnaire avec un voisinage en forme de L de profondeur $P = Q = 1$. C'est le plus proche du modèle de Corman and Kecman [2018] pour la prévision des retards. L'idée est de tester jusqu'à quel point nous pouvons le généraliser et le simplifier sans perdre en qualité de prévision. Les meilleurs modèles sélectionnés (cellules en vert), dans la Table 4.7, sont différents pour chaque ensemble de variables [RO] ou [PF]. Pour les variables d'exploitation ferroviaire [RO], la meilleure stratégie de prévision est d'utiliser un modèle stationnaire ayant des performances légèrement meilleures avec moins de paramètres que le modèle de référence. Le modèle sélectionné est le modèle avec sélection automatique du voisinage car il est le meilleur compromis entre performance et parcimonie. En effet, la prévision des retards serait légèrement meilleure avec $P = Q = 2$ toutefois le nombre de paramètres est en faveur de la sélection automatique du voisinage avec 99 paramètres contre 118 pour $P = Q = 2$.

Pour la prévision des variables de flux de voyageurs [PF], les performances en MAE sont les meilleures, en dehors du modèle non stationnaire, pour un modèle semi-stationnaire, excepté pour la prévision des descentes. Cependant, le nombre de descentes, en heures de pointe du matin vers Paris hors la gare terminus Gare du Nord, est négligeable. C'est pour cette raison que l'erreur associée au nombre A de descentes est deux fois plus faible que celle associée au nombre B de montées. Ainsi, nous choisissons un modèle semi-stationnaire avec $P = Q = 1$ car c'est le meilleur compromis entre performance et parcimonie.

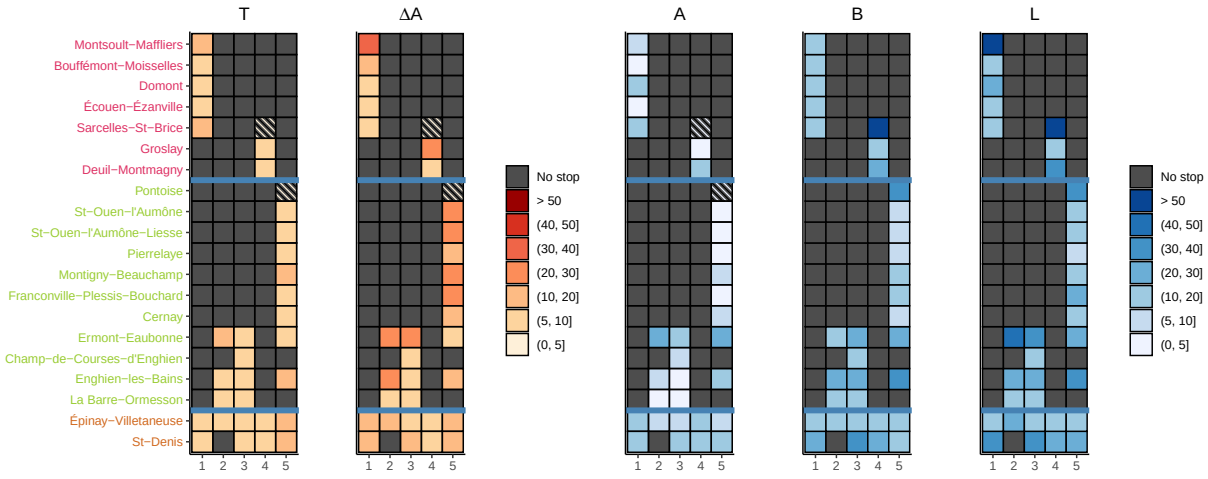


FIGURE 4.9 – Moyenne des erreurs absolues pour les méthodes en vert présentées dans la Table 4.7. Chaque graphique représente une variable d'intérêt. Chaque case est la moyenne pour un arrêt, caractérisé par une gare s en ordonnée et un indice de train projeté $k[M]$ en abscisse, pendant la période de pointe sur le jeu de données test. Les cases hachurées sont des gares origines. En rouge, les variables d'exploitation ferroviaire, et en bleu, les variables de flux voyageurs. Plus la couleur est claire plus l'erreur est faible.

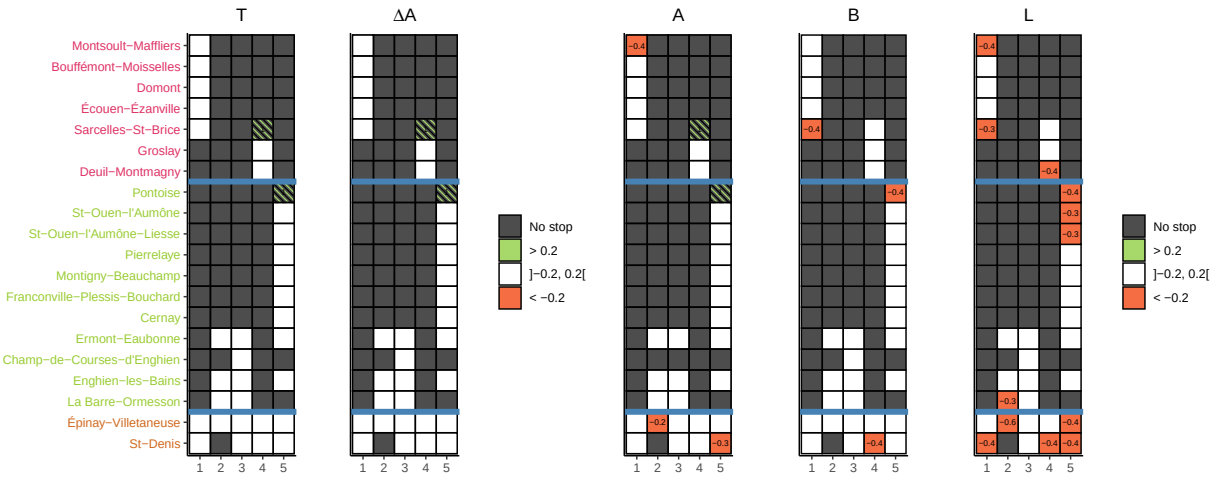


FIGURE 4.10 – Différence moyenne relative des erreurs entre la méthode de référence en violet dans la Table 4.7 et les méthodes sélectionnées en vert. Chaque motif représente une variable d'intérêt. Chaque case représente la moyenne relative de la différence pour un arrêt projeté sur un motif $(k[M], s)$ en heures de pointe pour le jeu de données test. Les améliorations par rapport à la méthode de référence sont en vert, les détérioration en orange et les équivalences au seuil de 20 % sont en blanc. Les valeurs portées dans les cases indiquent la différence moyenne d'erreur absolue relative entre les deux méthodes.

4.4.3 Résultats locaux

Sur la Figure 4.9, nous proposons une analyse des erreurs absolues moyennes pour chaque arrêt (k, s) (locale) des performances des modèles sélectionnés dans la Table 4.7. Le principal constat est que ce sont les gares en extrémité de périmètre qui ont les erreurs les plus importantes, par exemple, pour les arrêts en gare de Montsoult-Maffliers ou en gare de Sarcelles-Saint-Brice. Ces erreurs sont particulièrement importantes pour les variables de retards à l'arrivée et la charge à bord, ce qui confirme l'importance du voisinage en train pour ces deux variables. Sur la Figure 4.10, nous calculons pour chaque variable et pour chaque arrêt, la moyenne des différences relatives des MAE entre le modèle de référence et le modèle sélectionné. Nous considérons que la différence de performance est significative pour un arrêt si elle dépasse 20 % dans un sens ou dans un autre. Une différence relative négative montre que le modèle sélectionné commet localement plus d'erreur que le modèle de référence. Nous constatons que pour les temps de stationnement T et le retard à l'arrivée ΔA, les performances des deux modèles sont équivalentes pour tous les arrêts $(k[M], s)$. Pour la prévision du nombre B de montées et du nombre A de descentes les performances sont ponctuellement moins bonnes (3 cellules sur 32) pour la méthode sélectionnée mais le nombre de paramètres est deux fois plus faible. Pour la prévision de la charge à bord L, le modèle de référence est globalement meilleur (13 cellules sur 32) ce qui est cohérent avec les résultats de la Table 4.7.

4.4.4 Voisinages optimaux

Nous indiquons la dimension des voisinages optimaux sélectionnés dans la Table 4.8 pour les modèles *Automatique* de la Table 4.7. Ici, nous présentons ces voisinages optimaux pour chaque arrêt en indiquant l'intensité (le nombre de voisins) et la direction (d'où viennent les voisins de la même gare ou du même train) du voisinage. L'intensité est représentée par des flèches orientées vers Paris, de tailles et de couleurs différentes sur la Figure 4.11. La direction du voisinage est indiquée par l'angle de la flèche. Un angle droit (une flèche verticale) indique un voisinage exclusivement en train, à l'opposé un angle nul (une flèche horizontale) indique un voisinage exclusivement en gare, les angles entre ces deux valeurs extrêmes représentent l'ensemble des directions possibles. Sur la Figure 4.11, nous confirmons que les voisinages optimaux varient en fonction des variables. En particulier, l'intensité des voisinages quel que soit l'arrêt est plus faible pour le nombre de montées B que pour le retard à l'arrivée ΔA ou le nombre de descentes A. Par ailleurs, même si le voisinage maximal est $P = Q = 3$, tous les arrêts du graphe de gauche de la Figure 4.11 n'ont pas un voisinage de cette taille car il se peut, que sur leur branche, il n'y ait qu'une mission par motif qui passe par la gare (faible voisinage en gare), ou que le train parte de l'origine (aucun voisinage en train). Indépendamment de cette contrainte technique, on constate que les arrêts loin de Paris (en haut) ont tendance à atteindre le maximum de voisins, faute de choix peut-être, à l'inverse les arrêts proches de Paris (en bas) ont des voisinages très variables par rapport au voisinage maximal $P = Q = 3$, à gauche. Par exemple, pour le graphe à droite de la Figure 4.11 de la charge à bord L, le voisinage en gare de Saint-Denis pour la mission 4 est bien plus faible que le nombre maximal de la figure de gauche.

En résumé. Nous avons montré que l'introduction de motifs permet de simplifier le modèle de référence sans trop en dégrader ses performances, excepté pour la charge à bord. La différence des performances globale et locale permet de confirmer ce résultat. On note que les arrêts les plus difficiles à prédire sont les arrêts en origine et à Saint-Denis. Enfin, la sélection d'un voisinage optimal permet d'éviter d'avoir à fixer un voisinage a priori tout en réduisant globalement le nombre de paramètres et conservant de bonnes performances, en particulier pour les variables d'exploitation ferroviaire.

4.5 Conclusion et perspectives

Ce travail sur la prévision à court terme sur un réseau de transport est préliminaire, en particulier pour les variables de flux voyageurs à l'échelle du train. Nous avons obtenu plusieurs résultats intéressants grâce à notre jeu de données particulièrement riche.

1. **Qualité des données.** La stabilité et le taux de couverture de la grille

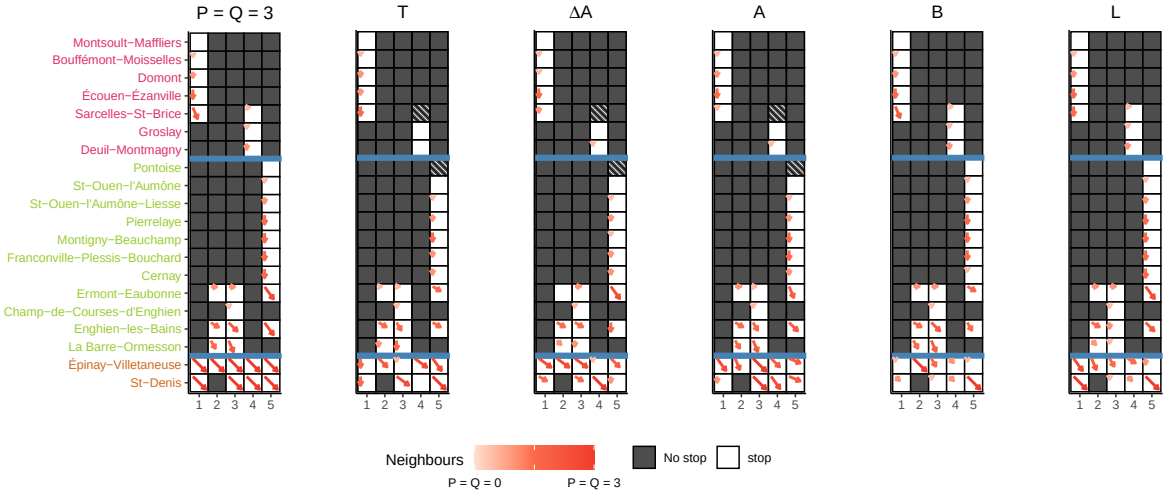


FIGURE 4.11 – Représentation par des flèches du voisinage optimal pour chaque arrêt. Le premier motif à gauche, $P = Q = 3$, est la borne supérieure i.e. le plus grand voisinage possible par arrêt. Le voisinage maximal pour un arrêt est atteint si et seulement s'il est entouré de suffisamment d'arrêts à l'intérieur d'un motif. Chaque motif représente une variable d'intérêt. Pour chaque arrêt la taille et l'intensité de la couleur des flèches représentent le nombre de voisins sélectionnés. L'orientation des flèches indique la direction privilégiée du voisinage i.e. un voisinage en gare (flèche horizontale) ou un voisinage en train (flèche verticale). Les cases hachurées sont les gares origines.

horaire à Transilien sont suffisants pour estimer des modèles de prévision dont le graphe est défini par la grille horaire.

2. **Généralisation.** Les modèles de prévision pour une variable, comme les modèles de Transilien ou le modèle non stationnaire inspiré de Corman and Kecman [2018], s'adaptent bien à la prévision d'autres variables. De même, la prévision à base de motifs permet de prédire avec moins de paramètres et des performances quasiment équivalentes. Une limite toutefois est que les temps de stationnement, à cause de leur dépendance aux horaires théoriques (voir Chapitre 3), ou le nombre de descentes, à cause de leur trop faible nombre en heures de pointe du matin, sont presque aussi bien prédits par un modèle non stationnaire $P = Q = 0$ que par un modèle plus sophistiqué.
3. **Parcimonie et performances.** L'exploitation des motifs dans la grille horaire permet de réduire le nombre de paramètres sans globalement diminuer les performances de prévision, excepté pour la charge à bord.
4. **Voisinage optimal.** La méthode proposée pour sélectionner le meilleur voisinage est efficace car elle permet de réduire le nombre de paramètres tout en ouvrant un large champ de recherche sur l'interprétation des voisinages sélectionnés.

Dans ce chapitre, nous avons construit des ponts entre la prévision court terme des retards et celle des flux de voyageurs. Nous avons posé les bases pour une discussion constructive sur les prochains défis de la prévision à court terme sur un réseau de transport avec une grille horaire, dont les trois principaux défis identifiés sont :

1. **Horizons de prévision.** Étendre la prévision à court terme aux gares $s + 2, s + 3, \ldots$;
2. **Prévisions conjointes des variables.** Prédire conjointement les différentes variables, par exemple les montées et les descentes ;
3. **Sensibilité au désordre.** Fournir des estimations et des prévisions dans des situations non conformes qui ont été exclues de ce travail.

D'autres suites possibles concernent enfin l'injection de saisonnalité entre motifs et l'utilisation de modèles non linéaires pour chaque arrêt.

5

Modélisation probabiliste des déplacements

Nous précisons les enjeux de la modélisation des déplacements à bord pour Transilien. Nous donnons des ordres de grandeur quant à l'ampleur des déplacements à bord. Nous proposons deux modèles simples des déplacements à l'échelle du trajet permettant de diviser par deux l'erreur d'estimation des descentes par rapport à un modèle sans déplacement. Nous exprimons la vraisemblance d'un modèle à variables latentes adapté aux comportements de déplacements différenciés selon les gares.

Contents

5.1 Introduction

Ce chapitre vise à poser certains fondements théoriques afin de faciliter l'utilisation des capteurs infra-rouges au-dessus des portes des nouvelles rames communicantes pour informer les voyageurs sur la *charge à bord* par zone. Le projet Hector [Jarrossay, 2021] est en ce sens la source d'inspiration de ce chapitre. L'enjeu scientifique est d'estimer la charge à bord cachée à partir du nombre de montées et de descentes par zone, sachant que les voyageurs se déplacent librement à bord des rames communicantes. Nous abordons ce problème d'estimation de la charge à bord cachée par la modélisation des proportions de *déplacements* de zone à zone pour chaque gare.

La modélisation des déplacements repose sur plusieurs hypothèses dont deux sont communes à l'ensemble du chapitre. La première suppose que les voyageurs descendent depuis la zone où ils ont voyagé *i.e.* qu'ils se sont déplacés immédiatement après être montés à bord du train. La seconde suppose que la charge à bord est un processus de Markov *i.e.* que la charge à bord en sortie de gare ne dépend que de la charge à bord à la gare précédente (ainsi que des montées et descentes à cette gare).

La modélisation des déplacements entre les zones est un problème suffisamment complexe pour nécessiter une résolution en deux temps. Dans un premier temps, nous résolvons un cas simplifié où tous les déplacements sont supposés identiques quelle que soit la gare. Cela revient à modéliser les proportions de déplacement à l'échelle du trajet *i.e.* en agrégeant le nombre de montées et de descentes suivant toutes les gares. Dans un deuxième temps, nous relâchons l'hypothèse d'uniformité des déplacements entre les gares en proposant un modèle des déplacements spécifique à chaque gare, appelé modèle à l'échelle de la gare. Dans ce cas, les déplacements, et par conséquent, la charge à bord au fil du trajet sont des variables *latentes/cachées* révélées au fur et à mesure par les descentes. Nous modélisons donc conjointement les déplacements et les descentes pour chaque gare.

Dans ce chapitre, nous avons testé sur des données réelles le modèle à l'échelle du trajet. Le modèle à l'échelle de la gare a été formalisé mais n'est pas encore testé sur des données réelles.

Plan du chapitre Nous montrons dans la Section 5.2 que les données de montées et descentes par zone sont peu exploitées dans la littérature. L'originalité de ces données nous amène à proposer une modélisation probabiliste novatrice des déplacements à l'intérieur des trains. Nous donnons des ordres de grandeur sur l'ampleur des déplacements à bord des rames dans la Section 5.3. Nous présentons la méthode et les résultats associés à la modélisation des déplacements à l'échelle du trajet dans la Section 5.4, puis nous présentons le modèle des déplacements à l'échelle de la gare à l'aide des variables latentes dans la Section 5.5. L'Annexe 5.A justifie l'approximation proposée pour la loi des déplacements issue d'une somme de lois multinomiales indépendantes mais non identiquement distribuées.

5.2 État de l'art et notations

Dans cette section, il s'agira dans un premier temps de faire un état de l'art sur la modélisation des déplacements dans les trains ou dans les bureaux pour, dans un deuxième temps, présenter les notations utilisées.

5.2.1 État de l'art

Nous montrons d'abord, qu'il existe des solutions de comptage directes de la charge à bord par zone mais nous rappelons que Transilien a privilégié une solution de mesure indirecte de la charge à bord à partir des montées et descentes par zone. Nous présentons ensuite les études et modèles des montées et descentes par zone dans les transports en commun, en insistant sur le fait que ces études ou modèles ne prennent quasiment jamais en compte les déplacements à bord. La partie suivante est l'occasion de présenter quelques travaux sur la modélisation des déplacements à bord à l'aide de modèles multi-agents pour simuler la trajectoire d'un voyageur jusqu'à son siège. Enfin, en ouvrant le sujet aux déplacements des employés dans un bâtiment, on constate que certains chercheurs développent des modèles probabilistes intéressants, capables de résoudre des problèmes proches du notre, par exemple le déplacement d'un groupe de personnes d'une pièce à l'autre d'un bâtiment au fil de la journée.

Mesure directe de la charge à bord par zone

La mesure directe de la charge à bord par zone est privilégiée par de nombreux opérateurs à l'aide de caméras vidéos ou de capteurs de pression. Thalès à Londres [Thales, 2021] utilise les caméras de vidéo-surveillance native pour mesurer le volume de voyageurs. Des lignes à Copenhague [Nielsen et al., 2014], Stockholm [Zhang et al., 2017], Londres [Schmitt, 2017, Rogers, 2019] ou Singapour [Ngauw, 2018] privilégient une mesure par la masse à l'essieu en utilisant les capteurs de pression au niveau des essieux. Nielsen et al. [2014] décrivent avec beaucoup de détails comment passer de la masse à l'essieu à la charge à bord à l'aide des comptages manuels. La charge à l'essieu permet de mesurer directement la charge à bord mais elle ne permet pas de mesurer les flux de voyageurs à l'interface quai-train. Ces flux de voyageurs sont pourtant importants pour modéliser les temps de stationnement. Pour palier cette limite, Peftitsi et al. [2020] couplent masse à l'essieu et matrices Origine-Destination pour reconstruire des flux de voyageurs par zone à partir de la masse à l'essieu. Pour éviter cette étape de reconstruction, Transilien privilégie une mesure indirecte de la charge à bord à partir du nombre de montées et de descentes. Ce nombre de montées et descentes par porte est mesuré à l'aide de capteurs infra-rouges ou vidéo au niveau des portes. Cette technologie se généralise grâce à un investissement massif de la part d'Île-de-France Mobilité (IdFM) qui permet à Transilien de renouveler environ 80% de ses rames entre 2008 et 2035. Ces

nouvelles rames sont toutes équipées de capteurs. Ces rames ont toutes des voitures communicantes permettant aux voyageurs de circuler librement à bord après qu'ils soient montés, ce qui complexifie la conversion des montées et descentes en charge à bord.

Modélisation des montées et des descentes à quai

Le problème de l'estimation de la charge à bord dans les transports en commun a souvent été abordé en négligeant les déplacements à bord. L'analyse de la répartition des montées a déjà fait l'objet de plusieurs études : Szplett and Wirasinghe [1984] montrent, sur deux gares de la ligne de C-Train de Calgary, que les montées se concentrent autour des entrées/sorties de quai. Plus récemment, Krstanoski [2014] a modélisé la distribution à quai des montées par zone du train à partir d'une loi multinomiale ou chaque classe correspond à une porte du métro. Seuls Szplett and Wirasinghe [1984], Krstanoski [2014] modélisent de façon probabiliste la position des voyageurs à quai. Il est plus courant d'utiliser des modèles de simulation de mouvement des piétons comme Seriani and Fujiyama [2019] ou Hänseler et al. [2020]. Cette dernière approche est non seulement coûteuse, car elle nécessite de modéliser chaque gare et quai, mais en plus ces modèles ne s'appliquent pas directement aux déplacements à bord.

Modélisation multi-agents des déplacements à bord

Une des approches pour modéliser les déplacements à bord est de modéliser le choix de chaque voyageur à l'aide d'un ensemble de règles abstraites. Schöttl et al. [2019] ont développé le modèle multi-agents `Vadere` qui permet de remplir les trains sans réservation. Ce modèle implique trois choix emboîtés : choix de la zone, choix du groupe de sièges et choix du siège. Ils intègrent des règles issues de recherches en psychologie sociale comme le fait de privilégier des sièges dans le sens de la marche [Trinkoff, 1985], ou le fait de s'assoir le plus loin possible des autres voyageurs [Evans and Wener, 2007]. Un tel niveau de détail n'est pas compatible avec les contraintes industrielles du modèle souhaité car il faut qu'il soit léger et facilement généralisable. Nous privilégions donc une modélisation probabiliste des déplacements d'un groupe de voyageurs d'une zone à l'autre. Les modèles d'occupation des bâtiments permettent justement de passer d'un agent à un groupe d'agents et d'un siège à une zone.

Modélisation probabiliste des déplacements dans un bâtiment

Dans la littérature sur l'occupation des bâtiments, une des approches utilisée consiste à simplifier la modélisation multi-agents à deux niveaux. D'une part, en ne considérant plus le choix d'un emplacement mais celui d'une pièce. D'autre part, en ne considérant pas le comportement individuel de chaque personne mais celui

d'un groupe de personnes. Par exemple, Wang et al. [2011] proposent un modèle de chaîne de Markov où ils fixent la matrice de passage $\boldsymbol{P}$ entre 6 pièces d'un bureau. Shelat et al. [2020] améliorent ce modèle en le linéarisant, sans proposer pour autant une méthode d'estimation de la matrice de passage. Dans cette littérature, les données utilisées proviennent souvent de capteurs de CO_2, de température, de son, etc. qui donnent accès directement au volume de personnes présentes dans une pièce. Dans notre cas, nous ne mesurons pas directement le nombre de personnes dans chaque zone. Nous mesurons les flux de voyageurs qui rentrent et sortent par zone du train sans pouvoir mesurer les déplacements à l'intérieur du train. C'est pourquoi nous proposons un modèle probabiliste des déplacements utilisant les entrées et sorties de chaque zone à chaque gare.

5.2.2 Notations

Pour rappel, nous observons des trajets qui sont une suite d'arrêts d'un train k pour un jour donné d. Pour chaque trajet, nous observons les vecteurs de montées $\boldsymbol{b}_{1:S}^{k,d} = \left(\boldsymbol{b}_1^{k,d}, \ldots, \boldsymbol{b}_S^{k,d}\right)$ et de descentes $\boldsymbol{a}_{1:S}^{k,d} = \left(\boldsymbol{a}_1^{k,d}, \ldots, \boldsymbol{a}_S^{k,d}\right)$. On remarque que le vecteur des descentes en origine $\boldsymbol{a}_1^{k,d}$ ainsi que celui des montées en terminus $\boldsymbol{b}_S^{k,d}$ sont nuls.

Dans ce chapitre, nous exploitons la mesure la plus précise des flux de voyageurs telle que nous observons les montées $b_{s,i}^{k,d}$ et les descentes $a_{s,i}^{k,d}$ définies pour une zone $i \in \{1, \ldots, I\}$ à la gare s pour un trajet (k, d). Ces observations comprennent, pour chaque arrêt, les vecteurs des montées $\boldsymbol{b}_s^{k,d} = (b_{s,1}^{k,d}, \ldots, b_{s,I}^{k,d})$ et des descentes $\boldsymbol{a}_s^{k,d} = (a_{s,1}^{k,d}, \ldots, a_{s,I}^{k,d})$. On remarque que l'estimation de la charge à bord $\ell_{s,i}^{k,d}$, qui est le nombre de personnes présentes dans une zone, est un problème compliqué. Nous rappelons que ce problème à l'échelle du train, que nous notons $\ell_{s,\bullet}^{k,d} = \sum_{i=1}^{I} \ell_{s,i}^{k,d}$, se calcule simplement grâce à l'équation 5.1 :

$$\ell_{s,\bullet}^{k,d} = \sum_{g=1}^{s} b_{g,\bullet}^{k,d} - a_{g,\bullet}^{k,d}. \tag{5.1}$$

Les notations présentées ci-dessus et celles introduites par la suite sont résumées dans les Tables 5.1–5.4.

TABLE 5.1 – Indices des observations

Indices des observations	
k	un train
d	un jour
$\mathcal{N}$	l'ensemble des couples (k, d) observés
Indices des variables	
s	une gare
i	une zone
$\{1, \ldots, S\}$	l'ensemble des gares
$\{1, \ldots, I\}$	l'ensemble des zones

TABLE 5.2 – Variables pour une zone (i) à une gare (s) observées pour un train (k) le jour (d)

Variables cachées	
$\ell_{s,i}^{k,d}$	la charge à bord
$w_{s,i}^{k,d}$	les déplacements jusqu'à la zone i
Variables observées	
$a_{s,i}^{k,d}$	le nombre de descentes
$b_{s,i}^{k,d}$	le nombre de montées

TABLE 5.3 – Indices de sommation illustrés avec une variable générique x

$\boldsymbol{x}_s = (x_{s,1} \ldots, x_{s,I})$	le vecteur par zone de taille I pour une gare donnée s
$\boldsymbol{x}_{1:s} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s\}$	l'ensemble des vecteurs pour les gares de 1 à s
$x_{s,\bullet} = \sum_{i=1}^{I} x_{s,i}$	la somme selon les zones pour une gare donnée s
$x_{\bullet,i} = \sum_{s=1}^{S} x_{s,i}$	la somme selon les gares pour une zone donnée i
$x_{\bullet,\bullet} = \sum_{i=1}^{I} \sum_{s=1}^{S} x_{s,i}$	la somme selon les gares et les zones

TABLE 5.4 – Paramètres de modélisation des déplacements à l'échelle d'une gare ou d'un trajet

Paramètres pour une gare s	
$p_{s,i,j}$	la probabilité de se déplacer de i vers j
$\boldsymbol{P}_s = [p_{s,i,j}]_{i,j}$	la matrice stochastique associée aux probabilités de déplacement
$r_{s,i} = b_{s,i}/b_{s,\bullet}$	la proportion de montées en zone i
$\alpha_{s,i}$	la probabilité de descentes en zone i
$\pi_{s,i,j} = r_{s,i} p_{s,i,j}$	la probabilité de se déplacer de i vers j pondérée par la proportion de montées en zone i
Des exemples de paramètres à l'échelle d'un trajet	
$r_i = b_{\bullet,i}/b_{\bullet,\bullet}$	la proportion globale de montées en zone i à l'échelle d'un trajet
$\pi_{i,j} = r_i p_{i,j}$	la probabilité globale de se déplacer de i vers j pondérée par la proportion globale de montées en i

5.3 Quelques ordres de grandeur

Avant de continuer la modélisation du problème, nous étudions les ordres de grandeur des phénomènes en jeu. La suite montrera que la détermination préliminaire de ces ordres de grandeur est importante pour justifier une approximation lors de la modélisation. Le volume de voyageurs ainsi que le nombre de trains sur un réseau de transport public sont des quantités variables dans le temps et dans l'espace. Nous donnons des ordres de grandeurs à partir des données de comptage non redressées de la ligne H, en particulier des trains allant de Gare du Nord à Pontoise de janvier à septembre 2021. Les trains de la ligne H sont composés d'une ou deux rames qui ne communiquent pas entre elles. Nous ne conservons que les trains à deux rames qui desservent la même zone de quai. Les zones sont numérotées de 1 à 8 pour la rame avant et de 9 à 16 pour la rame arrière. Ces deux rames sont imperméables car aucun déplacement n'est possible d'une rame à l'autre. Nous représentons cette séparation par un trait gris sur la Figure 5.1. Cette indépendance entre les rames fait que nous proposons un modèle par position (avant/arrière).

5.3.1 Différences des distributions de montées et descentes

Nous comparons les volumes moyens de descentes, notés $\boldsymbol{a}_\bullet$ et ceux de montées, notés $\boldsymbol{b}_\bullet$ par zone à l'échelle du trajet sur le graphique du milieu de la Figure 5.1. Pour simplifier l'écriture, nous supprimons les indices (k, d). $\boldsymbol{a}_\bullet = (a_{\bullet,1}, \ldots, a_{\bullet,I})$ est un vecteur ligne de taille I où chaque élément, noté $a_{\bullet,i}$, est la somme des descentes pour la zone i tout au long du trajet :

$$a_{\bullet,i} = \sum_{s=1}^{S} a_{s,i} \ .$$

Nous reprenons cette notation pour les montées et la charge à bord. Les deux graphiques en haut de la Figure 5.1 représentent les valeurs $b_{s,i}$ (à gauche) et $a_{s,i}$ (à droite), qui représentent des moyennes sur l'ensemble des trajets (k, d), noté $\mathcal{N}$. Leurs valeurs agrégées le long du trajet sont représentées sous forme de diagramme en barres sur la graphique du milieu de la Figure 5.1, puis fusionnées sur le graphique du bas. Nous constatons à la lecture de ce dernier graphique qu'il y a beaucoup plus de montées que de descentes agrégées pour les zones 8 et 16. Ce déséquilibre s'explique par la géographie des quais à Gare du Nord qui incite les voyageurs à monter dans la zone arrière d'une des deux rames puis à marcher dans le train jusqu'à trouver une place assise. Ce déséquilibre pose problème car pour calculer la charge à bord, nous calculons la somme de la différence du nombre de montées et de descentes au fil du trajet par zone. Ainsi, pour les zones 8 et 16, en appliquant brutalement ce cumul, il y aurait des charges à bord en terminus fortement négatives, ce qui signifie soit que des voyageurs ont disparu, soit qu'ils se sont déplacés d'une zone à l'autre.

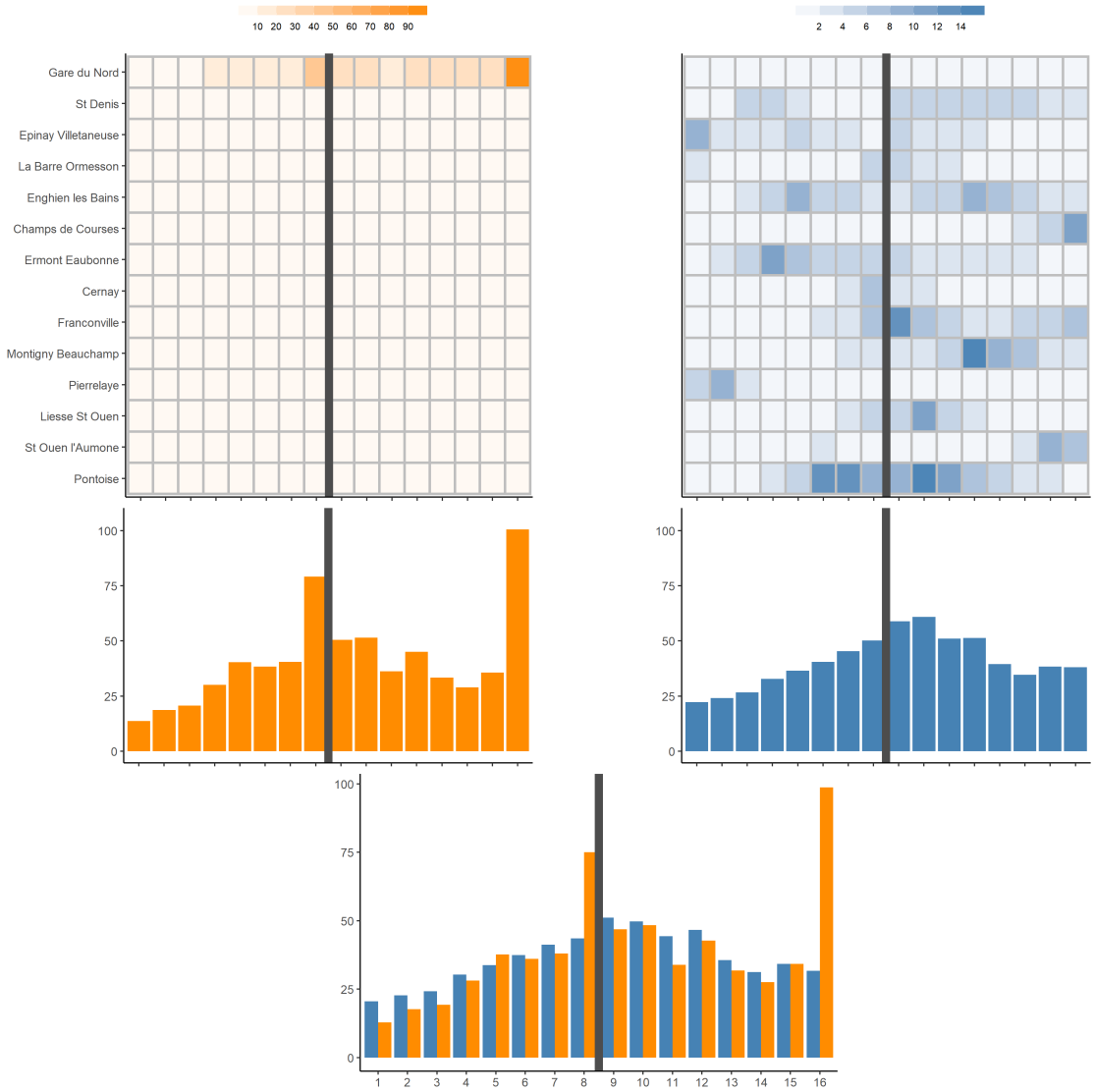


FIGURE 5.1 – Illustration de la somme du vecteur des montées $\boldsymbol{b}_s$ (en orange) et des descentes $\boldsymbol{a}_s$ (en bleu) par gare à l'échelle du trajet. En haut à gauche, les montées représentées par une carte de chaleur, où chaque entrée est la moyenne des montées par gare pour chaque zone, que nous sommons à l'échelle du trajet sur le diagramme en barres juste en-dessous. En haut à droite, les descentes représentées par une carte de chaleur puis sommées, en-dessous. Les barres grises verticales séparent la rame avant, numérotée de 1 à 8, de la rame arrière, numérotée de 9 à 16. Le diagramme en barres en bas représente la moyenne des montées et des descentes sommées à l'échelle du trajet pour toutes les zones $(\boldsymbol{b}_\bullet, \boldsymbol{a}_\bullet)$.

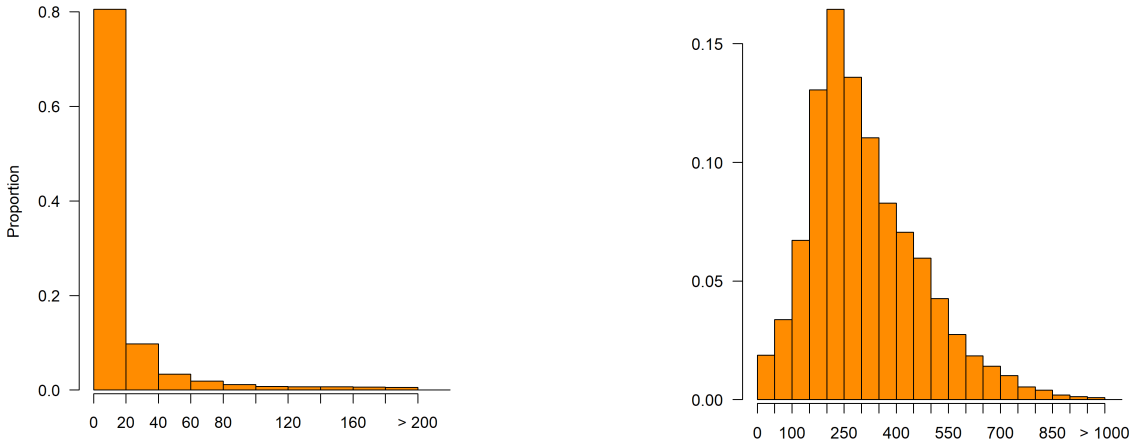


FIGURE 5.2 – Répartition du nombre de montées par train pour toutes les gares ($b_{s,\bullet}$) sur le graphique de gauche ou agrégé par trajet ($b_{\bullet,\bullet}$) sur le graphique de droite.

5.3.2 Volumes de montées à l'échelle du trajet et de la gare

Nous étudions les déplacements à bord de rames dont la capacité [1] varie en fonction du matériel roulant entre 726 et 1028 places. Nous rappelons que le nombre de montées à l'échelle du train à la gare s, noté $b_{s,\bullet}$ est égal à :

$$b_{s,\bullet} = \sum_{i=1}^{I} b_{s,i}.$$

Il ne peut jamais dépasser la capacité totale du matériel. En général, $b_{s,\bullet}$ est même beaucoup plus faible que cette capacité totale, comme en atteste la Figure 5.2 où le nombre de montées par gare est rarement supérieur à 60. En revanche, le nombre de montées $b_{\bullet,\bullet}$ par train à l'échelle d'un trajet est significativement plus important et rarement en dessous de 50, voir la Figure 5.2.

Résumé.

- Les déplacements à l'intérieur des rames transparaissent au travers d'un fort déséquilibre entre la distribution par zone des montées et des descentes en fin de trajet ;
- Nous observons rarement moins de 50 montées à l'échelle d'un trajet cependant il n'est pas rare d'observer moins de 20 montées par arrêt à l'échelle d'un train ;
- Nous estimons la répartition des voyageurs par zone qui n'est pas uniforme entre les zones, voir la Figure 5.1.

1. À Transilien, la capacité est le nombre de voyageurs qu'une rame peut accueillir. Elle est définie comme la somme de la capacité assise et de la capacité debout. La capacité debout est calculée en supposant que la rame peut accueillir au plus 4 personnes par mètre carré.

5.4 Modélisation des déplacements à l'échelle du trajet

Pour réduire ce déséquilibre entre le nombre de montées et de descentes par zone à l'échelle d'un trajet, nous proposons de modéliser les déplacements entre les différentes zones d'une rame à l'aide d'une matrice de passage $\boldsymbol{P}$ de dimension $I \times I$. Dans cette section, nous considérons que les déplacements des passagers de leur zone de montée à leur zone de trajet (et par hypothèse de descente) ne dépendent pas de la gare de montée *i.e.* $\boldsymbol{P}$ ne dépend pas de s. Dans ce cas simplifié, la matrice stochastique $\boldsymbol{P}$ est donc :

$$\boldsymbol{P} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,I} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,I} \\ \vdots & \vdots & \ddots & \vdots \\ p_{I,1} & p_{I,2} & \cdots & p_{I,I} \end{pmatrix}. \tag{5.2}$$

où $p_{i,j}$ est la probabilité de monter en i et de se déplacer en j. Nous proposons d'estimer ces probabilités de déplacement de deux manières. La première à partir d'un modèle de régression linéaire multivariée permettant d'estimer $\boldsymbol{P}$ par minimisation des moindres carrés sous la contrainte que $\boldsymbol{P}$ est une matrice stochastique. Dans ce cas, le vecteur $\boldsymbol{a}_\bullet$ des descentes est le vecteur réponse et le vecteur $\boldsymbol{b}_\bullet$ des montées contient les variables explicatives. La seconde à partir d'un modèle probabiliste posé sur le vecteur des descentes dont la loi, d'espérance $\boldsymbol{b}_\bullet \boldsymbol{P}$, est approximée par une loi multinomiale de même espérance. Ce modèle probabiliste permet d'estimer $\boldsymbol{P}$ par maximum de vraisemblance.

5.4.1 Modélisation des déplacements par moindres carrés sous contraintes

Nous considérons que le vecteur ligne des descentes en fin de trajet $\boldsymbol{a}_\bullet = (a_{\bullet,1}, \ldots, a_{\bullet,I})$ est le vecteur des variables à expliquer d'un problème de régression linéaire multivariée sous contraintes. Les variables explicatives sont le vecteur ligne des montées par zone $\boldsymbol{b}_\bullet = (b_{\bullet,1}, \ldots, b_{\bullet,I})$. Les coefficients sont les $p_{i,j}$ de la matrice de passage stochastique $\boldsymbol{P}$. Les contraintes sont que les probabilités en ligne se somment à un et que chaque entrée $p_{i,j}$ est comprise entre 0 et 1. Le problème à résoudre est le suivant :

$$\boldsymbol{A}_\bullet = \boldsymbol{b}_\bullet \boldsymbol{P} + \varepsilon, \tag{5.3}$$

où ε est le vecteur des résidus. Nos observations sont l'ensemble des trajets (k, d) qui permettent de résoudre le problème des moindres carrés de l'équation (5.4).

$$\sum_{(k,d)\in\mathcal{N}} \left\| \boldsymbol{a}_{\bullet}^{k,d} - \boldsymbol{b}_{\bullet}^{k,d}\boldsymbol{P} \right\|_2^2 = \sum_{(k,d)\in\mathcal{N}} \sum_{j=1}^{I} \left(a_{\bullet,j}^{k,d} - \sum_{i=1}^{I} b_{\bullet,i}^{k,d} p_{i,j} \right)^2 \tag{5.4}$$

L'estimation de $\boldsymbol{P}$ est la solution du problème (5.5) des moindres carrés :

$$\begin{aligned} \underset{\boldsymbol{P}}{\operatorname{argmin}} \quad & \sum_{(k,d)\in\mathcal{N}} \left\| \boldsymbol{a}_{\bullet}^{k,d} - \boldsymbol{b}_{\bullet}^{k,d}\boldsymbol{P} \right\|_2^2 \\ \text{s.c} \quad & \boldsymbol{P} \text{ est stochastique} \end{aligned} \tag{5.5}$$

Le problème (5.5) est un problème d'optimisation quadratique sous contraintes d'égalité et d'inégalité linéaires que nous résolvons numériquement avec la fonction `lsqlincon` du package `pracma`. L'ajout de contraintes d'inégalité empêche l'écriture d'une solution analytique simple du problème.

5.4.2 Modélisation des déplacements par maximum de vraisemblance

Nous proposons un modèle probabiliste associé aux descentes pour estimer une matrice de passage $\boldsymbol{P}$ en maximisant sa vraisemblance par rapport aux descentes. Ce modèle part du constat de la section 5.3 concernant le déséquilibre de montées et de descentes par zone en fin de trajet. Puisque qu'aucun voyageur n'a pu disparaître, il y a donc forcément eu des déplacements à bord des rames.

Déplacement des voyageurs après être montés

Une rame est décomposée en I zones, nous supposons que les déplacements depuis une zone suivent une loi multinomiale à I classes associées aux I zones accessibles. Ainsi, nous définissons les déplacements des voyageurs étant montés en zone i à l'aide d'un vecteur aléatoire, noté $\boldsymbol{U}_{\bullet,i}$, à I composantes. $\boldsymbol{U}_{\bullet,i}$ suit une loi multinomiale $\mathcal{M}(b_{\bullet,i}, p_{i,1}, \ldots, p_{i,I})$ où $b_{\bullet,i}$ est le nombre de montées en zone i et $p_{i,j}$ est la probabilité de se déplacer en zone j lorsqu'on est monté en zone i. L'ensemble des probabilités forme la matrice de passage, notée $\boldsymbol{P}$. Nous supposons que les déplacements des voyageurs montés par une zone sont indépendants des déplacements des voyageurs montés par une autre zone ($\forall i \neq j,\ \boldsymbol{U}_{\bullet,i} \perp\!\!\!\perp \boldsymbol{U}_{\bullet,j}$).

Vecteur des descentes égal à la somme des déplacements

Nous posons que le vecteur des descentes par zone, noté $\boldsymbol{A}_{\bullet}$, est la somme par zone des voyageurs s'étant déplacés jusqu'aux différentes zones, c'est-à-dire :

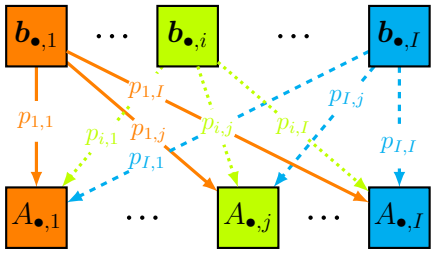


FIGURE 5.3 – Modélisation des descentes à partir des montées : l'importance d'une matrice de passage

$$\boldsymbol{A}_\bullet = \sum_{j=1}^{I} \boldsymbol{U}_{\bullet,j}.$$

$\boldsymbol{A}_\bullet$ est modélisée comme la somme de lois multinomiales indépendantes non identiquement distribuées. Nous ne connaissons pas la loi de cette somme, ainsi nous proposons l'approximation 1 de la loi de $\boldsymbol{A}$ dont la qualité est discutée dans l'Annexe 5.A.

Approximation 1 *La loi de $\boldsymbol{A}_\bullet$ est approchée par la loi multinomiale de même espérance : $\mathcal{M}(b_{\bullet,\bullet}, \pi_{\bullet,1}, \ldots, \pi_{\bullet,I})$ avec $\pi_{\bullet,j} = \sum_{i=1}^{I} r_{\bullet,i} p_{i,j}$ où $r_{\bullet,i} = b_{\bullet,i}/b_{\bullet,\bullet}$.*

Nous illustrons sur la Figure 5.3 les liens entre les montées et les descentes par zone. Le rôle de la matrice de passage $\boldsymbol{P}$ est représenté par des arcs orientés de $\boldsymbol{b}_\bullet$ vers $\boldsymbol{A}_\bullet$ et pondérés par la probabilité des voyageurs $(p_{i,j})$ de se déplacer de i vers j.

Suivant l'approximation 1, nous exprimons la densité de $\boldsymbol{A}_\bullet$ comme :

$$\mathbb{P}(\boldsymbol{A}_\bullet = \boldsymbol{a}_\bullet; \boldsymbol{b}_\bullet) = \prod_{j=1}^{I} \frac{b_{\bullet,\bullet}!}{a_{\bullet,j}!} \left(\sum_{i=1}^{I} r_{\bullet,i} p_{i,j} \right)^{a_{\bullet,j}},$$

où les montées sont des paramètres. La partie de la log-vraisemblance associée à une observation (k, d) ne dépendant que des paramètres est :

$$\ell(\boldsymbol{a}_\bullet^{k,d}; \boldsymbol{P}, \boldsymbol{b}_\bullet^{k,d}) = \sum_{j=1}^{I} a_{\bullet,j}^{k,d} \log \left(\sum_{i=1}^{I} r_{\bullet,i}^{k,d} p_{i,j} \right).$$

Puisque toutes les observations sont indépendantes et tirées suivant la même loi, nous obtenons que la log-vraisemblance, à facteurs additifs, est :

$$\ell(\boldsymbol{P}) = \sum_{(k,d)\in\mathcal{N}} \sum_{j=1}^{I} a_{\bullet,j}^{k,d} \log \left(\sum_{i=1}^{I} r_{\bullet,i}^{k,d} p_{i,j} \right).$$

Nous maximisons numériquement la log-vraisemblance, sous contraintes d'inégalités et d'égalités linéaires, issue du problème d'optimisation convexe (5.6) à l'aide du

package `Rsolnp`.

$$\begin{aligned} \underset{\boldsymbol{P}}{\text{argmax}} \quad & \sum_{(k,d)\in\mathcal{N}} \sum_{j=1}^{I} a_{\bullet,j}^{k,d} \log\left(\sum_{i=1}^{I} r_{\bullet,i}^{k,d} p_{i,j}\right) \\ \text{s.c} \quad & \boldsymbol{P} \text{ est stochastique} \end{aligned} \tag{5.6}$$

5.4.3 Résultats à l'échelle du trajet : moindres carrés ou log-vraisemblance, quelles différences ?

Nous comparons sur un jeu de données les deux méthodes d'estimation des matrices de passage par rame, la première par minimisation des moindres carrés, notée $\widehat{\boldsymbol{P}}_{\text{MLS}}$, la seconde par maximisation de la log-vraisemblance, notée $\widehat{\boldsymbol{P}}_{\text{MLE}}$. Nous comparons la qualité de prévision des descentes ($\boldsymbol{a}_\bullet$) par les montées ($\boldsymbol{b}_\bullet$) avec ou sans déplacements. La stratégie sans déplacements consiste à appliquer comme matrice de passage la matrice identité. Cette dernière stratégie est la solution retenue pour le moment par Transilien.

$$\begin{aligned} \text{MAE} &= \frac{1}{N_{\mathcal{T}\text{est}}} \sum_{(k,d)\in\mathcal{T}\text{est}} \left\| \boldsymbol{a}_\bullet^{k,d} - \widehat{\boldsymbol{a}}_\bullet^{k,d} \right\|_1 \\ \text{MAPE} &= \frac{1}{N_{\mathcal{T}\text{est}}} \sum_{(k,d)\in\mathcal{T}\text{est}} \left\| \frac{\boldsymbol{a}_\bullet^{k,d} - \widehat{\boldsymbol{a}}_\bullet^{k,d}}{\boldsymbol{a}_\bullet^{k,d}} \right\|_1 \\ \text{RMSE} &= \sqrt{\frac{1}{N_{\mathcal{T}\text{est}}} \sum_{(k,d)\in\mathcal{T}\text{est}} \left\| \boldsymbol{a}_\bullet^{k,d} - \widehat{\boldsymbol{a}}_\bullet^{k,d} \right\|_2^2} \end{aligned}$$

.

Le jeu de données comprend plus de 3 500 trajets de Gare du Nord à Pontoise sur la ligne H entre janvier et septembre 2021 pour les jours ouvrés hors vacances et jours fériés. Pour évaluer les performances de modélisation des deux méthodes à l'aide de métriques standards, nous découpons le jeu de données en un jeu de données d'apprentissage (75%) et un jeu de données test (25%), noté $\mathcal{T}\text{est}$ et de taille $N_{\mathcal{T}\text{est}}$. Les métriques calculées sur le jeu de données test pour évaluer la qualité d'estimation, avec $\widehat{\boldsymbol{a}}_\bullet = \boldsymbol{b}_\bullet \widehat{\boldsymbol{P}}$, sont la moyenne des erreurs absolues (MAE), la moyenne du pourcentage des erreurs absolues (MAPE) et la racine de la moyenne des erreurs au carré (RMSE).

Les déplacements pour correctement estimer le vecteur des descentes à l'échelle du trajet. La Table 5.5 permet de confirmer que la méthode naïve qui consiste à ne pas déplacer les voyageurs après qu'ils sont montés est mauvaise. En effet, ses performances d'estimation du vecteur des descentes sont deux fois moins bonnes qu'une méthode avec déplacements. Ce décalage est d'autant plus frappant

pour la rame arrière, ce qui s'explique par un quai terminus à Gare du Nord qui incite les voyageurs à entrer en queue de train, puis à marcher à bord du train pour trouver une place assise. Les performances de P_{MLS} et P_{MLE} sont très similaires, bien que les matrices de passage ne soient pas parfaitement identiques.

La matrice de passage révèle les déplacements à bord. La force de notre modélisation est de donner une interprétation simple des coefficients estimés. Nous illustrons de deux manières les déplacements entre les zones sur la Figure 5.4. La première, plus synthétique permet d'identifier des formes géométriques de dépendance entre les zones. Par exemple, la ligne de carrés bleu clair de la zone 8, qui indique qu'il y a beaucoup de déplacements depuis cette zone, ou la diagonale de carrés bleu foncé, qui indique que la majorité des voyageurs reste là où ils sont montés. La représentation en graphe permet de mieux appréhender les distances qui séparent les différentes zones. Le passage des zones 8 à 1, qui sont très éloignées, est courant tandis que le déplacement de 1 vers 8 est rare.

Des différences faibles entre méthodes mais notables entre rames arrière et avant. Lorsque nous comparons les matrices de passage sur la Figure 5.5, nous remarquons que les différences entre les méthodes sont faibles, toutefois pour la rame arrière les différences sont un peu moins négligeables que pour la rame avant. Ce constat est particulièrement vrai pour la dernière zone qui est aussi la zone la plus sujette au déplacement des passagers. Par ailleurs, nous notons que les matrices de passage entre la rame avant et la rame arrière sont différentes. Ce dernier constat fait que nous ne pouvons pas utiliser une seule matrice de passage quelle que soit la position de la rame. La connaissance de la position de la rame et son sens de circulation sont donc nécessaires pour pouvoir modéliser correctement les déplacements.

Résumé. La comparaison des performances d'estimation des descentes et des matrices de passage montre d'une part que la modélisation des déplacements permet de diminuer par deux l'erreur d'estimation de descentes, et d'autre part, qu'il n'y a pas de différence majeure entre les deux méthodes. Le problème des

TABLE 5.5 – Performances mesurées en MAE, MAPE et RMSE en fonction de la matrice de passage estimée pour les rames avant et arrière. En ligne, les différentes méthodes d'estimation : moindres carrés (Section 5.4.1), maximum de vraisemblance (Section 5.4.2) ou sans déplacements (matrice identité).

	Avant			Arrière		
Modèles	MAE [voy]	MAPE (%)	RMSE [voy]	MAE [voy]	MAPE (%)	RMSE [voy]
Sans déplacements	10.9	35.5	16.6	17.5	55.7	29.9
$\widehat{\boldsymbol{P}}_{\text{MLS}}$	6	33.1	8.3	8.5	38.8	11.4
$\widehat{\boldsymbol{P}}_{\text{MLE}}$	6	31.9	8.4	8.5	37.8	11.6

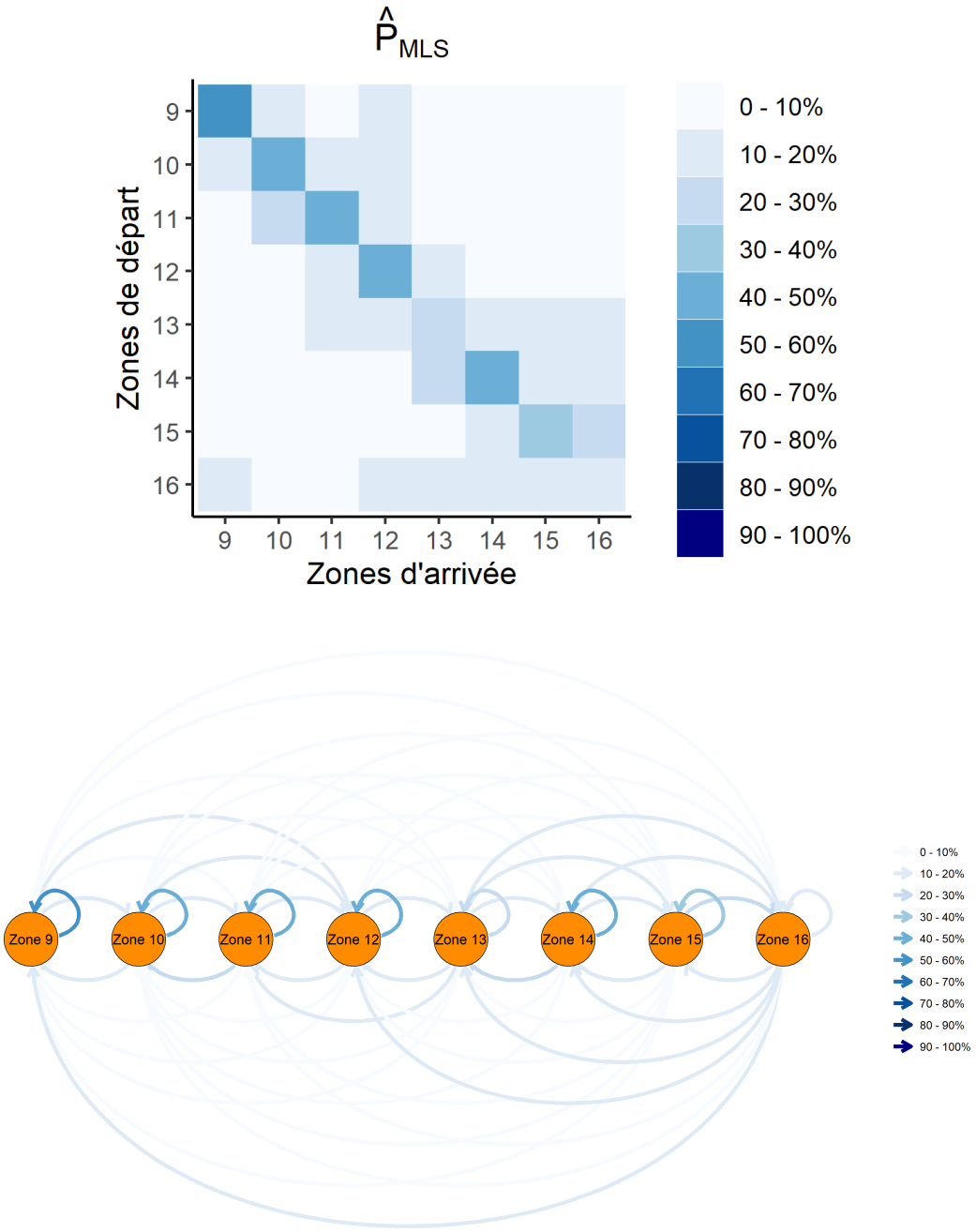


FIGURE 5.4 – Représentation de la matrice de passage estimée par moindres carrés pour la rame arrière (9-16) des trains allant de Gare du Nord vers la banlieue. Graphique du haut : représentation sous forme matricielle. Graphique du bas : représentation sous forme de graphe.

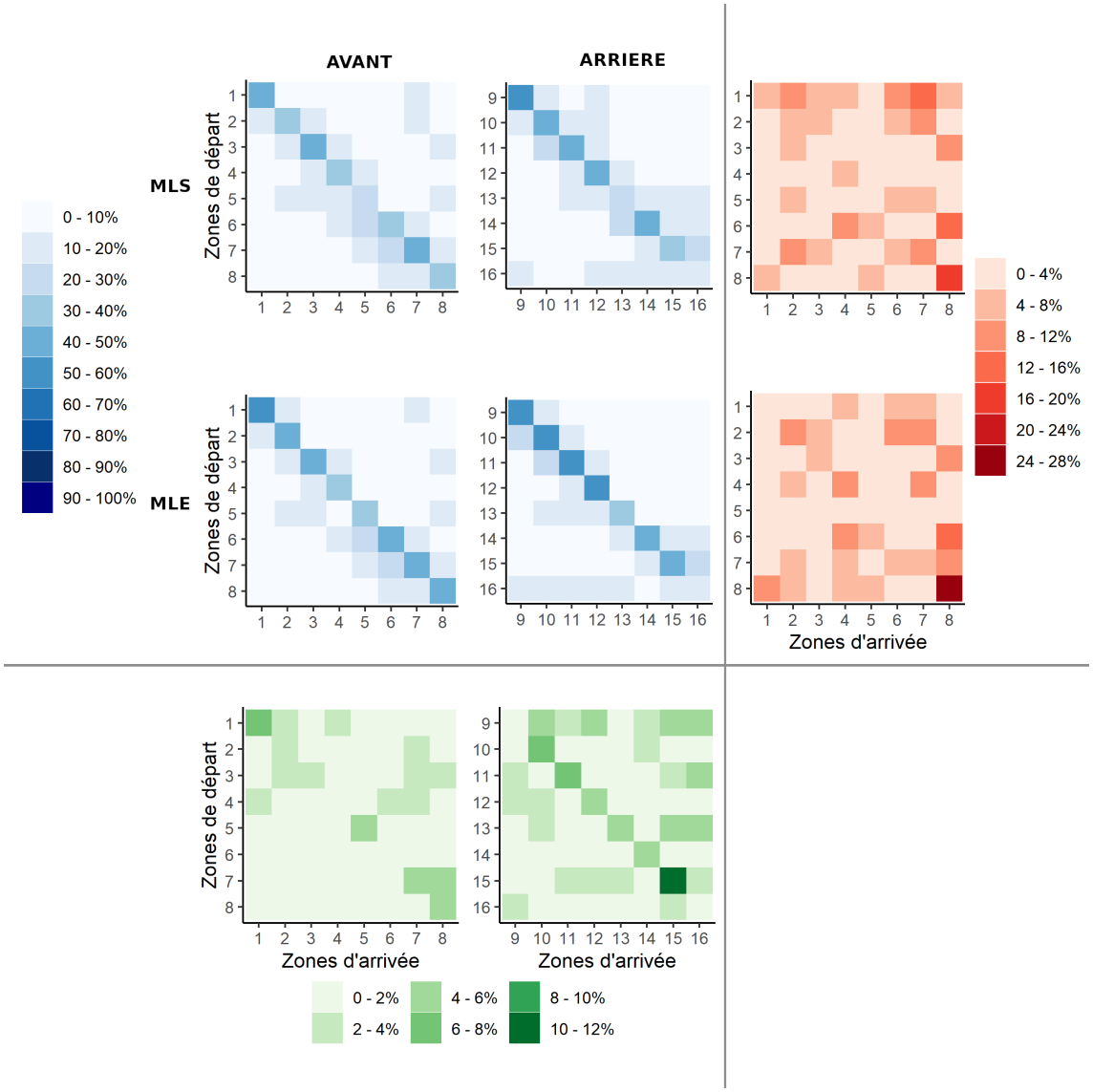


FIGURE 5.5 – Comparaison des matrices de passage suivant la méthode d'estimation et/ou la position de la rame. Les quatre graphiques centraux représentent les matrices de passage pour chaque méthode (en ligne) et chaque position de la rame (en colonne). Les deux graphiques rouges (à droite) représentent pour chaque méthode la différence en points de pourcentage des coefficients entre la rame avant et arrière. Les deux graphiques verts (en bas) représentent la différence en points de pourcentage entre les deux méthodes pour une même position de rame.

moindres carrés est beaucoup plus rapide à résoudre mais l'intégration de variables latentes est compliquée. Enfin, il faut prendre en compte le sens de circulation et la position de la rame pour modéliser les déplacements des voyageurs.

5.5 Modélisation des déplacements à l'échelle de la gare

Dans la Section 5.4, nous avons considéré les montées et descentes par zone agrégées à l'échelle du trajet, c'est-à-dire sommées sur l'ensemble des S gares du trajet entre la gare origine $s = 1$ et la gare terminus $s = S$. La modélisation associée reposait, quelle que soit la gare de montée, sur un seul type de déplacement de zone à zone. Par exemple, elle négligeait l'impact de la position des entrées/sorties du quai sur la répartition des montées et des descentes. Or, ces effets sont bien documentés [Szplett and Wirasinghe, 1984, Krstanoski, 2014], et ils engendrent des déplacements à bord différents selon les gares où les voyageurs montent ou descendent.

Dans cette section, nous levons la limitation principale de la modélisation précédente en permettant des déplacements spécifiques à chaque gare. Nous travaillons donc à partir des montées et descentes à chaque gare et à chaque zone. Ainsi, l'observation pour un trajet (k, d) donné du train k le jour d est formée de l'ensemble des vecteurs du nombre de montées à chaque gare, noté $\boldsymbol{b}^{k,d}_{1:(S-1)} = (\boldsymbol{b}^{k,d}_1, \dots, \boldsymbol{b}^{k,d}_{S-1})$ et celui des vecteurs du nombre de descentes à chaque gare, noté $\boldsymbol{a}^{k,d}_{2:S} = (\boldsymbol{a}^{k,d}_2, \dots, \boldsymbol{a}^{k,d}_S)$. Les $\boldsymbol{b}_s$ et $\boldsymbol{a}_s$ sont des vecteurs de dimension I, le nombre de zones considérées. Le vecteur du nombre de montées au terminus $\boldsymbol{b}^{k,d}_S$ et celui du nombre de descentes à la gare origine $\boldsymbol{a}^{k,d}_1$ étant toujours nuls, nous avons choisi de ne pas les inclure dans les observations. Nous commençons par définir le modèle et ses hypothèses, puis nous exprimons sa vraisemblance. Nous discutons *in fine* de l'estimation par maximum de vraisemblance des paramètres du modèle qui incluent des probabilités de déplacements $p_{s,i,j}$.

Notation. Afin d'alléger l'écriture, nous supprimons la référence au trajet (k, d) dans la notation pour l'étude d'un trajet générique, et nous ne la réintroduisons que lorsque l'ensemble des trajets sera impliqué.

5.5.1 Modèles et hypothèses

Notre objectif est de modéliser les descentes en fonction des montées, qui sont considérées comme des variables explicatives. La modélisation doit prendre en compte la spécificité de chaque gare concernant aussi bien les déplacements d'une zone à l'autre que les descentes. En première approche, nous rappelons que nous considérons que les voyageurs, une fois qu'ils sont montés à la gare s en zone i, se déplacent instantanément à la zone j, dans laquelle ils feront leur voyage et de

laquelle ils descendront, sans aucun autre déplacement ultérieur. De plus, nous considérons que les trajets sont indépendants entre eux.

Des descentes qui révèlent la charge à bord

Il est naturel d'envisager que le nombre de descentes à une zone i et une gare s dépend de la charge à bord, qui est la somme cumulée de la différence entre le nombre de voyageurs s'étant déplacés jusqu'à cette zone depuis l'origine et le nombre de voyageurs descendus aux gares précédentes. Notons $\ell_{s-1,i}$ la charge à bord en sortie de gare $s-1$, après les descentes et le déplacement des voyageurs montés en gare $s-1$. Cette charge à bord, définie en tenant compte des descentes et des déplacements de voyageurs montés jusqu'en gare $s-1$, est identique à celle en entrée de gare s puisque les voyageurs se déplacent pour aller directement à leur zone de descente. C'est donc elle qui influe sur le nombre de descentes à la gare suivante s.

Nous considérons ainsi que, connaissant le vecteur des charges à bord $\boldsymbol{\ell}_{s-1}$ en sortie de gare $s-1$, les coordonnées du vecteur des descentes $\boldsymbol{A}_s = (\boldsymbol{A}_{s,1}, \ldots, \boldsymbol{A}_{s,I})$ depuis les différentes zones à la gare s sont indépendantes, de loi binomiale dépendant de la zone i et la gare s, plus précisément,

$$\boldsymbol{A}_{s,i} \sim \mathcal{B}(c_{s-1,i}, \alpha_{s,i}), \quad \forall i = 1, \ldots I, \forall s = 2, \ldots, S$$

où $\alpha_{s,i}$ est la probabilité de descendre. En effet, tous les voyageurs montant à une gare ne descendent pas tous simultanément à la gare suivante. Ainsi, la charge à bord est une variable inaccessible directement (variable latente/cachée), dépendant d'une part des déplacements successifs vers les différentes zones (dépendant eux-mêmes des montées) et qui sont inconnus, d'autre part des descentes aux gares précédentes (qui sont elles observées et connues). Il est donc important de définir un modèle de déplacement permettant d'inférer la charge à bord.

Modélisation des déplacements

Nous reprenons l'idée introduite dans la Section 5.4, concernant la modélisation des déplacements d'une zone à l'autre après être monté. Les voyageurs montant à la gare s en zone i sont supposés se déplacer vers les autres zones suivant une loi multinomiale $\boldsymbol{U}_{s,i} \sim \mathcal{M}(b_{s,i}, p_{s,i,1}, \ldots, p_{s,i,I})$. Les paramètres de chaque loi multinomiale ($i = 1, \ldots, I$) dépendent non seulement de la zone (comme à la Section 5.4) mais aussi de la gare $s = 1, \ldots, S-1$ (ce qui est nouveau par rapport à la Section 5.4). Nous supposons aussi que les voyageurs se déplaçant d'une zone à l'autre ne se gênent pas mutuellement, et que les déplacements depuis les différentes zones sont indépendants. Ainsi, le vecteur des montées après déplacement à la gare s (que nous appelons plus simplement déplacements), noté $\boldsymbol{W}_s$, est la somme des I lois multinomiales indépendantes mais non identiquement

distribuées :

$$\boldsymbol{W}_s = \sum_{i=1}^{I} \boldsymbol{U}_{s,i}.$$

Nous proposons pour la loi jointe des déplacements $\boldsymbol{W}_s = (W_{s,1}, \ldots, W_{s,I})$ la même approximation que celle utilisée dans la Section 5.4 pour les descentes, approximation qu'il faudrait justifier. Ainsi, $\boldsymbol{W}_s \sim \mathcal{M}(b_{s,\bullet}, \pi_{s,1}, \ldots, \pi_{s,I})$, les probabilités $\pi_{s,j} = \sum_{i=1}^{I} \left(b_{s,i}/b_{s,\bullet}\right) p_{s,i,j}$ dépendant maintenant de la gare s. Le vecteur aléatoire $\boldsymbol{W}$ est, tout comme le vecteur de la charge à bord, non observé. Il permet d'exprimer la charge à bord de la façon suivante :

$$\boldsymbol{L}_s = \sum_{g=1}^{s} \boldsymbol{W}_g - \boldsymbol{A}_g, \tag{5.7}$$

faisant le lien entre la modélisation des déplacements et celle des descentes.

Formulation du modèle

L'encart suivant résume les hypothèses que nous venons d'introduire pour définir le modèle des descentes, en utilisant les indices de trajet (k, d).

Modèle 1 *Le modèle des descentes à une gare s combine la prise en compte des montées b et des descentes A aux gares précédentes grâce à l'introduction, comme dans l'équation (5.7), de variables latentes de déplacement W.*

(A_0) *Les trajets sont indépendants suivant les jours d et les trains k.*

Pour un trajet (k, d) :

(A_{1a}) *Les déplacements conditionnellement aux montées suivent une loi multinomiale*

$$\boldsymbol{W}_s^{k,d} \sim \mathcal{M}(b_{s,\bullet}^{k,d}, \pi_{s,1}^{k,d}, \ldots, \pi_{s,I}^{k,d}), \quad s = 1, \ldots, S-1$$

où $\pi_{s,j}^{k,d} = \sum_{i=1}^{I} r_{s,i}^{k,d} p_{s,i,j}$ avec $r_{s,i}^{k,d} = b_{s,i}^{k,d}/b_{s,\bullet}^{k,d}$.

(A_{1b}) *Les déplacements aux différentes gares $\boldsymbol{W}_s^{k,d}$, $s = 1, \ldots, S-1$, sont indépendants entre eux et indépendants des montées aux gares $s' \neq s$.*

(A_{2a}) *La loi des descentes en zone i à la gare s conditionnellement au passé ne dépend que de la charge à bord en sortie de gare $s-1$:*

$$\mathbb{P}\left(A_{s,i}^{k,d} \middle| a_{2:(s-1),i}^{k,d}, z_{1:(s-1),i}^{k,d}\right) = \mathbb{P}\left(A_{s,i}^{k,d} \middle| \ell_{s-1,i}^{k,d}\right), \quad s = 2, \ldots, S.$$

(A_{2b}) *Les descentes en zone i à la gare s conditionnellement à la charge à bord $\ell_{s-1,i}^{k,d}$ en sortie de gare $s-1$ suivent une loi binomiale :*

$$A_{s,i}^{k,d} \sim \mathcal{B}(\ell_{s-1,i}^{k,d}, \alpha_{s,i}), \quad s = 2, \ldots, S-1.$$

(A_{2c}) *Pour toute gare s, les coordonnées du vecteur des descentes $\boldsymbol{A}_s^{k,d}$ à cette gare sont indépendantes conditionnellement à la charge bord en entrée de gare $\boldsymbol{\ell}_s^{k,d}$.*

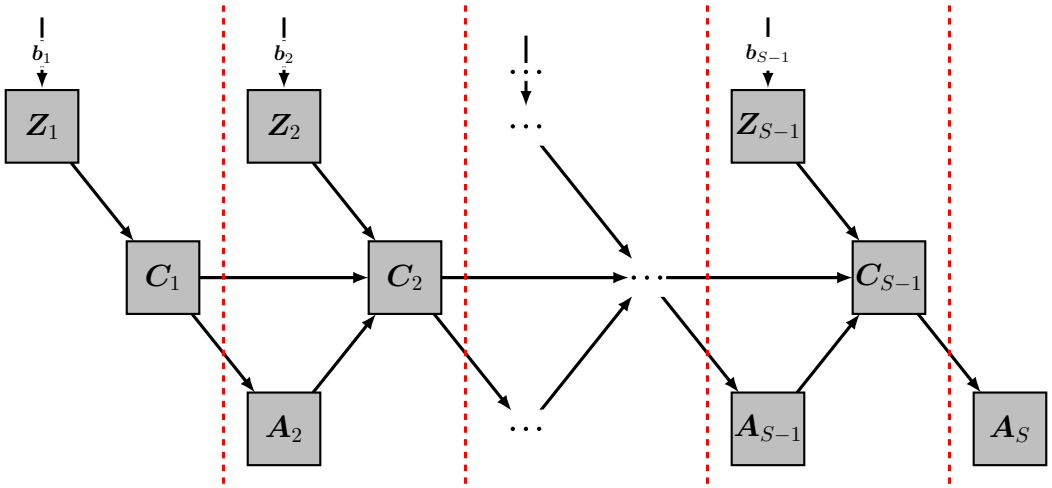


FIGURE 5.6 – Relation séquentielle entre les variables au fil du trajet.

Remarques.

- Les paramètres de la loi multinomiale $\pi_{s,1}^{k,d}$ en (A_1) ne dépendent du trajet que par le nombre de montées. Les paramètres $p_{s,i,j}$ quant à eux sont les mêmes pour tous les trajets (k,d). Ils peuvent être regroupés en $S-1$ matrices stochastiques $\boldsymbol{P}_{1:(S-1)} = (\boldsymbol{P}_1, \ldots, \boldsymbol{P}_{S-1})$.
- La loi proposée en (A_2) fait uniquement dépendre le nombre de descentes à la gare s de la charge à bord à la gare précédente, ce qui est une hypothèse markovienne.
- (A_3) explicite la contrainte temporelle de déroulement du trajet.

5.5.2 Expression de la vraisemblance avec variables cachées

Le modèle 1 définit la loi des observations des vecteurs $\boldsymbol{A}_{2:S}^{k,d}$ des descentes à toutes les gares des trajets (k,d), avec les vecteurs des montées comme variables explicatives. Les paramètres à inférer sont les coefficients des $S-1$ matrices stochastiques $\boldsymbol{P}_{1:(S-1)}$ représentant les probabilités de déplacement d'une zone i à une zone j à chaque gare, et les probabilités $\boldsymbol{\alpha}_{2:S}$ des descentes de chaque zone à chaque gare. On remarque que puisque tous les voyageurs descendent au terminus $\boldsymbol{\alpha}_S = (1, \ldots, 1)$, il y a donc $I[(I-1)(S-1)+(S-2)]$ paramètres à estimer, regroupés sous la notation synthétique $\boldsymbol{\theta} = (\boldsymbol{P}_{1:(S-1)}, \boldsymbol{\alpha}_{2:(S-1)})$. Il faut de plus inférer les variables latentes de charge à bord $\boldsymbol{L}_{1:(S-1)}$ (ou de façon équivalente de déplacements $\boldsymbol{W}_{1:(S-1)}$) utilisées pour définir le modèle.

Fonctions de masse et vraisemblance

Ayant défini une modélisation statistique des observations, la méthode du maximum de vraisemblance est utilisée pour l'estimation. Il faut ainsi déterminer la forme de la vraisemblance des descentes observées, qui est une fonction des paramètres calculée

sur les trajets observés. Les observations étant discrètes, la vraisemblance s'exprime comme la fonction de masse vue comme une fonction du paramètre $\boldsymbol{\theta}$. Les trajets étant indépendants selon l'hypothèse (A_0), la fonction de masse associée à l'ensemble des descentes est le produit des fonctions de masse associées aux descentes $\boldsymbol{a}_{2:S}^{k,d}$ pour chaque trajet (k,d) :

$$\prod_{k,d} \mathbb{P}\Big(\boldsymbol{A}_{2:S}^{k,d} = \boldsymbol{a}_{2:S}^{k,d}; \boldsymbol{b}_{1:(S-1)}^{k,d}, \boldsymbol{\theta}\Big).$$

Il s'agit maintenant de déterminer la fonction de masse associée à un trajet (k,d), que nous notons génériquement $\mathbb{P}(\boldsymbol{a}_{2:S}; \boldsymbol{b}_{1:(S-1)}, \boldsymbol{\theta})$ pour alléger les notations. Les descentes n'étant pas indépendantes les unes des autres le long du trajet, l'expression ne se factorise pas en fonction des gares. De plus, la présence des variables latentes de déplacement $\boldsymbol{W}_{1:(S-1)}$ complique la situation. Nous utilisons alors une stratégie typique des modèles à variables latentes en intégrant la loi jointe des déplacements cachés et des descentes, celle-ci étant accessible. Ainsi,

$$\mathbb{P}\Big(\boldsymbol{a}_{2:S}; \boldsymbol{b}_{1:(S-1)}, \boldsymbol{\theta}\Big) = \sum_{\boldsymbol{w}_{1:(S-1)} \in \mathcal{W}} \mathbb{P}\Big(\boldsymbol{a}_{2:S}, \boldsymbol{w}_{1:(S-1)}; \boldsymbol{b}_{1:(S-1)}, \boldsymbol{\theta}\Big), \tag{5.8}$$

où $\mathcal{W}$ est l'ensemble des combinaisons possibles de déplacements des voyageurs montés $\boldsymbol{b}_s$ à chaque gare $s \in \{1, \ldots, S-1\}$ entre les I zones. Cet ensemble croit de façon exponentielle avec le nombre d'arrêts, de zones et de montées.

Loi jointe des déplacements et des descentes

Pour déterminer l'expression de la loi jointe des déplacements et des descentes, nous procédons par récurrence en conditionnant les observations à chaque gare par son passé, c'est-à-dire par les événements aux gares précédentes, en commençant par la gare terminus. Ainsi, en appliquant la règle de Bayes à la gare terminus S, nous obtenons :

$$\begin{aligned}
&\mathbb{P}\Big(\boldsymbol{a}_{2:S}, \boldsymbol{w}_{1:(S-1)}; \boldsymbol{b}_{1:(S-1)}, \boldsymbol{\theta}\Big) \\
&\quad = \mathbb{P}\Big(\underbrace{\boldsymbol{a}_S}_{\text{Gare S}} \,\Big|\, \underbrace{\boldsymbol{a}_{2:(S-1)}, \boldsymbol{w}_{1:(S-1)}}_{\text{Valeurs jusqu'à la gare S-1}}; \boldsymbol{b}_{1:(S-1)}, \boldsymbol{\theta}\Big) \mathbb{P}\Big(\boldsymbol{a}_{2:(S-1)}, \boldsymbol{w}_{1:(S-1)}; \boldsymbol{b}_{1:(S-1)}, \boldsymbol{\theta}\Big).
\end{aligned}$$

En appliquant la règle de Bayes deux fois sur le terme de droite, on a premièrement que :

$$\begin{aligned}
&\mathbb{P}\Big(\boldsymbol{a}_{2:(S-1)}, \boldsymbol{w}_{1:(S-1)}; \boldsymbol{b}_{1:(S-1)}, \boldsymbol{\theta}\Big) \\
&\qquad = \mathbb{P}\Big(\boldsymbol{a}_{2:(S-1)}, \boldsymbol{w}_{1:(S-2)}; \boldsymbol{b}_{1:(S-1)}, \boldsymbol{\theta}\Big) \times \mathbb{P}\Big(\boldsymbol{w}_{S-1}; \boldsymbol{b}_{1:(S-1)}, \boldsymbol{\theta}\Big),
\end{aligned} \tag{5.9}$$

que l'on peut simplifier grâce à l'hypothèse (A_{1b}) :

$$\mathbb{P}\Big(\boldsymbol{w}_{S-1}; \boldsymbol{b}_{1:(S-1)}, \boldsymbol{\theta}\Big) = \mathbb{P}\Big(\boldsymbol{w}_{S-1}; \boldsymbol{b}_{S-1}, \boldsymbol{\theta}\Big).$$

Puis deuxièmement, en appliquant la règle de Bayes au terme de gauche et en utilisant le fait que les montées sont non aléatoires, on obtient que :

$$\mathbb{P}\Big(\boldsymbol{a}_{2:(S-1)},\boldsymbol{w}_{1:(S-2)};\boldsymbol{b}_{1:(S-1)},\boldsymbol{\theta}\Big) = \mathbb{P}\Big(\underbrace{\boldsymbol{a}_{S-1}}_{\text{Gare S-1}} \Big| \underbrace{\boldsymbol{a}_{2:(S-2)},\boldsymbol{w}_{1:(S-2)}}_{\text{Valeurs jusqu'à la gare S-2}};\boldsymbol{b}_{1:(S-2)},\boldsymbol{\theta}\Big)\mathbb{P}\Big(\boldsymbol{a}_{2:(S-2)},\boldsymbol{w}_{1:(S-2)};\boldsymbol{b}_{1:(S-1)},\boldsymbol{\theta}\Big).$$

Le terme de droite de l'équation (5.9) est égal à :

$$\mathbb{P}\Big(\boldsymbol{a}_{2:(S-1)},\boldsymbol{w}_{1:(S-1)};\boldsymbol{b}_{1:(S-1)},\boldsymbol{\theta}\Big) = \mathbb{P}\Big(\underbrace{\boldsymbol{a}_{S-1}}_{\text{Gare S-1}} \Big| \underbrace{\boldsymbol{a}_{2:(S-2)},\boldsymbol{w}_{1:(S-2)}}_{\text{Valeurs jusqu'à la gare S-2}};\boldsymbol{b}_{1:(S-2)},\boldsymbol{\theta}\Big) \times \mathbb{P}\Big(\boldsymbol{w}_{S-1};\boldsymbol{b}_{1:(S-1)},\boldsymbol{\theta}\Big)\times \mathbb{P}\Big(\boldsymbol{a}_{2:(S-2)},\boldsymbol{w}_{1:(S-2)};\boldsymbol{b}_{1:(S-1)},\boldsymbol{\theta}\Big).$$

En itérant successivement, la loi jointe devient :

$$\mathbb{P}\Big(\boldsymbol{a}_{2:S},\boldsymbol{w}_{1:(S-1)};\boldsymbol{b}_{1:(S-1)},\boldsymbol{\theta}\Big) = \prod_{s=2}^{S}\mathbb{P}\Big(\boldsymbol{a}_s\Big|\boldsymbol{w}_{1:(s-1)};\boldsymbol{b}_{1:(s-1)},\boldsymbol{\theta}\Big) \times \mathbb{P}\Big(\boldsymbol{w}_{s-1};\boldsymbol{b}_{s-1},\boldsymbol{\theta}\Big). \tag{5.10}$$

Dans l'équation (5.10), les descentes à chaque gare s sont conditionnelles à tout le passé, de l'origine à la gare s. En introduisant les charges à bord $\boldsymbol{\ell}_{s-1}$ en entrée de gare s, puis en utilisant l'hypothèse (A_{2a}), nous écrivons :

$$\mathbb{P}\Big(\boldsymbol{a}_s\Big|\boldsymbol{w}_{1:(s-1)};\boldsymbol{b}_{1:(s-1)},\boldsymbol{\theta}\Big) = \mathbb{P}\Big(\boldsymbol{a}_s\Big|\boldsymbol{\ell}_{s-1};\boldsymbol{\theta}\Big).$$

Il reste à utiliser l'hypothèse (A_{2c}) pour factoriser la loi jointe par zone :

$$\mathbb{P}\Big(\boldsymbol{a}_s\Big|\boldsymbol{\ell}_{s-1};\boldsymbol{\theta}\Big) = \prod_{i=1}^{I}\mathbb{P}\Big(a_{s,i}\Big|\ell_{s-1,i};\boldsymbol{\theta}\Big).$$

En résumé, l'expression de la loi jointe des déplacements et des descentes est égale à :

$$\mathbb{P}\Big(\boldsymbol{a}_{2:S},\boldsymbol{w}_{1:(S-1)};\boldsymbol{b}_{1:(S-1)},\boldsymbol{\theta}\Big) = \prod_{s=2}^{S}\left(\prod_{i=1}^{I}\mathbb{P}\Big(a_{s,i}\Big|\ell_{s-1,i};\boldsymbol{\theta}\Big)\right)\mathbb{P}\Big(\boldsymbol{w}_{s-1};\boldsymbol{b}_{s-1},\boldsymbol{\theta}\Big)$$

En appliquant le modèle des déplacements (A_{1a}) de loi multinomiale et le modèle de descentes (A_{2a}) de loi binomiale, la loi jointe des déplacements et des descentes s'écrit :

$$\mathbb{P}\Big(\boldsymbol{a}_{2:S},\boldsymbol{w}_{1:(S-1)};\boldsymbol{b}_{1:(S-1)},\boldsymbol{\theta}\Big) = \prod_{s=2}^{S}\underbrace{\left(\prod_{i=1}^{I}\binom{\ell_{s-1,i}}{a_{s,i}}(\alpha_{s,i})^{a_{s,i}}(1-\alpha_{s,i})^{(\ell_{s-1,i}-a_{s,i})}\right)}_{\mathbb{P}(\boldsymbol{a}_s|\boldsymbol{\ell}_{s-1};\boldsymbol{\theta})}\underbrace{\left(\prod_{i=1}^{I}\frac{(b_{s-1,\bullet}!)}{(w_{s-1,i}!)}(\pi_{s-1,i})^{w_{s-1,i}}\right)}_{\mathbb{P}(\boldsymbol{w}_{s-1};\boldsymbol{b}_{s-1},\boldsymbol{\theta})}. \tag{5.11}$$

La vraisemblance observée

La fonction de masse des descentes d'un trajet s'exprime en sommant la vraisemblance jointe de l'équation (5.11) des déplacements et des descentes sur l'ensemble des états cachés possibles comme dans l'équation (5.8). La vraisemblance observée associée à l'observation d'un trajet (k,d) est donc la fonction L de $\boldsymbol{\theta} = (\boldsymbol{P}_{1:(S-1)}, \boldsymbol{\alpha}_{2:S})$ définie par :

$$\begin{aligned} & L\Big(\boldsymbol{a}_{2:S}^{k,d}; \boldsymbol{b}_{1:(S-1)}^{k,d}, \boldsymbol{\theta}\Big) \\ & \quad = \sum_{\boldsymbol{w}_{1:(S-1)}^{k,d} \in \mathcal{W}} \prod_{s=2}^{S} \left(\left(\prod_{i=1}^{I} \binom{\ell_{s-1,i}^{k,d}}{a_{s,i}^{k,d}} (\alpha_{s,i})^{a_{s,i}^{k,d}} (1-\alpha_{s,i})^{\left(\ell_{s-1,i}^{k,d} - a_{s,i}^{k,d}\right)} \right) \right. \\ & \qquad \left. \left(\prod_{i=1}^{I} \frac{(b_{s-1,\bullet}^{k,d}!)}{(w_{s-1,i}^{k,d}!)} \left(\pi_{s-1,i}^{k,d}\right)^{w_{s-1,i}^{k,d}} \right) \right). \end{aligned} \tag{5.12}$$

La vraisemblance associée à tous les trajets $(k,d) \in \mathcal{N}$ s'en déduit par indépendance des trajets (A_0) :

$$L_N(\boldsymbol{\theta}) = \prod_{(k,d)\in\mathcal{N}} L\Big(\boldsymbol{a}_{2:S}^{k,d}; \boldsymbol{b}_{1:(S-1)}^{k,d}, \boldsymbol{\theta}\Big). \tag{5.13}$$

Le nombre d'états cachés croit exponentiellement avec le nombre de voyageurs et le nombre de gares et le calcul de la somme impliquée dans la vraisemblance n'est rapidement plus calculable numériquement. Cependant, le calcul de la vraisemblance n'est pas forcément nécessaire pour obtenir l'estimateur de son maximum. En effet, dans le cas de modèles à variables cachées ou latentes, une stratégie classique consiste à travailler avec l'espérance du logarithme de la vraisemblance complète (observations et états latents) conditionnellement aux observations, qui est en général beaucoup plus accessible. C'est en particulier la stratégie de l'EM que nous discutons dans la section suivante.

5.5.3 Estimation par l'algorithme *Expectation-Maximisation*

L'algorithme EM (voir l'article fondateur de Dempster et al. [1977]) ne nécessite pas de calculer la vraisemblance observée, mais travaille avec la vraisemblance complète (ou complétée) des observations et des variables latentes, qui est en général plus simple à déterminer. Cet algorithme garantit l'augmentation de la vraisemblance à chaque itération, et converge vers un point stationnaire de la vraisemblance. Comme de nombreux algorithmes déterministes, la convergence n'est que locale, et des stratégies d'initialisation doivent être mises en place pour explorer l'espace des paramètres. L'algorithme EM est un algorithme de choix dans le cas de variables latentes. La log-vraisemblance complétée est au cœur de l'algorithme EM. Dans notre cas, son expression pour un trajet (k,d) se déduit de l'équation 5.12. En prenant le logarithme et en ne conservant que les termes dépendant des paramètres, l'expression utile de la log-vraisemblance complétée est

la fonction ℓ_c du paramètre $\boldsymbol{\theta}$ suivante :

$$\ell_c\Big(\boldsymbol{a}^{k,d}_{2:S}, \boldsymbol{w}^{k,d}_{1:(S-1)}; \boldsymbol{b}^{k,d}_{1:(S-1)}, \boldsymbol{\theta}\Big) = \sum_{s=2}^{S}\sum_{j=1}^{I} a^{k,d}_{s,i}\log(\alpha_{s,i}) + \left(\ell^{k,d}_{s-1,i} - a^{k,d}_{s,i}\right)\log(1-\alpha_{s,i}) + w^{k,d}_{s-1,j}\log\left(\pi^{k,d}_{s-1,j}\right)$$

où nous rappelons la définition de la charge à bord :

$$\ell^{k,d}_{s,i} = \sum_{g=1}^{s} w^{k,d}_{g,i} - a^{k,d}_{g,i}.$$

Les trajets étant indépendants et identiquement distribués, l'expression prenant en compte la totalité des observations devient :

$$\ell_c(\boldsymbol{\theta}) = \sum_{(k,d)\in\mathcal{N}}\sum_{s=2}^{S}\sum_{j=1}^{I} a^{k,d}_{s,i}\log(\alpha_{s,i}) + \left(\ell^{k,d}_{s-1,i} - a^{k,d}_{s,i}\right)\log(1-\alpha_{s,i}) + w^{k,d}_{s-1,j}\log\left(\pi^{k,d}_{s-1,j}\right)$$

Cette expression ne fait intervenir que des sommes, et est bien plus facile à appréhender que l'expression (5.13).

Résumé Dans cette section, nous avons formulé un modèle des déplacements cachés à l'échelle de la gare. Ce modèle repose sur la modélisation conjointe des descentes et des déplacements au fil du trajet. Nous avons posé une loi d'émission sur le nombre de descentes le liant à la charge à bord en entrée de gare. Nous avons également approché la loi des déplacements à chaque gare en suivant la logique de la modélisation à l'échelle du trajet. Nous avons pu exprimer la vraisemblance du modèle en se ramenant, grâce à un ensemble d'hypothèses, aux deux lois des descentes et des déplacements. Nous avons ébauché une stratégie d'estimation des paramètres du modèle à partir de l'optimisation de la vraisemblance complétée. Le travail de l'immédiat après-thèse consistera à formaliser complètement un tel algorithme pour notre problème et à le résoudre sur des données réelles.

5.A Justification par simulation de l'Approximation 1 à l'échelle du trajet de la page 9

L'Approximation 1 de la page 9 consiste à modéliser la somme de variables aléatoires de lois multinomiales indépendantes mais non identiquement distribuées par une unique loi multinomiale de même espérance.

Pouvons-nous utiliser la loi de Poisson multinomiale ? Hong [2013] étudie la loi de la somme de variables aléatoires de Bernoulli indépendantes et non identiquement distribuées, appelée loi de Poisson binomiale. Il montre que la densité de la loi Poisson binomiale n'est pas calculable numériquement. Dès lors, il propose deux approximations, la première par une loi de Poisson, efficace pour n grand et np plutôt faible, et la seconde par une loi gaussienne efficace, pour n grand et np grand. Ces deux approximations ont été étendues aux cas multi-classes par Lin et al. [2022]. Ces approximations sont efficaces pour un n grand, or dans notre cas, le nombre de voyageurs qui monte dans une rame, qui correspond ici à n, peut être faible par gare et même à l'échelle du trajet, voir section 5.3.

Évaluation par la simulation de l'approximation proposée. Puisque aucune des deux solutions n'est convaincante dans notre cas, nous proposons une troisième voie qui est l'approximation 1. Nous justifions cette approximation par une expérience numérique. Pour construire cette expérience, nous introduisons un ensemble de concepts et de notations génériques. Soit un ensemble $\{M_1, \ldots, M_I\}$ de vecteurs indépendants, chacun de loi multinomiale : $\boldsymbol{M}_i \sim \mathcal{M}(n_i, q_{i,1}, \ldots, q_{i,p})$ avec des notations parlant d'elles-mêmes. Nous définissons $\boldsymbol{S}$ comme la somme des $\boldsymbol{M}_i$:

$$\boldsymbol{S} = \sum_{i=1}^{I} \boldsymbol{M}_i. \tag{5.14}$$

Puisque nous ne connaissons pas la loi de $\boldsymbol{S}$, nous proposons comme approximation la variable aléatoire $\tilde{\boldsymbol{S}}$ de loi $\mathcal{M}(n, w_1, \ldots, w_p)$ où $n = \sum_{i=1}^{I} n_i$ et la probabilité w_j est définie comme :

$$w_j = \sum_{i=1}^{I} \frac{n_i}{n} q_{i,j} \tag{5.15}$$

avec n_i et $q_{i,j}$ définis précédemment.

5.A.1 Schéma des expériences numériques

Dans nos expériences numériques pour simuler $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$, nous fixons la matrice

$$\boldsymbol{Q} = \begin{pmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,p} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ q_{I,1} & q_{I,2} & \cdots & q_{I,p} \end{pmatrix}$$

de transition, égale à $\boldsymbol{P}_{\text{MLS}}$ et $\boldsymbol{P}_{\text{MLE}}$ estimées en Section (5.4.1, 5.4.2) et nous fixons aussi les effectifs n_i par zone. Une première solution, trop coûteuse en temps et en espace mémoire, pour fixer ces effectifs serait d'étudier la qualité de l'approximation selon toutes les valeurs possibles du vecteur $(n_1, \ldots, n_I)$. Une deuxième possibilité plus simple est de simuler $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ en deux temps, le premier en tirant des vecteurs $(n_1, \ldots, n_I)$ selon une certaine loi, puis le second en simulant $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$. Une idée que nous avons vite rejetée consisterait à tirer les montées par zone n_i suivant une loi multinomiale donnée. Une rapide étude numérique de la matrice de covariance des montées infirme le fait que les montées suivent une loi multinomiale. En effet, la matrice de covariance empirique ne prend pas de valeurs négatives en dehors de la diagonale. Nous proposons donc, ne connaissant pas la loi des montées par zone, de tirer les n_i suivant leur loi empirique. Cela revient à tirer avec remise N couples (k, d) parmi les 3 500 observations. Cependant, cette stratégie n'est pas sans limites. En effet, avec un tel tirage les effectifs n varient pour chaque couple (k, d). Cette variation impacte directement les matrices de covariance estimées car elles ont plus de poids que les variations intrinsèques à la loi multinomiale. Ainsi, la matrice de covariance estimée n'est plus identifiable à celle d'une loi multinomiale. Pour remédier à ce problème, nous tirons les N couples formés par un numéro de train k et un jour d : (k, d), sous condition de taille des effectifs $b^{k,d}_{\bullet,\bullet}$. Ainsi, nous ne conservons que les couples (k, d) tels que $b^{k,d}_{\bullet,\bullet}$ appartienne à un intervalle d'effectifs admissible. Cette méthode peut faire penser à une stratégie de simulation par rejet même si elle est beaucoup moins coûteuse en espace mémoire, car elle repose sur un filtre des observations en amont de la simulation. Nous définissons deux intervalles d'effectifs admissibles, un faible tel que $b^{k,d}_{\bullet,\bullet}$ appartient à l'intervalle $[50, 80]$ et un moyen tel que $b^{k,d}_{\bullet,\bullet}$ appartient à l'intervalle $[200, 210]$. Ces plages de valeurs représentent respectivement 3% et 5% de l'ensemble des observations. Pour conclure, l'expérience numérique justifiant l'approximation $\tilde{\boldsymbol{S}}$ repose sur quatre scénarios complémentaires résumés dans la Table 5.6 : deux matrices de transition et deux intervalles d'effectif différents. Un scénario est caractérisé par un intervalle admissible, noté n, et une matrice de transition, noté $\boldsymbol{Q}$. Nous verrons que l'approximation dévie légèrement de la loi de $\boldsymbol{S}$ pour des effectifs faibles, et que la différence entre les scénarios avec l'une ou l'autre des matrices de passage ($\boldsymbol{P}_{\text{MLS}}$, $\boldsymbol{P}_{\text{MLE}}$) est nulle.

Schéma de simulation. Nous simulons $N = 10\,000$ observations de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ avec un même schéma de simulation à deux étapes : une première pour fixer les effectifs et

TABLE 5.6 – Ensemble des intervalles admissibles et des matrices de passage utilisés pour la simulation de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$

n	effectifs faibles : $[50, 80]$ ou effectifs moyens : $[200, 210]$
$\boldsymbol{Q}$	$\boldsymbol{P}_{\text{MLS}}$ ou $\boldsymbol{P}_{\text{MLE}}$

une seconde pour simuler les deux lois. Ce schéma de simulation impose de fixer au préalable un scénario. Ainsi, pour illustration, nous fixons un intervalle admissible d'effectifs faibles $[50, 80]$ et une matrice de passage $\boldsymbol{P}_{\text{MLS}}$. Pour chaque itération de 1 à N, nous calculons les deux étapes :

1. Fixer les effectifs :
   (a) Tirer un couple (k, d) d'effectif $b^{k,d}_{\bullet,\bullet}$ appartenant à l'intervalle admissible des effectifs faibles, c'est-à-dire $b^{k,d}_{\bullet,\bullet} \in [50, 80]$;
   (b) Allouer le vecteur des montées de l'observation (k, d) aux n_i, c'est-à-dire $(n_1, \ldots, n_I) = (b^{k,d}_{\bullet,1}, \ldots, b^{k,d}_{\bullet,I})$.
2. Simuler les deux lois $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$:
   (a) Pour $\boldsymbol{S}$, faire en deux temps. (i) Tirer I réalisations indépendantes des $\boldsymbol{M}_i$ suivant $\mathcal{M}(n_i, q_{i,1}, \ldots, q_{i,p})$. (ii) Calculer la somme $\boldsymbol{S}$ avec les I réalisations $\{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_I\}$.
   (b) Pour $\tilde{\boldsymbol{S}}$, tirer $\tilde{\boldsymbol{S}}$ suivant $\mathcal{M}(n, w_1, \ldots, w_p)$ où w_i est calculé grâce à l'équation (5.15)

5.A.2 Résultats

Pour justifier l'Approximation 1 de $\tilde{\boldsymbol{S}}$ par $\boldsymbol{S}$, nous testons quatre scénarios avec un même schéma de simulation. Nous présentons dans la Section 5.A.2 les résultats en comparant $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ avec leurs densités, fonctions de répartition marginales et matrices de covariance, pour un premier scénario d'effectifs faibles $[50, 80]$ et de matrice de passage $\boldsymbol{P}_{\text{MLS}}$. Puis dans la Section 5.A.2, nous comparons $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ pour l'ensemble des scénarios de la Table 5.6.

Comparaison de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ pour un scénario

Nous comparons, pour des effectifs faibles $[50, 80]$ et la matrice de passage $\boldsymbol{P}_{\text{MLS}}$, l'approximation multinomiale $\tilde{\boldsymbol{S}}$ de la somme, notée $\boldsymbol{S}$, de lois multinomiales indépendantes et non identiquement distribuées. Nous comparons les densités et les fonctions de répartition marginales empiriques de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$. Sur la Figure 5.7, nous remarquons que les densités marginales sont très similaires excepté pour quelques valeurs autour de la moyenne. Nous remarquons également que les fonctions de répartition de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ sont quasiment indistinguables sur la Figure 5.8. Par ailleurs,

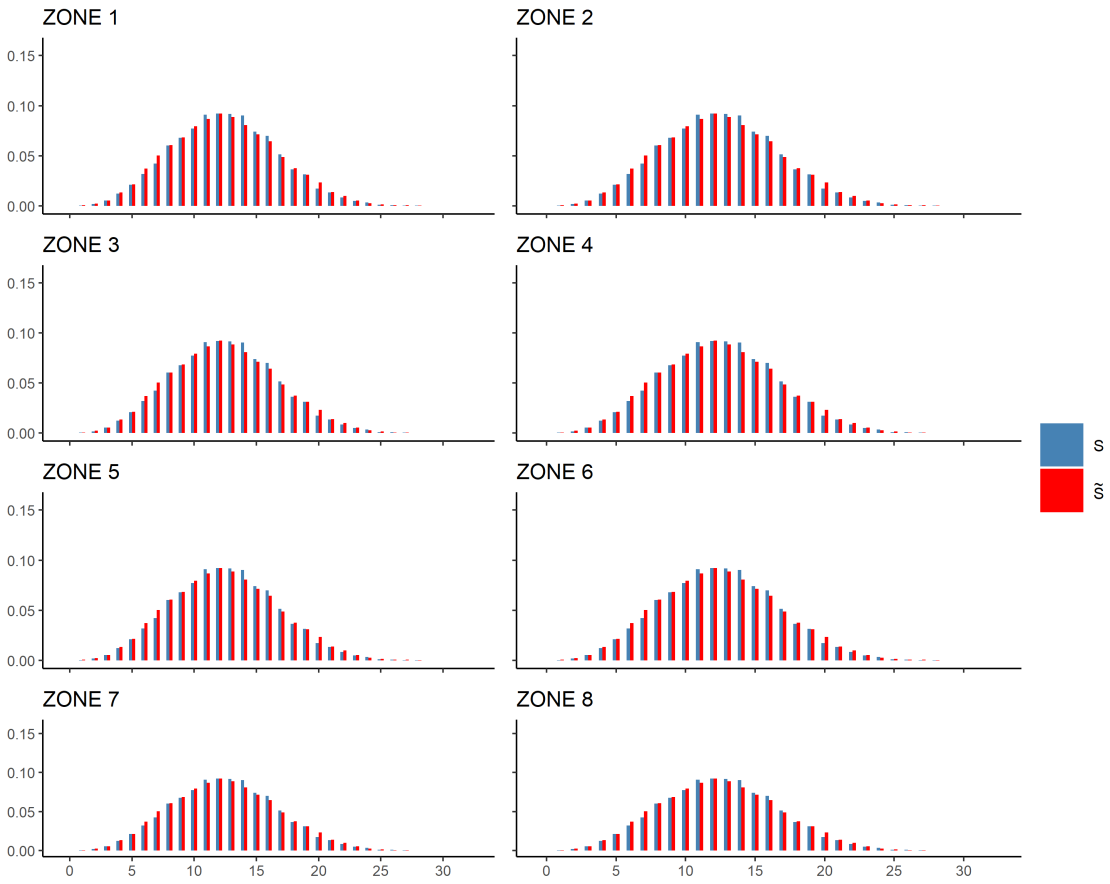


FIGURE 5.7 – Diagramme en bâton des densités marginales empiriques de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ de la zone 1 en haut à gauche à la zone 8 en bas à droite. Pour chaque graphique, en abscisse, le même support pour toutes les zones, en ordonnée, les proportions estimées

il apparait évident que la comparaison par les fonctions de répartition est très utile car elle permet de comparer les distributions marginales de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ de façon très compacte. La représentation par les fonctions de répartition empirique va nous permettre de généraliser cette comparaison qualitative à tous les scénarios dans la section suivante. Au delà des comportements des lois marginales, nous nous intéressons aussi aux relations entre zones. Pour ce faire, nous étudions les matrices de covariance empiriques calculées à partir des données simulées sur la Figure 5.9. Nous confirmons que les matrices de covariance sont celles de lois multinomiales (variances positives sur la diagonale et covariances négatives en dehors). Par ailleurs pour ce scénario les matrices de covariance de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ sont très proches.

Comparaison de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ pour tous les scénarios

Les fonctions de répartition empiriques sont indistinguables sur la Figure 5.10 ce qui confirme la proximité forte des matrices de passage $\boldsymbol{P}_{\text{MLS}}$ et $\boldsymbol{P}_{\text{MLE}}$ déjà constatée sur la Figure 5.5. De même, nous n'arrivons pas, pour un scénario donné, à distinguer les fonctions de répartition de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$, ce qui implique que leurs lois simulées sont très proches. Sur la Figure 5.11, nous constatons que quelle que soit la zone considérée, l'approximation de $\boldsymbol{S}$ par $\tilde{\boldsymbol{S}}$ est bonne au sens des fonctions de répartition marginales. Sur cette même figure, nous constatons que la taille d'effectif ne vient pas dégrader l'approximation proposée. En effet, que les effectifs soient faibles : par exemple en haut à gauche, en zone 1, ou forts : par exemple en bas à droite, en zone 8, les fonctions de répartition de $\boldsymbol{S}$ par $\tilde{\boldsymbol{S}}$ sont indistinguables. Les matrices de covariance de la Figure 5.12 indiquent que pour un même intervalle d'effectifs admissible (la première ligne par rapport à la seconde ligne) les matrices de covariance sont très proches. Par ailleurs, les matrices de gauche de chaque paire sont très proches de celles de droite, c'est-à-dire que quel que soit le scénario, les variations des vecteurs $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ sont cohérentes.

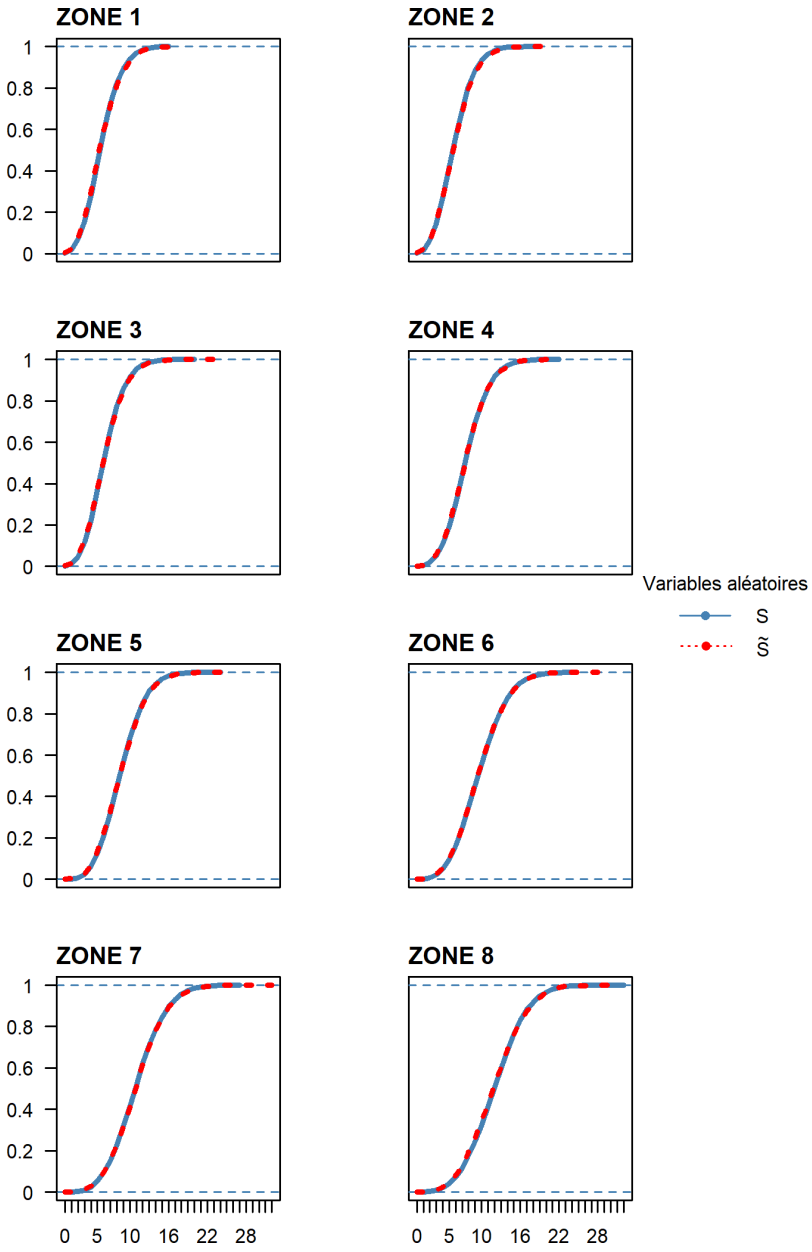


FIGURE 5.8 – Fonctions de répartition marginales empiriques de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ de la zone 1 en haut à gauche à la zone 8 en bas à droite. Pour chaque graphique, en abscisse, le même support pour toutes les zones, en ordonnée les proportions estimées

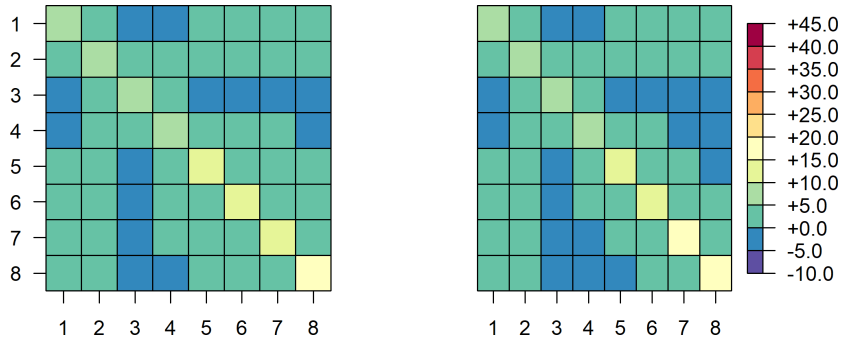


FIGURE 5.9 – Matrices de covariance de $\boldsymbol{S}$ (à gauche) et de $\tilde{\boldsymbol{S}}$ (à droite) pour des effectifs faibles $[50, 80]$ et la matrice de passage $\boldsymbol{P}_{\text{MLS}}$

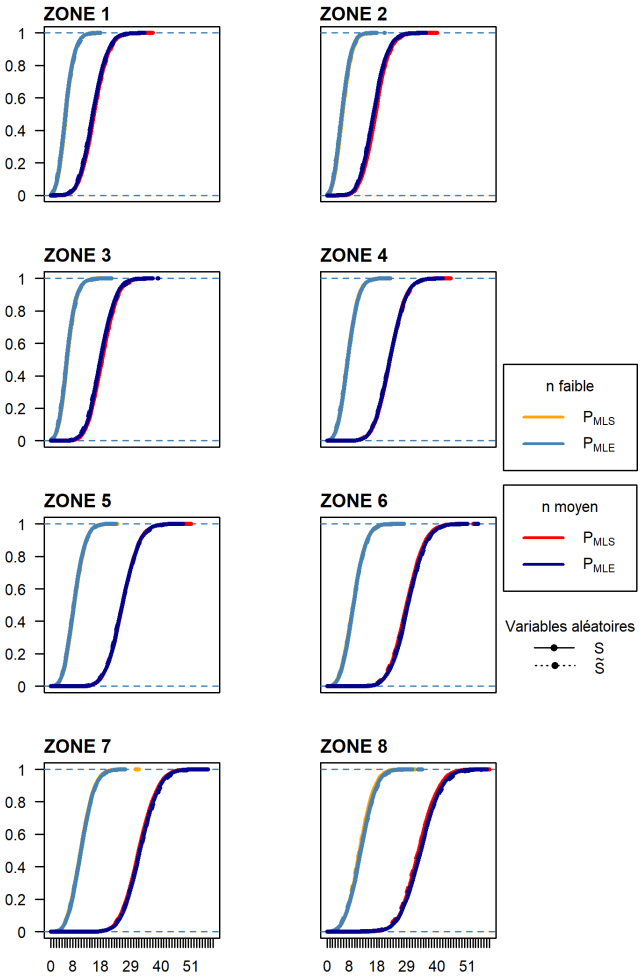


FIGURE 5.10 – Fonctions de répartition marginales empiriques de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ pour les quatre scénarios pour chacune des huit zones. Pour chaque graphique par zone, en abscisse, le même support pour toutes les zones, en ordonnée, les proportions estimées

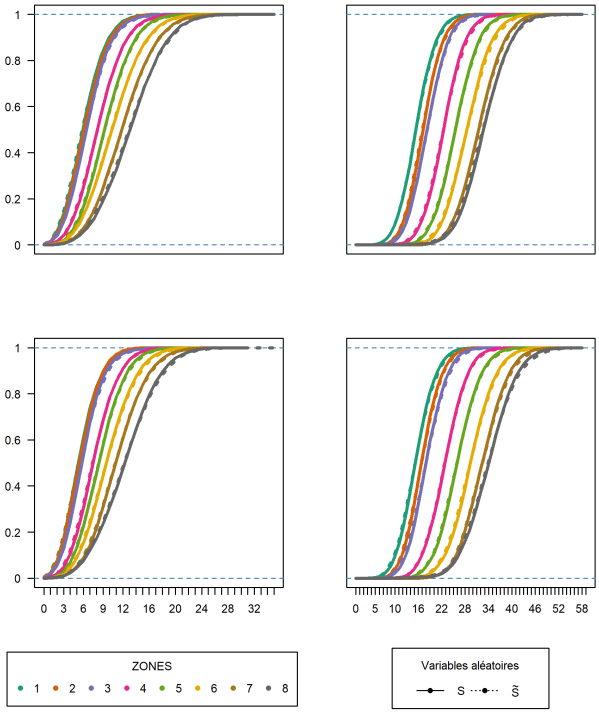


FIGURE 5.11 – Fonctions de répartition marginales empiriques de $\boldsymbol{S}$ et $\tilde{\boldsymbol{S}}$ pour les huit zones pour chacun des quatre scénarios. En haut : $\boldsymbol{P}_{\text{MLS}}$ et en bas $\boldsymbol{P}_{\text{MLE}}$. À gauche les faibles effectifs $[50, 80]$ et à droite les effectifs moyens $[200, 210]$. Pour chaque graphique, en abscisse, le même support pour toutes les zones, en ordonnée, les proportions estimées

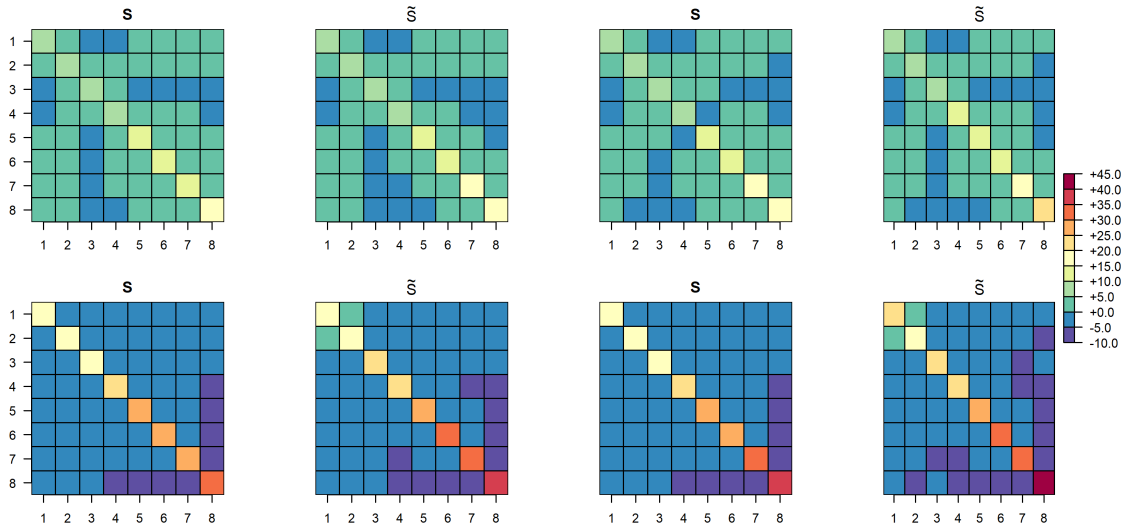


FIGURE 5.12 – Quatre paires de matrices de covariance $\boldsymbol{S}$ (à gauche) et de $\tilde{\boldsymbol{S}}$ (à droite) pour les scénarios suivants : la première ligne est dédiée aux faibles effectifs i.e. $[50, 80]$ avec $\boldsymbol{P}_{\text{MLS}}$ (la paire de gauche) et $\boldsymbol{P}_{\text{MLE}}$ (la paire de droite), la seconde ligne est dédiée aux effectifs moyens i.e. $[200, 210]$ avec $\boldsymbol{P}_{\text{MLS}}$ (la paire de gauche) et $\boldsymbol{P}_{\text{MLS}}$ (la paire de droite)

5.A.3 Conclusion

Nous avons illustré graphiquement que l'approximation $\tilde{\boldsymbol{S}}$ de $\boldsymbol{S}$ est raisonnable. Pour notre problème, nous pouvons utiliser l'approximation proposée sans crainte. Pour généraliser ce résultat, il faudrait tester des scénarios plus extrêmes. Par exemple, nous n'avons pas testé de valeurs d'effectifs très faibles autour de 10 ou 20 ce qui n'arrive quasiment jamais à l'échelle du trajet mais régulièrement à l'échelle d'un arrêt. Par ailleurs, nous avons utilisé des matrices de passage lisses car issues de l'optimisation des paramètres de notre problème concret. Nous n'avons pas testé de déplacements extrêmes, par exemple une seule colonne de 1 ce qui revient à tous les voyageurs se déplacent vers une même zone. Ces matrices de passage un peu particulières seraient probablement susceptibles de mettre en défaut l'approximation proposée. Cependant, elles n'ont aucune pertinence dans le cadre des déplacements des voyageurs dans les trains. Pour aller plus loin, une étude théorique de l'approximation pourrait être particulièrement utile surtout si elle peut permettre de tenir compte des liaisons possibles entre zones au-delà des matrices de covariance.

Bibliographie

Alfieri, Arianna, Groot, Rutger, Kroon, Leo, et Schrijver, Alexander. Efficient circulation of railway rolling stock. *Transportation Science*, 40(3) :378–391, 2006.

Altazin, Estelle, Dauzère-Pérès, Stéphane, Ramond, François, et Tréfond, Sabine. A multi-objective optimization-simulation approach for real time rescheduling in dense railway systems. *European Journal of Operational Research*, 286(2) :662–672, 2020.

Amita, Johar, Singh, Jain Sukhvir, et Kumar, Garg Pradeep. Prediction of bus travel time using artificial neural network. *International Journal for Traffic and Transport Engineering*, 5(4) :410–424, 2015. doi: 10.7708/ijtte.2015.5(4).06.

Anwar, Afian, Odoni, Amedeo, et Toh, Nelson. Busviz : Big data for bus fleets. *Transportation Research Record*, 2544(1) :102–109, 2016.

Assis, Wanderson O. et Milani, Basilio E.A. Generation of optimal schedules for metro lines using model predictive control. *Automatica*, 40(8) :1397–1404, 2004.

Bapaume, Thomas, Côme, Etienne, Roos, Jérémy, Ameli, Mostafa, et Oukhellou, Latifa. Image inpainting and deep learning to forecast short-term train loads. *IEEE Access*, 9 :98506–98522, 2021.

Baro, Johanna et Khouadjia, Mostepha. Passenger flow forecasting on transportation network : sensitivity analysis of the spatiotemporal features. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 734–741. IEEE, 2021.

Barry, Mike et Cardl, Brian. Visualizing MBTA data an interactive exploration of Boston's subway system. http://mbtaviz.github.io/, 2014. Accessed : 2022-08-22.

Berger, Annabell, Gebhardt, Andreas, Müller-Hannemann, Matthias, et Ostrowski, Martin. Stochastic delay prediction in large train networks. In *11th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2011.

Breiman, Leo. Random forests. *Machine Learning*, 45(1) :5–32, 2001. doi: 10.1023/A:1010933404324.

Brethomé, Lucile. *Modélisation et optimisation d'un plan de transport ferroviaire en zone dense du point de vue des voyageurs*. PhD thesis, École Centrale de Lille, 2018.

Briand, Anne-Sarah, Come, Etienne, Coulombel, Nicolas, El Mahrsi, Mohamed Khalil, Munch, Emmanuel, Richer, Cyprien, et Oukhellou, Latifa. Projet MOBILLETIC. Données billettiques et analyse des mobilités urbaines : le cas rennais, 2017.

Buchmüller, Stefan, Weidmann, Ulrich, et Nash, Andrew. Development of a dwell time calculation model for timetable planning. *WIT Transactions on The Built Environment*, 103 :525–534, 2008. doi: 10.2495/CR080511.

Büker, Thorsten et Seybold, Bernhard. Stochastic modelling of delay propagation in large networks. *Journal of Rail Transport Planning & Management*, 2(1-2) : 34–50, 2012.

Burch, Michael, Staudt, Yves, Frommer, Sina, Uttenweiler, Janis, Grupp, Peter, Hähnle, Steffen, Scheytt, Josia, et Kloos, Uwe. Pasvis : enhancing public transport maps with interactive passenger data visualizations. In *Proceedings of the 13th International Symposium on Visual Information Communication and Interaction*, pages 1–8, 2020.

Cacchiani, Valentina, Huisman, Dennis, Kidd, Martin, Kroon, Leo, Toth, Paolo, Veelenturf, Lucas, et Wagenaar, Joris. An overview of recovery models and algorithms for real-time railway rescheduling. *Transportation Research Part B : Methodological*, 63 :15–37, 2014.

Caprara, Alberto, Fischetti, Matteo, Toth, Paolo, Vigo, Daniele, et Guida, Pier Luigi. Algorithms for railway crew management. *Mathematical programming*, 79(1) :125–141, 1997.

Caprara, Alberto, Kroon, Leo, Monaci, Michele, Peeters, Marc, et Toth, Paolo. Passenger railway optimization. *Handbooks in operations research and management science*, 14 :129–187, 2007.

Ceder, Avishai. *Public Transit Planning and Operation : Modeling, Practice and Behavior*. CRC press, 2016.

Chen, Tianqi et Guestrin, Carlos. XGBoost : A scalable tree boosting system. In *Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Chu, Wen-jun, Zhang, Xing-chen, Chen, Jun-hua, et Xu, Bin. An ELM-based approach for estimating train dwell time in urban rail traffic. *Mathematical Problems in Engineering*, 2015 :Article ID 473432, 2015. doi: 10.1155/2015/473432.

Corman, Francesco et Henken, Jonas. Estimating aggregate railway performance from realized empirical data : Literature review, a test case and a research roadmap. *Journal of Rail Transport Planning & Management*, 22 :100316, 2022.

Corman, Francesco et Kecman, Pavle. Stochastic prediction of train delays in real-time using Bayesian networks. *Transportation Research Part C : Emerging Technologies*, 95 :599–615, 2018.

Cornet, Sélim, Buisson, Christine, Ramond, François, Bouvarel, Paul, et Rodriguez, Joaquin. Methods for quantitative assessment of passenger flow influence on train dwell time in dense traffic areas. *Transportation Research Part C : Emerging Technologies*, 106 :345–359, 2019. doi: 10.1016/j.trc.2019.05.008.

Coulaud, Rémi et Grangé, Martine. Modélisation de l'impact des flux voyageurs sur les temps d'échange pour la simulation des marges d'exploitation : une application à la ligne N de transilien. In *4èmes Rencontres Francophones Transport Mobilité (RFTM)*, 2022.

Coulaud, Rémi et Vimont, Mathilde. How to use APC data to model passenger movement on-board ? An application to Paris suburban train network. In *8th International Symposium On Transport Network Reliability (INSTR)*, 2021.

Coulaud, Rémi, Keribin, Christine, et Stoltz, Gilles. Quels modèles pour le temps de stationnement des trains en île de france ? In *SFdS 2020-52èmes Journées de Statistiques de la Société Française de Statistiques*, 2020.

Coulaud, Rémi, Keribin, Christine, et Stoltz, Gilles. One-station-ahead forecasting of dwell time, arrival delay and passenger flows on trains equipped with automatic passenger counting (apc) device. In *13th World Congress on Rail Research (WCRR)*, 2022.

Coulaud, Rémi, Keribin, Christine, et Stoltz, Gilles. Modeling dwell time in a data-rich railway environment : with operations and passenger flows data. Re-soumis à *Transportation Research Part C* (TRC) après corrections. Preprint accessible ici hal.archives-ouvertes.fr/hal-03651835/, 2022.

Coulaud, Rémi, Mazon, Valentine, Sanchis, Laura, et Cats, Oded. Share of strategic alighting passengers combining automatic passenger counting and OpenStreeMap. In *Conference on Advanced Systems in Public Transport (CASPT)*, 2022.

Daamen, Winnie, Lee, Yu-chen, et Wiggenraad, Paul. Boarding and alighting experiments : Overview of setup and performance and some preliminary results. *Transportation Research Record*, 2042(1) :71–81, 2008. doi: 10.3141/2042-08.

Daamen, Winnie, Goverde, Rob M. P., et Hansen, Ingo A. Non-discriminatory automatic registration of knock-on train delays. *Networks and Spatial Economics*, 9(1) :47–61, 2009.

D'Acierno, Luca, Botte, Marilisa, Placido, Antonio, Caropreso, Chiara, et Montella, Bruno. Methodology for determining dwell times consistent with passenger flows in the case of metro services. *Urban Rail Transit*, 3(2) :73–89, 2017. doi: 10.1007/s40864-017-0062-4.

Darsena, Donatella, Gelli, Giacinto, Iudice, Ivan, et Verde, Francesco. Enabling and emerging sensing technologies for crowd management in public transportation systems : A review, 2020. Available on the open-archive ARXIV at https://arxiv.org/abs/2009.12619.

Deau, Dominique. Nexteo. *Revue générale des Chemins de Fer*, 2015.

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

Ding, Chuan, Wang, Donggen, Ma, Xiaolei, et Li, Haiying. Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. *Sustainability*, 8(11) :1100, 2016. doi: 10.3390/su8111100.

Dueker, Kenneth J., Kimpel, Thomas J., Strathman, James G., et Callas, Steve. Determinants of bus dwell time. *Journal of Public Transportation*, 7(1) :21–40, 2004. doi: 10.5038/2375-0901.7.1.2.

Evans, Gary W. et Wener, Richard E. Crowding and personal space invasion on the train : Please don't make me sit in the middle. *Journal of Environmental Psychology*, 27(1) :90–94, 2007.

Farandou, Jean-Pierre. Le fer contre le carbone, 2022. Fondation Jean Jaurès.

Friedman, Jerome H. Greedy function approximation : a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001. doi: 10.1214/aos/1013203451.

Ghofrani, Faeze, He, Qing, Goverde, Rob M. P., et Liu, Xiang. Recent applications of big data analytics in railway transportation systems : A survey. *Transportation Research Part C : Emerging Technologies*, 90 :226–246, 2018.

Goodfellow, Ian, Bengio, Yoshua, et Courville, Aaron. *Deep Learning*. MIT Press, Cambridge, 2016.

Goverde, Rob M. P. *Punctuality of railway operations and timetable stability analysis*. PhD thesis, Delft University of Technology, 2005.

Goverde, Rob M. P. et Meng, Lingyun. Advanced monitoring and management information of railway operations. *Journal of Rail Transport Planning & Management*, 1(2) :69–79, 2011.

Graffagnino, Thomas. Ensuring timetable stability with train traffic data. *Computers in Railways XIII : Computer System Design and Operation in the Railway and Other Transit Systems*, 127 :427, 2013.

Hänseler, Flurin S., van den Heuvel, Jeroen P.A., Cats, Oded, Daamen, Winnie, et Hoogendoorn, Serge. A passenger-pedestrian model to assess platform and train usage from automated data. *Transportation Research Part A : Policy and Practice*, 132 :948–968, 2020.

Hansen, Ingo A. et Pachl, Jörn. *Railway Timetabling and Operations : Analysis, Modelling, Optimisation, Simulation, Performance, Evaluation*. Eurail press, 2014.

Hansen, Ingo A., Goverde, Rob M.P., et van der Meer, Dirk J. Online train delay recognition and running time prediction. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 1783–1788, 2010. doi: 10.1109/ITSC.2010.5625081.

Harris, Nigel G. et Anderson, Richard J. An international comparison of urban rail boarding and alighting rates. *Proceedings of the Institution of Mechanical Engineers, Part F : Journal of Rail and Rapid Transit*, 221(4) :521–526, 2007. doi: 10.1243/09544097JRRT115.

Hastie, Trevor, Tibshirani, Robert, et Friedman, Jerome H. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer, New-York, 2nd edition, 2009.

Hong, Yili. On computing the distribution function for the poisson binomial distribution. *Computational Statistics & Data Analysis*, 59 :41–51, 2013.

IdFM, . Contrat entre Île-de-France Mobilités, SNCF Voyageurs et SNCF-Gares & Connexions 2020 – 2023, 2020.

Jarrossay, Anaïs. *Pour éviter de voyager serré, pensez HECTOR !*, 2021. https://maligneh.transilien.com/2021/03/09/pour-eviter-de-voyager-serre-pensez-hector [Last accessed on 2022-06-26].

Jenelius, Erik. Data-driven metro train crowding prediction based on real-time load data. *IEEE Transactions on Intelligent Transportation Systems*, 21(6) :2254–2265, 2019.

Kecman, Pavle. *Models for Predictive Railway Traffic Management*. PhD thesis, Delft University of Technology, 2014.

Kecman, Pavle et Goverde, Rob M.P. Predictive modelling of running and dwell times in railway traffic. *Public Transport*, 7(3) :295–319, 2015. doi: 10.1007/s12469-015-0106-7.

Kecman, Pavle, Corman, Francesco, et Meng, Lingyun. Train delay evolution as a stochastic process. In *6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015)*. IVT, ETH Zurich ; Orange Labs, 2015.

Kim, Hyunmi, Kwon, Sohee, Wu, Seung Kook, et Sohn, Keemin. Why do passengers choose a specific car of a metro train during the morning peak hours ? *Transportation Research Part A : Policy and Practice*, 61 :249–258, 2014.

Koutsopoulos, Haris N., Ma, Zhenliang, Noursalehi, Peyman, et Zhu, Yiwen. Transit data analytics for planning, monitoring, control, and information. In *Mobility patterns, big data and transport analytics*, pages 229–261. 2019.

Krstanoski, Nikola. Modelling passenger distribution on metro station platform. *International Journal for Traffic & Transport Engineering*, 4(4) :456–465, 2014.

Kuipers, Ruben A., Palmqvist, Carl-William, Olsson, Nils O.E., et Hiselius, Lena Winslott. The passenger's influence on dwell times at station platforms : a literature review. *Transport Reviews*, 41(6) :721–741, 2021. doi: 10.1080/01441647.2021.1887960.

Kuipers, Ruben A., Palmqvist, Carl-William, Olsson, Nils O.E., et Winslott Hiselius, Lena. The passenger's influence on dwell times at station platforms : a literature review. *Transport Reviews*, 41(6) :721–741, 2021. doi: 10.1080/01441647.2021.1887960.

Lam, William H.K., Cheung, C.Y., et Poon, Y.F. A study of train dwelling time at the Hong Kong mass transit railway system. *Journal of Advanced Transportation*, 32(3) :285–295, 1998.

Laurent, Sophie, Prédali, Frédérique, et Boichon, Nicolas. Comparaison de réseaux mass transit francilien et internationaux, avec un zoom sur l'accueil de grands événements, 2018. IAU Île-de-France.

Lessan, Javad, Fu, Liping, et Wen, Chao. A hybrid Bayesian network model for predicting delays in train operations. *Computers & Industrial Engineering*, 127 : 1214–1222, 2019.

Levinson, Herbert S. Analyzing transit travel time performance. *Transportation Research Record*, 915 :1–6, 1983.

Li, Dewei, Daamen, Winnie, et Goverde, Rob M.P. Estimation of train dwell time at short stops based on track occupation event data : a study at a Dutch railway station. *Journal of Advanced Transportation*, 50(5) :877–896, 2016. doi: 10.1002/atr.1380.

Li, Yaguang, Yu, Rose, Shahabi, Cyrus, et Liu, Yan. Diffusion convolutional recurrent neural network : Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJiHXGWAZ.

Lin, Tyh-ming et Wilson, Nigel H.M. Dwell time relationships for light rail systems. *Transportation Research Record*, 1361 :287–295, 1992.

Lin, Zhengzhi, Wang, Yueyao, et Hong, Yili. The poisson multinomial distribution and its applications in voting theory, ecological inference, and machine learning. *arXiv preprint arXiv :2201.04237*, 2022.

Litman, Todd. Transit price elasticities and cross-elasticities. *Journal of Public Transportation*, 7(2) :3, 2004.

Massoni, Michel, Raoul, Emmanuel, et Ayong Le Kama, Alain. Modélisation des déplacements de voyageurs en Île-de-France, 2015.

McNally, Michael G. The four-step model. In *Handbook of transport modelling*. Emerald Group Publishing Limited, 2007.

Medeossi, Giorgio et Nash, Andrew. Reducing delays on high-density railway lines : London–Shenfield case study. *Transportation Research Record*, 2674(7) :193–205, 2020. doi: 10.1177/0361198120921159.

Ngauw, Brian. *Is the passenger loading system beneficial ?*, 2018. https://blog.sgtrains.com/2018/06/passenger-loading-systems [Last accessed on 2022-06-17].

Nielsen, Bo Friis, Frølich, Laura, Nielsen, Otto Anker, et Filges, Dorte. Estimating passenger numbers in trains using existing weighing capabilities. *Transportmetrica A : Transport Science*, 10(6) :502–517, 2014.

Noursalehi, Peyman, Koutsopoulos, Haris N., et Zhao, Jinhua. Predictive decision support platform and its application in crowding prediction and passenger information generation. *Transportation Research Part C : Emerging Technologies*, 129 :103–139, 2021.

OMNIL, DRIEA. EGT H2020-Île-de-France Mobilités/Résultats partiels, 2018.

Palmqvist, Carl-William, Tomii, Norio, et Ochiai, Yasufumi. Explaining dwell time delays with passenger counts for some commuter trains in Stockholm and Tokyo. *Journal of Rail Transport Planning & Management*, 14 :100189, 2020. doi: 10.1016/j.jrtpm.2020.100189.

Parbo, Jens, Nielsen, Otto Anker, et Prato, Carlo Giacomo. Passenger perspectives in railway timetabling : a literature review. *Transport Reviews*, 36(4) :500–526, 2016.

Pasini, Kevin, Khouadjia, Mostepha, Same, Allou, Ganansia, Fabrice, et Oukhellou, Latifa. LSTM encoder-predictor for short-term train load forecasting. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 535–551. Springer, 2019. doi: 10.1007/978-3-030-46133-1_32.

Pedersen, Timothy, Nygreen, Thomas, et Lindfeldt, Anders. Analysis of temporal factors influencing minimum dwell time distributions. *WIT Transactions on the Built Environment*, 181 :447–458, 2018. doi: 10.2495/CR180401.

Peftitsi, Soumela, Jenelius, Erik, et Cats, Oded. Determinants of passengers' metro car choice revealed through automated data sources : a stockholm case study. *Transportmetrica A : Transport Science*, 16(3) :529–549, 2020.

Pelletier, Marie-Pier, Trépanier, Martin, et Morency, Catherine. Smart card data use in public transit : A literature review. *Transportation Research Part C : Emerging Technologies*, 19(4) :557–568, 2011.

Pinna, Ivano et Dalla Chiara, Bruno. Automatic passenger counting and vehicle load monitoring. *Ingegneria Ferroviaria*, 65(2) :101–138, 2010.

Pritchard, James, Sadler, Jason, Blainey, Simon, Waldock, Ian, et Austin, Jeremy. Predicting and mitigating small fluctuations in station dwell times. *Journal of Rail Transport Planning & Management*, 18 :100249, 2021. doi: 10.1016/j.jrtpm.2021.100249.

Puong, Andre. Dwell time model and analysis for the MBTA red line. Technical report, Massachusetts Institute of Technology Research Memo, 2000.

Robenek, Tomáš, Maknoon, Yousef, Azadeh, Shadi Sharif, Chen, Jianghang, et Bierlaire, Michel. Passenger centric train timetabling problem. *Transportation Research Part B : Methodological*, 89 :107–126, 2016.

Rogers, Stella. *Arriva Rail London trial real-time passenger information, every step of the journey*, 2019. https://www.globalrailwayreview.com/article/78467/arriva-rail-london-passenger-information, Last accessed on 2022-06-17.

Roos, Jérémy, Gavin, Gérald, et Bonnevay, Stéphane. A dynamic Bayesian network approach to forecast short-term urban rail passenger flows with incomplete data. *Transportation research procedia*, 26 :53–61, 2017.

Rößler, David, Reisch, Julian, Hauck, Florian, et Kliewer, Natalia. Discerning primary and secondary delays in railway networks using explainable ai. *Transportation Research Procedia*, 52 :171–178, 2021.

Schmitt, Angie. *These London Trains Have Real-Time Displays to Reduce Crowding*, 2017. https://usa.streetsblog.org/2017/08/03/these-london-trains-have-real-time-displays-to-reduce-crowding [Last accessed on 2022-06-17].

Schöttl, Jakob, Seitz, Michael J, et Köster, Gerta. Investigating the randomness of passengers' seating behavior in suburban trains. *Entropy*, 21(6) :600, 2019.

Seriani, Sebastian et Fujiyama, Taku. Modelling the distribution of passengers waiting to board the train at metro stations. *Journal of Rail Transport Planning & Management*, 11 :100141, 2019.

Shelat, Sanmay, Daamen, Winnie, Kaag, Bjorn, Duives, Dorine, et Hoogendoorn, Serge. A markov-chain activity-based model for pedestrians in office buildings. *Collective Dynamics*, 5 :423–430, 2020.

Szplett, David et Wirasinghe, S.C. An investigation of passenger interchange and train standing time at lrt stations :(i) alighting, boarding and platform distribution of passengers. *Journal of advanced transportation*, 18(1) :1–12, 1984.

Thales, . *Reflecting passengers' top public transport experience priorities, Thales to provide real-time passenger density insights to public transport operators*, 2021. https://www.thalesgroup.com/en/group/journalist/press-release/reflecting-passengers-top-public-transport-experience-priorities [Last accessed on 2022-06-17].

Tirachini, Alejandro, Hensher, David A., et Rose, John M. Crowding in public transport systems : effects on users, operation and implications for the estimation of demand. *Transportation research part A : Policy and practice*, 53 :36–52, 2013.

Toqué, Florian. *Prévision et visualisation de l'affluence dans les transports en commun à l'aide de méthodes d'apprentissage automatique*. PhD thesis, IFSTTAR/GRETTIA - Génie des Réseaux de Transport Terrestres et Informatique Avancée, 2019.

Trépanier, Martin, Tranchant, Nicolas, et Chapleau, Robert. Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1) :1–14, 2007.

Trinkoff, Alison M. Seating patterns on the washington, dc metro rail system. *American journal of public health*, 75(6) :657–658, 1985.

UITP, . World report on metro automation : Statistics brief, 2019.

Ulak, Mehmet Baran, Yazici, Anil, et Zhang, Yun. Analyzing network-wide patterns of rail transit delays using bayesian network learning. *Transportation Research Part C : Emerging Technologies*, 119 :102749, 2020.

Wang, Chuang, Yan, Da, et Jiang, Yi. A novel approach for building occupancy simulation. In *Building simulation*, volume 4, pages 149–167, 2011.

Wang, Yihui, Tang, Tao, Ning, Bin, Van Den Boom, Ton JJ, et De Schutter, Bart. Passenger-demands-oriented train scheduling for an urban rail transit network. *Transportation Research Part C : Emerging Technologies*, 60 :1–23, 2015.

Wang, Zhenzhen, He, Sylvia Y, et Leung, Yee. Applying mobile phone data to travel behaviour research : A literature review. *Travel Behaviour and Society*, 11 : 141–155, 2018.

Wiggenraad, Paul B. L. Alighting and boarding times of passengers at Dutch railway stations. In *TRAIL Research School*, 2001.

Wirasinghe, S. Chan et Szplett, David. An investigation of passenger interchange and train standing time at LRT stations : (ii) estimation of standing time. *Journal of Advanced Transportation*, 18(1) :13–24, 1984. doi: 10.1002/atr.5670180103.

Wood, Simon. *Generalized Additive Models : An Introduction with R*. CRC Press, 2006.

Wright, Marvin N. et Ziegler, Andreas. ranger : A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1) :1–17, 2017. doi: 10.18637/jss.v077.i01.

Yaghini, Masoud, Khoshraftar, Mohammad M., et Seyedabadi, Masoud. Railway passenger train delay prediction via neural network model. *Journal of Advanced Transportation*, 47(3) :355–368, 2013. doi: 10.1002/atr.193.

Zeng, Wei, Fu, Chi-Wing, Arisona, Stefan Müller, Erath, Alexander, et Qu, Huamin. Visualizing mobility of public transportation system. *IEEE transactions on visualization and computer graphics*, 20(12) :1833–1842, 2014.

Zhang, Cen et Teng, Jing. Bus dwell time estimation and prediction : a study case in Shanghai-China. *Procedia-Social and Behavioral Sciences*, 96 :1329–1340, 2013.

Zhang, Yanru et Haghani, Ali. A gradient boosting method to improve travel time prediction. *Transportation Research Part C : Emerging Technologies*, 58 :308–324, 2015. doi: 10.1016/j.trc.2015.02.019.

Zhang, Yizhou, Jenelius, Erik, et Kottenhoff, Karl. Impact of real-time crowding information : a stockholm metro pilot study. *Public Transport*, 9(3) :483–499, 2017.

Zhong, Chen, Batty, Michael, Manley, Ed, Wang, Jiaqiu, Wang, Zijia, Chen, Feng, et Schmitt, Gerhard. Variability in regularity : Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data. *PloS one*, 11(2) :e0149222, 2016.

A

One-Station-Ahead Forecasting of Dwell Time, Arrival Delay and Passenger Flows on Trains Equipped with Automatic Passenger Counting (APC) Device

Ce travail est une version préliminaire du Chapitre 4. Il a été soumis et présenté à la conférence : World Congress on Railway Research (WCRR) en juin 2022.

Coulaud, Rémi, Keribin, Christine, et Stoltz, Gilles. One-station-ahead forecasting of dwell time, arrival delay and passenger flows on trains equipped with automatic passenger counting (apc) device. In *13th World Congress on Rail Research (WCRR)*, 2022.

One-Station-Ahead Forecasting of Dwell Time, Arrival Delay and Passenger Flows on Trains Equipped with Automatic Passenger Counting (APC) Device

Rémi COULAUD[1,2], Christine KERIBIN[2], Gilles STOLTZ[2]

[1]SNCF Voyageurs – Transilien, 10 rue Camille Moke, 93220, Saint-Denis, France
[2]Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France
Corresponding Author: Rémi Coulaud (remi.coulaud@sncf.fr)

Abstract

We consider a suburban railway network line in the greater Paris area, with sub-branches. Trains of the line are equipped both with automatic vehicle localization (AVL) and automatic passenger counting (APC) devices, leading to a rich data set with simultaneous measurements of variables related to railway operations (arrival delay, dwell time) and passenger flows (numbers of passengers alighting and boarding, total load at departure). We aim for one-station-ahead forecasting of each of these five variables independently from each other. To do so, we build a bi-auto-regressive approach consisting of using the past values of the variable of interest along a first dimension, given by past stations along the train ride, and along a second dimension, given by past trains at the station. A building block of this approach is a train-station representation that accommodates different types of train services. We identify repeated patterns in this representation and exploit this fact. Indeed, the proposed bi-auto-regressive models are based on linear regressions whose coefficients depend on the stations and possibly only on the location of the train ride within a repeated pattern. This results in models that have a smaller complexity than extremely local models tailored to the timetables, with no significant decrease in accuracy.

Keywords: dwell time, arrival delay, passenger flow, forecasting, bi-auto-regressive models

1. Introduction

We consider a suburban railway network line, namely, line H of the SNCF railway network of the greater Paris area, in the direction from suburbs (from the origin station "Pontoise" and from the intermediate station "Montsoult-Mafflier") to Paris (terminus station "Gare du Nord"), see Figure 1.

We are interested in the short-term (next station or next train) forecasting of some variables related to train stops in a given station: numbers A and B of passengers alighting and boarding, load L at departure, as well as dwell time T and arrival delay ΔA. We use X to refer to any of these quantities in a generic way. At a high-level, our approach consists of predicting the value $X_{k,s}$ of a quantity for the k-th train in the s-th station based on recent observations of the same quantity at the same station for earlier trains and at earlier stations for the same train, i.e., we consider some auto-regressive modelling, where "time" is measured through pairs (k, s) of trains and stations.

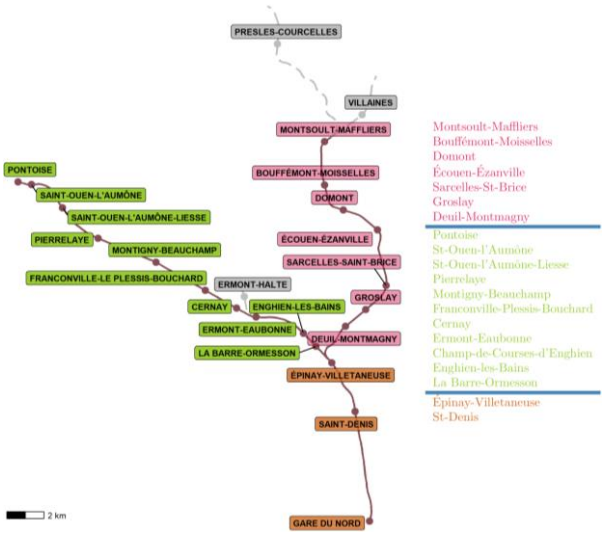


Figure 1*:* Left picture: The subset of the railway line considered, consisting of two branches (in green and in pink) merging for the final three stations (in brown). Stations not considered are in grey. Right column: List of the corresponding stations except for the terminus station, with the same colour code.

We do this in a novel way: we introduce our methodology in Sections 2.1 and 2.2 and then compare it to existing ones in Section 2.3.

2. Methodology

Our simplified representation of line H is composed of two branches that merge for the final three stations. As there is no scheduled train overtake, we may reindex the scheduled timetable from time in x-axis (left part of Figure 2) to train number in x-axis (right part). Trains are ranked according to their stops at the second station of the common part of line H. We discard the terminus station in the timetable and in our predictions, as the forecasting of most quantities of interest (dwell time, number of passengers alighting and boarding, load after departure) is of no interest or not applicable.

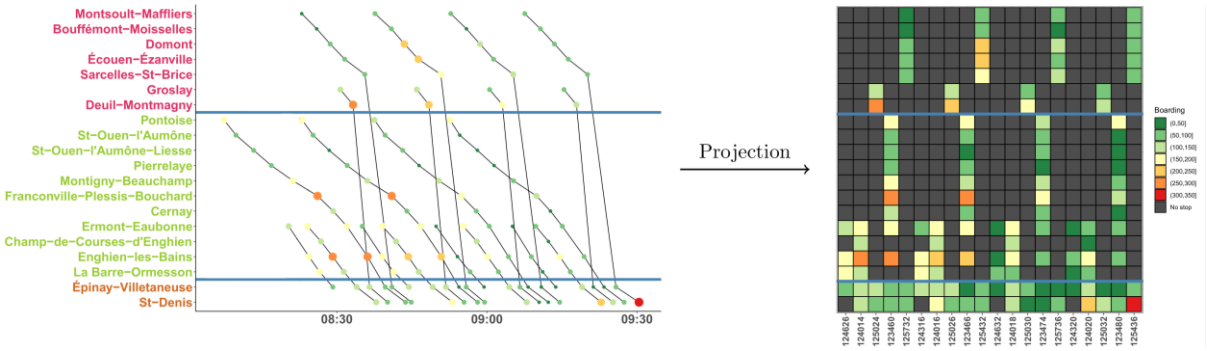


Figure 2: Projection of time-station timetable (left graph) into a train-station timetable (right graph). Colours and sizes of circles indicate the level of the boarding. Trains do not stop at stations marked in dark grey on the right graph.

2.1. L-shaped Neighbourhoods for Prediction

To forecast some information (for example, for train 29 at station Épinay-Villetaneuse: purple cell on Figure 3), we may use past information at the station of interest (for earlier trains: pink cells) and along the past stations of the given train ride (blue cells). Of course, we may restrict our attention to a shorter memory range ($P = 2$ past trains in dark pink and $Q = 3$ earlier stations in dark blue). Strikethrough cells contain future information and cannot be used for prediction. All in all, the information to be used for prediction is shaped like an inverse-L: we will refer to it as an L-shaped neighbourhood with sizes P and Q.

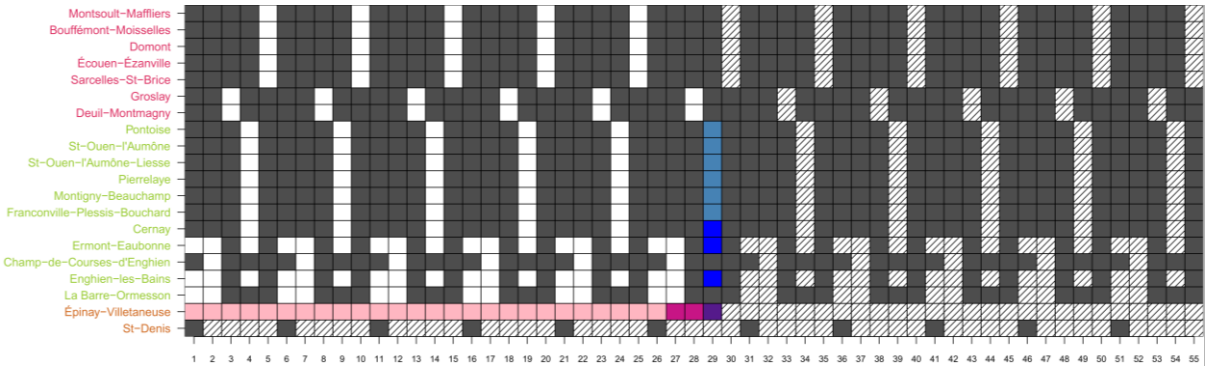


Figure 3: Identification of the underlying structure in the train-station graph of the morning peak hour. Dark grey cells mean no stop for corresponding train-station pairs. Information available at earlier stations and for earlier trains are in blue and pink, respectively. Dark blue and dark pink denote the most recent information. Strikethrough cells contain future information and cannot be used for prediction.

In the right graph of Figure 2, we read an underlying structure that consists of 4 repetitions of a given pattern. structure, see Figure 3. We denote by M the periodicity of the repetitions: here, the same sub-structures arise every $M = 5$ trains. On top of these intra-day repetitions, we also consider inter-days repetitions, i.e., each day

may be seen as a realization of a given stochastic process.

For the sake of clarity, Figure 4 depicts the repetitions of the L-shaped neighbourhood introduced in Figure 3, during the morning peak hour of a given day.

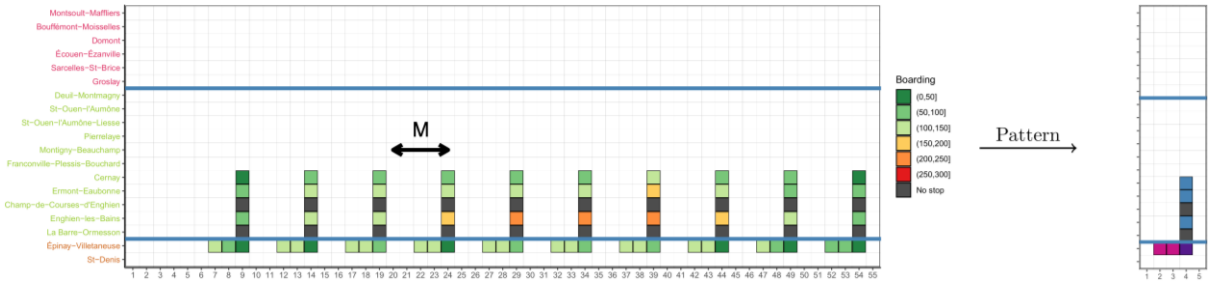
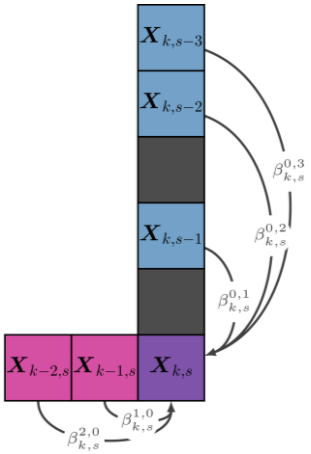


Figure 4: Illustration of the repetitions of a given L-shaped neighbourhood in the train-station graph of the morning peak hour.

2.2. L-shaped Regression Models, with Different Degrees of Stationarity

We model a quantity $X_{k,s}$ of interest (dwell time T, number B of passengers boarding , etc.) at a station s for train k as an affine function of the same quantities of interest $X_{k-p,s}$ and $X_{k,s-q}$ in the L-shaped neighbourhood considered – hence the name of L-shaped regression models. The three models introduced differ by the flexibility allowed for the coefficients. In the "stationary" model, all instances of the L-shaped neighbourhoods within the peak hours and among days are considered repetitions of the same stochastic process, while in the "non-stationary" model, days are considered repetitions of the same stochastic process, but no specific structure is assumed within the peak hour. The respective linear regression models have coefficients that only depend on the station s and the location of k within a pattern (which is given by the value of k modulo M, denoted by $k[M]$), and, on the contrary, that may fully depend on k and s. A model lying between these two extremes is referred to as "semi-stationary", where the intercept coefficient may fully depend on k and s (which models some variation of the level within the morning peak hour) but the coefficients for explanatory variables only depend on $k[M]$ and s (which models some intrinsic relationship with neighbouring values). Formally, the modelling equations read as follows, where $\varepsilon_{k,s}$ denote the error terms:



Non-stationary: $$X_{k,s} = \beta^{0,0}_{k,s} + \sum_{p=1}^{P} \beta^{p,0}_{k,s} X_{k-p,s} + \sum_{q=1}^{Q} \beta^{0,q}_{k,s} X_{k,s-q} + \varepsilon_{k,s}$$

Semi-stationary: $$X_{k,s} = \beta^{0,0}_{k,s} + \sum_{p=1}^{P} \beta^{p,0}_{k[M],s} X_{k-p,s} + \sum_{q=1}^{Q} \beta^{0,q}_{k[M],s} X_{k,s-q} + \varepsilon_{k,s}$$

Stationary: $$X_{k,s} = \beta^{0,0}_{k[M],s} + \sum_{p=1}^{P} \beta^{p,0}_{k[M],s} X_{k-p,s} + \sum_{q=1}^{Q} \beta^{0,q}_{k[M],s} X_{k,s-q} + \varepsilon_{k,s}$$

2.3. Literature Review and Main Contributions

Main contribution #1: Assessing forecasting methods on 5 variables. In the public transportation literature, short-term forecasting methods are typically built for one specific variable at a time: e.g., dwell time in Kecman

& Goverde [4], Li et al. [5], arrival delay in Corman & Kecman [2], passenger load at departure in Bapaume et al. [1], Jenelius [3], Pasini et al. [6]. Although the underlying methods are often generic, they were each tested only for a specific variable. The richness of our data set allows for the assessment of each model considered on 5 different variables. We chose to evaluate performance one-single-step ahead as Li et al. [5] did but keep in mind that this is merely a first step towards operational solutions, which require rather multiple-step-ahead forecasts as noted by Bapaume et al. [1].

Main contribution #2: Train-station representation despite sub-branches. Conversions of time-station representations into train-station ones were already performed in Bapaume et al. [1], Jenelius [3] on simpler networks with a unique branch and simpler train rides, with no overtake. We extend such conversions to a case where there are sub-branches and several train services types. We note however that the more complex approach by Corman & Kecman [2] allows for such multiple sub-branches, but leads to models with significantly more parameters (see below), while we aim for frugal models.

Main contribution #3: Bi-auto-regressive modelling. To the best of our knowledge, a systematic exploration of the power of auto-regressive modelling using both the recent past along the train ride and the recent past at the station considered was not offered by the literature so far. Instead, Jenelius [3] built auto-regressive models along the train ride using levels of the variable of interest at the station considered, formed by averages over some possibly short- and long-term past.

Main contribution #4: A balance between frugality and complexity. Some models of the literature are possibly too frugal and hold independently of the stations and train rides considered (i.e., do not depend on s and k, with our notation), as in Li et al. [5]. Some other approaches rely on possibly too many parameters, as the one in Corman & Kecman [2] which proposes a modelling even more complex than what we termed the "non-stationary approach" above, where all coefficients depended on s and k. Thanks to the identification of repeated patterns, possibly combined with the existence of a trend taken into account by the non-stationary approach, we propose an intermediate modelling, where coefficients depend on s and on $k[M]$, the location of the train ride within a repeated pattern. However, our approach does not accommodate well deviations to the scheduled timetable so far, just as the one by Corman & Kecman [2]; it is thus somehow less flexible to these deviations than the approach of Li et al. [5], for instance, which however ignores a significant part of information available and rely on a local view given by a single train ride.

Note. Some ad hoc, possibly very complex models (e.g., deep-learning based treatment of images, see Bapaume et al. [1], Pasini et al. [6]), were also proposed for some of the variables considered, but they are out of the scope of this contribution, which targets generality and simplicity of the models constructed.

3. Presentation of the Data Set

Our approach crucially relies on having identical timetables from day to day: we therefore have to restrict our attention to working days. We do so for the morning peak hours (55 trains daily during the 6h33 - 9h28 time range) and for the 106-day-long period ranging from January 7, 2019 to July 5, 2019. We recall that we consider 20 stations. All in all, the data set consists of about 34,000 observed stops. Passenger flow variables (numbers A and B of passengers alighting and boarding, load at departure L) are measured by automatic passenger counting (APC) sensors. Railway operations variables (dwell time T and arrival delay ΔA) are measured by automatic vehicle localization (AVL) and track data. There are few missing values (10% for passenger flow variables and 6.5% for railway operations variables).

As explained in Section 2, abidance by the timetable is key to run our methodology. When trains take over or are suppressed, we clear locally the corresponding data (e.g., both train rides in case of a takeover). Doing so, 78% of the 34,000 potential observations are available and used.

We split the dataset into two data sets: a train set (January 7 - May 20) and a test set (May 21 - July 5), accounting for 70% and 30% of the observations, respectively. We estimate the parameters of the L-shaped regression models on the train set, by ordinary least squares, and compute their associated performance on the test set, which we report next. Our main indicator of performance is the mean absolute error (MAE).

4. Results

For the sake of space, we only report results for symmetric L-shaped neighbourhoods, i.e., with $P = Q$. The case $P = Q = 0$ corresponds to predictions given by average values on the train set, per cell (k, s). "Real" predictions based on the local context use $P = Q \geq 1$. Table 1 reports the global MAE (i.e., the MAE averaged over all possible cells of Figure 2) of the methods introduced in Section 2.2, for various values of $P = Q$. The reference method consists of the non-stationary L-shaped regression models with $P = Q = 1$, which is the closest to what the literature considered so far (see Section 2.3). We however note that for two variables at least (dwell time T and number of passengers alighting A), reporting average values (i.e., using $P = Q = 0$) results in decent predictions.

One issue of the reference method is the large number of coefficients to be estimated – around 900. We now address the wish to reduce the complexity of the method while preserving performance. There is a general balance in statistical models between their intrinsic ability to model the phenomenon at stake, which usually requires more coefficients, and the need to properly estimate these coefficients, which requires not having to estimate too many of them given a data set of fixed size.

Models			**Railway operations**		**Passenger flow**		
Name	*L-Shape*	Number of coefficients	T [s]	ΔA [s]	A [count]	B [count]	L [count]
Non-stationary	P = Q = 0	327	9,7	35,8	10	21	70
	P = Q = 1	915	9,5	16,1	9	18	22
Semi-stationary	P = Q = 1	403	9,2	16,1	10	18	23
	P = Q = 2	440	9,2	15,8	9	18	23
	P = Q = 3	466	9,1	15,8	9	18	23
Stationary	P = Q = 1	76	9,3	16,2	10	21	30
	P = Q = 2	113	9,2	15,8	9	20	29
	P = Q = 3	139	9,2	15,9	9	20	29

Table 1: Global MAE (mean absolute error) of some of the forecasting methods considered (indicated in the first two columns). For each method, we report the number of coefficients it uses (column 3), as well as the global MAE achieved for each of the five variables to be predicted (columns 4-8): dwell time T, arrival delay ΔA, numbers A and B of passengers alighting and boarding, load at departure L. The reference method is in blue and good alternative methods are in green.

For railway operations variables (dwell time T and arrival delay ΔA) stationary L-shaped regression models with $P = Q = 2$ rely on only about 100 coefficients while obtaining slightly better performance than the reference model. For passenger flow variables, a good alternative model consists of semi-stationary L-shaped regression

models with $P = Q = 1$: it requires about twice fewer coefficients than the reference method while obtaining an only slightly worse performance.

We move to a more local study of the performance, and report the difference of MAE between the reference method (non-stationary model with $P = Q = 1$) and the alternative methods (green cells in Table 1). We observe that for a majority of train-station pairs, the MAE (over repetitions within peak hours and over days) are almost equivalent, i.e., differ by at most 1 s or 1 passenger; see the white cells on Figure 5. This is especially remarkable for the dwell time T. The arrival delay ΔA and the number B of passengers boarding are locally better predicted by the alternative models. On the contrary, the alternative models are less accurate for the local prediction of the load at departure L (especially in one of the sub-branches) and the number A of passengers alighting (especially in the common part of the line).

References

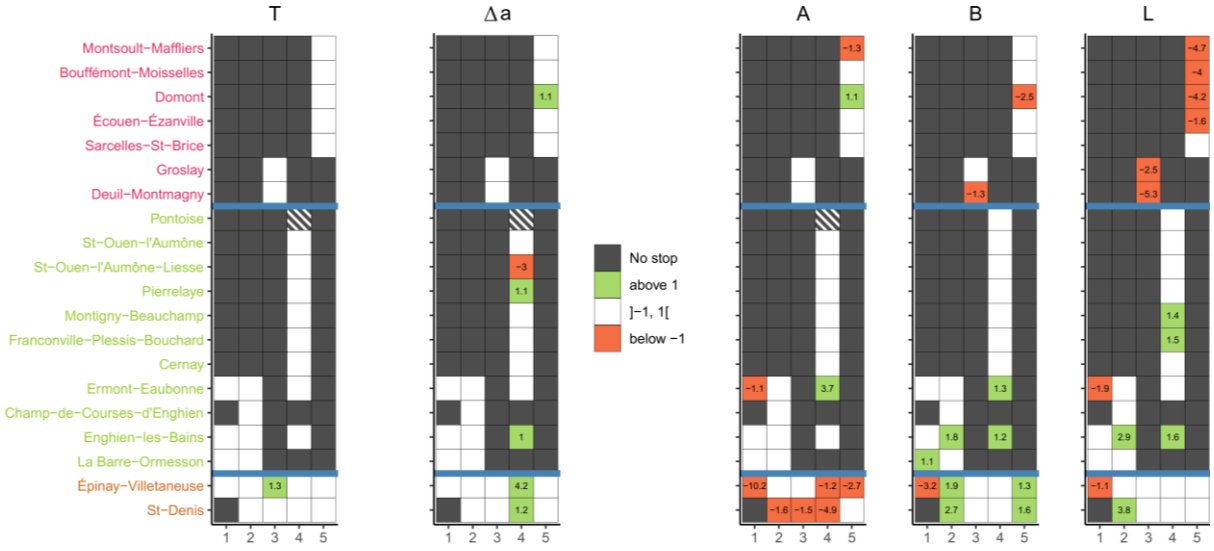


Figure 5: Average difference of performance between the alternative methods introduced in Table 1 and the reference method. Each column corresponds to a variable of interest. Each cell reports the average over corresponding train-station pairs during the peak hour and over the days. Significant improvements over the reference method are in green, deteriorations are in orange, while white denotes equivalence. The numbers indicate the average difference in MAE.

[1] Bapaume, T., Côme, E., Roos, J., Ameli, M., & Oukhellou, L., "Image Inpainting and Deep Learning to Forecast Short-Term Train Loads", *IEEE Access*, vol. 9, pp. 98506-98522, 2021.

[2] Corman, F., & Kecman, P., "Stochastic prediction of train delays in real-time using Bayesian networks", *Transportation Research Part C: Emerging Technologies*, vol. 95, pp. 599-615, 2018.

[3] Jenelius, E., "Data-driven metro train crowding prediction based on real-time load data", *IEEE Transactions on Intelligent Transportation Systems*, vol. 21(6), pp. 2254-2265, 2019.

[4] Kecman, P., & Goverde, R. M., "Predictive modelling of running and dwell times in railway traffic". *Public Transport*, vol. 7(3), pp. 295-319, 2015.

[5] Li, D., Daamen, W., & Goverde, R. M., "Estimation of train dwell time at short stops based on track occupation event data: A study at a Dutch railway station", *Journal of Advanced Transportation*, vol. 50(5), pp. 877-896, 2016.

[6] Pasini, K., Khouadjia, M., Same, A., Ganansia, F., & Oukhellou, L., "LSTM encoder-predictor for short-term train load forecasting", In *Proceedings of the ECML PKDD conference*, volume III, pp. 535-551, 2019.

How to use APC data to model passenger movement on-board? An application to Paris suburban train network

Ce travail est une version préliminaire du Chapitre 5. Il a été soumis et présenté à la conférence : international symposium on transport network reliability (INSTR) en juin 2021.

Coulaud, Rémi et Vimont, Mathilde. How to use APC data to model passenger movement on-board? An application to Paris suburban train network. In *8th International Symposium On Transport Network Reliability (INSTR)*, 2021.

How to use APC data to model passenger movement on-board? An application to Paris suburban train network

Rémi Coulaud[1,2] & Mathilde Vimont[2]

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

Extended abstract submitted for presentation at the 8th International Conference on Transport Network Reliability (Stockholm, 16-18 June, 2021)

1. Introduction

The continuous increase of passengers in Île-de-France transportation network attracts attention on non-uniform passenger distribution among coaches which reduces both passenger comfort and carrying capacity. The COVID-19 epidemic increases even more this need for precise information on passenger load inside trains (Tirachini & Cats 2020), especially if we consider the relatively high impact of passenger information on on-board crowding distribution (Zhang et al. 2017). Yet, allowing passengers to choose a boarding coach with respect to crowding can be done only if reliable on-board load information is available.

Numerous tools exist to estimate on-board crowding. Most of the literature relies on weight measurement in the air suspension system of the rolling stocks that allows a direct load measure (Jenelius 2019, Peftitsi et al. 2020). In some papers, the load measure is also obtained through infra-red or video sensors positioned on top of the train doors, as it is the case in Munich (Khomchuk et al. 2018). Though Automatic Passenger Counting (APC) gives an indirect measure of on-board crowding at a large scale (i.e, consist or train scale), it is difficult to get a reliable estimation at a smaller scale (i.e, coach scale) without taking into account passenger movement on-board.

In our case, passengers can board each coach through doors equipped with infra-red APC systems counting the number of alighting and boarding passengers at each stop. Within a consist (see Figure 2), coaches communicate to allow passengers to spread more uniformly between coaches. To our knowledge, few research studies have been led in such a context. Indeed, a large part of transportation literature studies the motivation behind boarding a specific coach as in Kim et al. (2014) or behind waiting for trains at a specific position along the platform (Hänseler et al. 2020). Schöttl et al. (2019) got closer to our problem and analysed the passenger seating strategy, yet not studying how to model their movements between coaches. There is also a large part of the pedestrian literature which took over quite similar issues, especially to understand alighting and boarding times thanks to either experimentation (Daamen et al. 2008) or automate cellular model (Seriani & Fujiyama 2019).

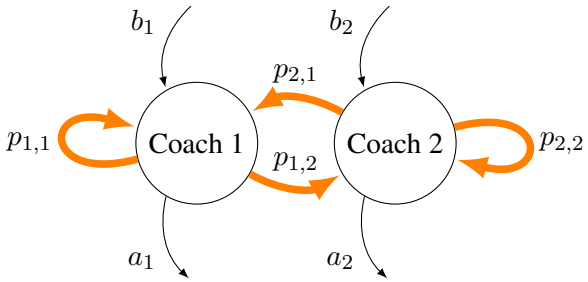


Figure 1: Illustration of the problem for a consist with two coaches

We propose to discretize a consist through a graphical model where each node represents a coach and each link between nodes is pondered by the probability to board one node and alight from the other node (see Figure 1). It naturally leads us to model passenger movement from one coach to another through a

multinomial probability distribution. Our model finds itself right in the middle of microscopic pedestrian simulation for platform-train or station pedestrian movement (Tang et al. 2017, Seriani & Fujiyama 2019) and macroscopic passenger movement in a transportation network through Origin-Destination matrix estimation (Van Zuylen & Willumsen 1980, Kuusinen et al. 2015).

2. Problem formulation

We recall that all doors of each consist are equipped with APC systems measuring the number of alighting and boarding passengers. For each coach $i = \{1, \ldots, C\}$, with C the maximum number of coaches per consist, let us define $b_i^{k,s}$ and $a_i^{k,s}$ the number of boarding and alighting passengers from coach i at station s for trip k[1]. We write the number of boarding and alighting passengers for trip k at station s as follows: $b^{k,s} = \sum_{i=1}^{C} b_i^{k,s}$ and $a^{k,s} = \sum_{i=1}^{C} a_i^{k,s}$.

S_k is the set of all stations during a trip k. Based on the conservation flow property of Kuusinen et al. (2015) our data-set is corrected such that there is no measurement error at the consist scale at the end of trip k:

$$\sum_{s \in S_k} b^{k,s} = \sum_{s \in S_k} a^{k,s}. \tag{1}$$

We also define the cumulative quantities by trip for coach i: $b_i^k = \sum_{s \in S_k} b_i^{k,s}$ and $a_i^k = \sum_{s \in S_k} a_i^{k,s}$, (see Table 1). The random variable $X_{i,j}$ is the number of passengers boarding coach i and alighting from coach j. Seemingly, $p_{i,j}$ is the probability for passengers to board coach i and alight from coach j. We interpret X margins as follows:

- $X_{i,\cdot} = (X_{i,1}, \ldots, X_{i,C})$ corresponds to the destination coaches (alighting coaches) vector for passengers that boarded coach i;
- $X_{\cdot,j} = (X_{1,j}, \ldots, X_{C,j})$ corresponds to the origin coaches (boarding coaches) vector for passengers that alighted from coach j.

Table 1: Notations and variables description

Notation	Description
i,j	coach number among {1, . . . , C} with C the maximum number of coaches
k	a unique trip id defined by the triplet: (consist (lead:1, rear:2), train number, day)
s	a station id
$b_i^{k,s}$	number of passengers boarding coach i for trip k at station s
$a_i^{k,s}$	number of passengers alighting from coach i for trip k at station s
b_i^k	total number of passengers boarding coach i for trip k
a_i^k	total number of passengers alighting from coach i for trip k
$b^{k,s}$	number of boarding passengers for trip k at station s
$a^{k,s}$	number of alighting passengers for trip k at station s
$X_{i,j}^k$	number of shifted passengers i.e, passengers moving from coach i to coach j for trip k

[1]Defined in Table 1

We used k in notations for the sake of completeness but we will only use it if needed in the following paragraphs. Contrary to the models used to estimate OD matrices, passengers move either to the left or the right side of the consist such that each coach j is accessible when boarding coach i. We suppose that passenger movement is modelled by a multinomial probability distribution conditionally to the number of boarding passengers for each door such that $X_{i,\cdot} \sim \mathcal{M}(b_i, p_{i,1}, \cdots, p_{i,C})$. The associated distribution function is defined $\forall x_{i,1}, \cdots, x_{i,C} \in (0, b_i)$ such that $\forall i \in \{1, \ldots, C\}, \sum_{j=1}^{C} x_{i,j} = b_i$:

$$\mathbb{P}(X_{i,1} = x_{i,1}, \cdots, X_{i,C} = x_{i,C} | b_i) = \frac{b_i!}{x_{i,1}! \cdots x_{i,C}!} p_{i,1}^{x_{i,1}} \cdots p_{i,C}^{x_{i,C}}.$$

We model passenger movement for each coach i such that we obtain a transition matrix given by, $p \in \mathbb{R}^{C \times C}$:

$$p = \begin{pmatrix} p_{1,1} & \cdots & p_{1,C} \\ \vdots & \ddots & \vdots \\ p_{C,1} & \cdots & p_{C,C} \end{pmatrix} \tag{2}$$

It verifies for all $i \in \{1, \ldots, C\}, \sum_{j=1}^{C} p_{i,j} = 1$ and for all $i, j \in \{1, \ldots, C\}^2$ $p_{i,j} \in [0, 1]$. The conditional expectancy of a single element of a multinomial random variable is $\mathbb{E}[X_{i,j} | b_i] = b_i p_{i,j}$. Each observation $k \in K$ corresponds to a unique trip. A classical estimator of $p_{i,j}$ is the maximum likelihood estimator defined as follows, $\forall i, j \in \{1, \ldots, C\}^2$:

$$\hat{p}_{i,j} = \frac{1}{K} \sum_{k=1}^{K} \frac{x_{i,j}^k}{b_i^k}.$$

This equation is used by Krstanoski (2014) to estimate passenger distribution on platform and by Ben-Akiva et al. (1985) to estimate OD matrices. However in both cases they observed $x_{i,j}^k$ while in our case there is no available measure of passenger movement inside consists. To overcome this difficulty, we generalise the conservation flow property (1) at the coach scale such that the total number of alighting passengers from coach i for a trip is supposed to be equal to the total number of passengers which moved to i:

$$\forall j \in \{1, \ldots, C\}, \quad a_j - \sum_{i=1}^{C} x_{i,j} = 0. \tag{3}$$

This leads us naturally to the following objective function defined by the conditional expectancy of the Euclidean distance between the total number of alighting passengers a and the random number of shifted passengers X:

$$\mathbb{E}\left[\sum_{j=1}^{C} \left(a_j - \sum_{i=1}^{C} X_{i,j} \right)^2 \middle| a, b \right],$$

of which an empirical version would be:

$$\frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{C} \left(a_j^k - \sum_{i=1}^{C} x_{i,j}^k \right)^2.$$

However, as we cannot observe $x_{i,j}^k$, we replace it by its theoretical conditional expectancy $b_i^k p_{i,j}$. The number of shifted passengers from coach i to j corresponds to a fixed proportion of the number of passengers having boarded i. It converts the problem into a general least square model under constraints:

$$
\begin{aligned}
\min_{p} \quad & \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{C}\left(a_j^k - \sum_{i=1}^{C} b_i^k p_{i,j}\right)^2 \\
\text{s.t} \quad & \forall i,j \in \{1,\dots,C\}^2,\ 0 \leq p_{i,j} \leq 1 \\
& \forall i \in \{1,\dots,C\}, \sum_{j=1}^{C} p_{i,j} = 1
\end{aligned} \tag{4}
$$

We easily find an analytical solution to this problem since it is equivalent to a regression problem where parameters are optimised under constraints.

To sum up the reasoning behind this model:

- we have observations of the number of boarding and alighting passengers for each door i but we do not observe passenger movement between communicating coaches;
- we assume that this passenger movement is driven by a multinomial probability distribution and we verify the generalised conservation flow property (1) at the coach scale;
- passenger movement is simplified as a parametric equation using transition matrix and boarding passengers by door.

We then introduce an additional assumption based on the willingness of passengers to move from their boarding coach. Indeed, it is classic to consider that passengers prefer to stay near their boarding coach (Kim et al. 2014). To take into account this hypothesis on passengers' behaviour, we add the following constraint to the optimisation problem (4), $\forall i,j \in \{1,\dots,C\}^2$ we have $\alpha_i \in [0,1]$ such that:

$$
p_{i,j} = p_{i,i} \times \alpha^{|i-j|}. \tag{5}
$$

α embedded the passengers' willingness to move. In this case, we loose the convexity of the problem and thus need to use a non convex optimisation solver.

To briefly conclude, we will compare three models:

1. A free movement model based on equation (4);
2. A constrained movement model based on equations (4) and (5);
3. A naive model, also called no-movement model, based on the identity transition matrix, which corresponds to a situation where all passengers stay in their boarding coach.

3. Case study

We applied this methodology framework to lines H and L of the Paris suburban train network (France). On those lines, trains are composed of two consists, each separated in either seven (line L) or eight (line H) connected coaches (see Figure 2).

(a) Line L

(b) Line H

Figure 2: Consist models for line L and H. Each coach is defined by a door (in purple)

Regarding line H, we only considered trains running on its western section, composed of 14 stations located between Paris Gare du Nord and Pontoise. Similarly, we only considered trains running on the

southern section of line L, composed of 16 stations located between Paris Saint-Lazare and Versailles-Rive-Droite (see Figure 3).

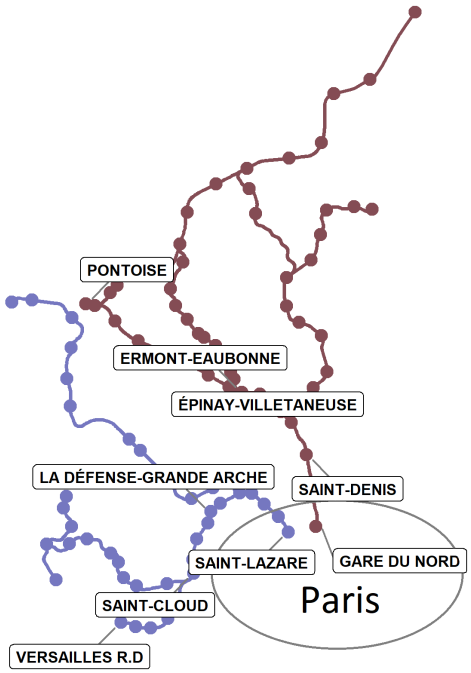


Figure 3: Geographic representation of line H and line L. Main stations of the considered sections are highlighted

We used APC data available for each coach at each train stop and aggregated the countings by trip. We respectively have data from the 1^{st} of September 2018 to the 31^{st} of August 2019 for the train data-set (line H $\approx$ 21,000 trips vs. line L $\approx$ 21,600 trips), and from the 1^{st} of September 2019 to the 30^{th} of November 2019 for the test data-set (line H $\approx$ 5,100 trips vs. line L $\approx$ 8,100 trips). This leads to a train/test ratio of 80/20. A sample data-set can be found in Table 5.

Transition probabilities are estimated on the train data-set, then we use the test data-set to evaluate models' performances. To estimate these probabilities, we also wanted to take stations' layout into consideration, as it has been shown to have an impact on passengers choices regarding boarding and alighting coaches (Kim et al. 2014, Fang et al. 2019). But as our objective function is defined at the trip scale, we hardly can take into consideration stations' layout as such in the optimisation problem. We partly overcame this issue by separating each data-set according to train ways of circulation (even: from suburb to Paris, odd: from Paris to suburb) and consist position (rear or lead consist) to estimate transition probabilities. Indeed, passengers' behaviour may differ between consists since they may be located differently relatively to platforms' entrances and exits. Similarly, passengers circulating on one way do not board and alight at the same stations and thus may behave differently within the train compared to the other way.

4. Results and conclusions

In this section, we present three main preliminary results:

- On-board passenger movement models decrease the difference between total alighting passengers and shifted passengers by coach;
- Passengers on both H and L lines rarely move by more than one or two coaches, which justifies the simpler constrained movement model;
- Some specific behaviours are revealed through the analysis of transition matrices.

Global performances and on-board behaviour

Naive, free and constrained movement models are evaluated through their objective function value (4) which is the difference between total alighting passengers (a) and total shifted passengers ($\hat{pb}$). We relied on the root of this quantity (i.e, RMSE) for ease of interpretation. They are gathered in Table 2.

Firstly, both movement models perform better than the naive model for the two lines, though the free movement model is the best, with an error divided by two for line H (**8.33** vs. 19.56) and almost by three for line L (**7.63** vs. 21.31). The improved performances of the free movement model compared to the constrained one are consistent with the more numerous parameters it relies on, thus allowing it to better fit the data.

Table 2: Performances of the three passenger movement models studied

	RMSE	
Models	Line H	Line L
Free movement	**8.33**	**7.63**
Constrained movement (α)	8.61	8.13
Naïve	19.56	21.31

We observe that performances of free and constrained movement models are very similar, with an error difference of only 0.28 and 0.50 for lines H and L respectively. The transition probabilities estimated through the free movement model also highlight that the probabilities of moving near the boarding coach (i.e, staying in the boarding coach or switching to one of the neighbour coaches) are high, with a probability averaged over all coaches comprised between 0.69 and 0.87 (see Table 3). This illustrates the passengers willingness to stay relatively close to their boarding coach.

Table 3: Probabilities of moving near the boarding coach (i.e staying in the boarding coach or switching to one of the neighbour coaches). To recall the circulation way, even: from suburb to Paris and odd: from Paris to suburb.

	Line H				Line L			
	Consist 1		Consist 2		Consist 1		Consist 2	
	Even	Odd	Even	Odd	Even	Odd	Even	Odd
Coach 1	0.90	0.88	0.70	0.73	1.00	0.94	0.95	1.00
Coach 2	0.73	0.75	0.83	0.90	0.97	0.74	0.91	0.91
Coach 3	0.75	0.75	0.58	0.71	0.85	0.83	0.82	0.90
Coach 4	0.72	0.65	0.79	0.75	0.86	0.72	0.82	0.79
Coach 5	0.72	0.62	0.70	0.74	0.86	0.81	0.81	0.87
Coach 6	0.52	0.85	0.89	0.79	0.95	0.74	0.89	0.93
Coach 7	0.55	0.59	0.90	0.80	0.58	0.48	0.56	0.40
Coach 8	0.59	0.66	0.26	0.27	-	-	-	-
Mean Coach	**0.69**	**0.72**	**0.71**	**0.71**	**0.87**	**0.75**	**0.82**	**0.83**

Specific movement strategies

If we take a closer look at these probabilities, we can spot a singularity affecting the rear coaches of trains circulating on both lines. Regarding line H, we notice a very low probability of staying near the 8^{th} coach of the second consist for both ways of circulation (0.26 for odd trains vs. 0.27 for the others). Figure 4 shows how passengers boarding this coach in trains circulating in the odd way tend to propagate towards the front of the consist. Indeed, transition probabilities for this coach are uniform (0.126 ± 0.031).

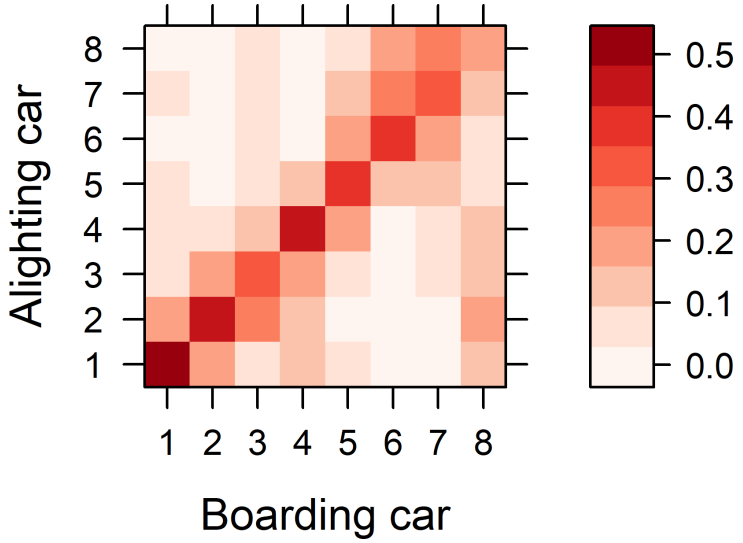


Figure 4: Transition matrices estimated with the free movement model for rear consists circulating on line H in the odd way

Yet, most of the passengers on this line section board at Paris Gare du Nord station, which is composed of a major platform entrance located near the rear of the train as shown in Figure 5. Passengers on this section also tend to alight at multiple stations along the trip, each having different layouts when it comes to platform exits.

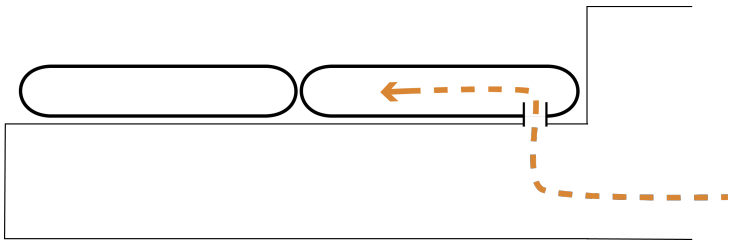


Figure 5: Drawing of Paris Gare du Nord platform station

The mean shares of boarding passengers in each coach at Paris Gare du Nord station presented in Figure 6 clearly show that most passengers board the 8^{th} coach of one of the two consists. Thus, the singular aforementioned probability could be due to passengers boarding the closest coach to the entrance platform at Paris Gare du Nord and then propagating themselves in the train, searching either for an available sit or a place near destination exit. This behaviour would be consistent with the literature that exists on passenger strategy in choosing boarding and alighting coaches (Kim et al. 2014, Fang et al. 2019). Also, the strong impact of Paris Gare du Nord station layout on estimated transition probabilities on line H highlights the relevance of trying to go one step further and to find a way of directly including stations' layout in the optimisation problem.

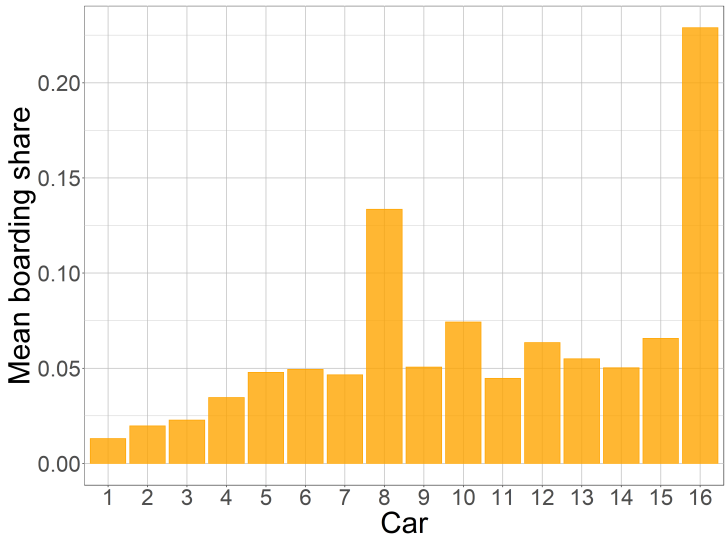


Figure 6: Mean boarding shares in each coach at Paris Gare du Nord station

Local performances

As the free movement model reduces the most the error made on the actual position of passengers within the train, we went further and studied how this model may correct for inconsistent load observations during the trip. More precisely, we noticed that not taking into account passenger movement within the consist led to numerous negative and extreme coach loads[2] during the trip, as well as plenty of non-null coach loads at terminus (i.e, the conservation passenger flow at the coach scale isn't satisfied). Those occurrences for both naive and free movement models are gathered in Table 4.

Table 4: Inconsistent coach loads observed during the trips on lines H and L for naive and free movement models

		Line H		Line L	
		Naive	Free movement	Naive	Free movement
Before terminus	Negative loads occurrences (%)	10.9	4.8	12.1	5.6
	Extreme loads occurrences[2]	4294	1	3346	231
At terminus	Non null loads occurrences (%)	85.0	83.0	83.4	81.3

One of the main improvements provided by the use of the model, is the almost disappearance of extreme loads for both H and L lines. Regarding line H, it is easily explained by the fact that most of these extreme loads occur at origin station Paris Gare du Nord for trains circulating in the odd way and affect the 8^{th} coach of the rear consist. The aforementioned study of transition matrices showed a clear effect of our model that is, the strong propagation of passengers from this coach to the rest of the consist. It is also consistent with the platform layout in Paris Gare du Nord.

[2]Above 120% of the coach capacity which is 138 passengers for line H and 130 passengers for line L

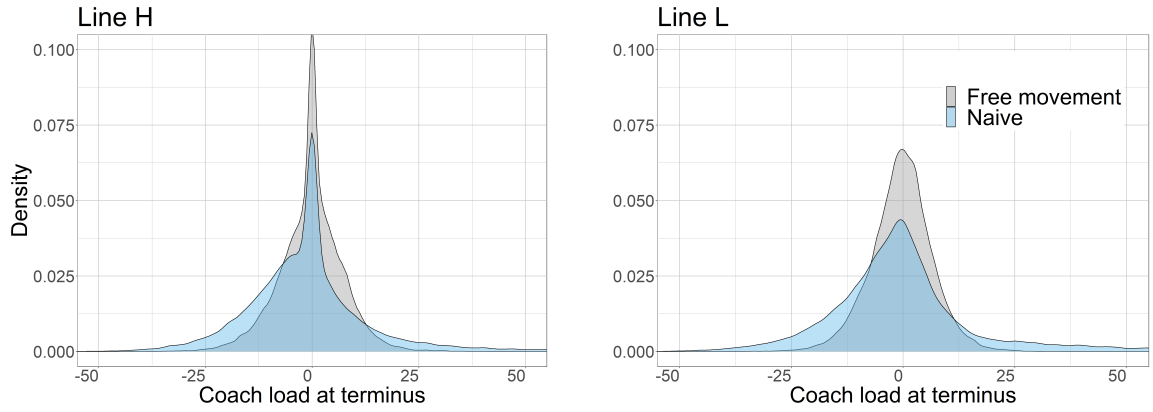


Figure 7: Distribution of coaches load at terminus for H and L lines

Our model also allows to improve the load estimation by reducing the number of negative coach loads during the trip by more than two for both lines: 4.8% vs. 10.9% occurrences in line H, and 5.6% vs. 12.1% occurrences in line L. Also, though the decrease in the number of non-null coach loads observed at terminus is quite weak (around 2% for both lines), Figure 7 shows that these loads are less scattered when the free movement model is used. This is consistent with the effect of the model on extreme loads, since reducing occurrences of extreme loads during the trip relies on homogenising the load along the consist.

References

Ben-Akiva, M., Macke, P. P. & Hsu, P. S. (1985), *Alternative methods to estimate route-level trip tables and expand on-board surveys*, number 1037.

Daamen, W., Lee, Y.-c. & Wiggenraad, P. (2008), 'Boarding and alighting experiments: Overview of setup and performance and some preliminary results', *Transportation Research Record* **2042**(1), 71–81.

Fang, J., Fujiyama, T. & Wong, H. (2019), 'Modelling passenger distribution on metro platforms based on passengers' choices for boarding cars', *Transportation Planning and Technology* **42**(5), 442–458.

Hänseler, F. S., van den Heuvel, J. P., Cats, O., Daamen, W. & Hoogendoorn, S. P. (2020), 'A passenger-pedestrian model to assess platform and train usage from automated data', *Transportation research part A: policy and practice* **132**, 948–968.

Jenelius, E. (2019), 'Data-driven metro train crowding prediction based on real-time load data', *IEEE Transactions on Intelligent Transportation Systems* **21**(6), 2254–2265.

Khomchuk, P., Tuladhar, S. R. & Sivananthan, S. (2018), 'Predicting passenger loading level on a train car: A bayesian approach', *arXiv preprint arXiv:1808.06962* .

Kim, H., Kwon, S., Wu, S. K. & Sohn, K. (2014), 'Why do passengers choose a specific car of a metro train during the morning peak hours?', *Transportation research part A: policy and practice* **61**, 249–258.

Krstanoski, N. (2014), 'Modelling passenger distribution on metro station platform', *International Journal for Traffic & Transport Engineering* **4**(4).

Kuusinen, J.-M., Sorsa, J. & Siikonen, M.-L. (2015), 'The elevator trip origin-destination matrix estimation problem', *Transportation Science* **49**(3), 559–576.

Peftitsi, S., Jenelius, E. & Cats, O. (2020), 'Determinants of passengers' metro car choice revealed through automated data sources: a stockholm case study', *Transportmetrica A: Transport Science* **16**(3), 529–549.

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

Seriani, S. & Fujiyama, T. (2019), 'Modelling the distribution of passengers waiting to board the train at metro stations', *Journal of Rail Transport Planning & Management* **11**.

Tang, T.-Q., Shao, Y.-X. & Chen, L. (2017), 'Modeling pedestrian movement at the hall of high-speed railway station during the check-in process', *Physica A: Statistical Mechanics and its Applications* **467**, 157–166.

Tirachini, A. & Cats, O. (2020), 'Covid-19 and public transportation: Current assessment, prospects, and research needs', *Journal of Public Transportation* **22**(1), 1.

Van Zuylen, H. J. & Willumsen, L. G. (1980), 'The most likely trip matrix estimated from traffic counts', *Transportation Research Part B: Methodological* **14**(3), 281–293.

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

Appendix

Table 5: First rows of train dataset on line H

Consist	Train Number	Date	b_1	a_1	b_1	a_2	b_3	a_3	b_4	a_4	b_5	a_5	b_6	a_6	b_7	a_7	b_8	a_8	Way
1	123400	05/03/2019	28	51	29	32	27	29	34	39	50	54	36	30	73	46	49	45	0
1	123409	22/02/2019	11	13	3	9	11	19	22	30	18	17	34	24	27	50	77	41	1
1	123419	16/12/2018	2	2	1	0	8	8	6	6	4	5	14	13	13	20	16	10	1
1	123499	14/06/2019	15	24	25	20	43	45	32	23	30	46	52	39	51	64	97	84	1
1	123543	16/08/2019	17	24	16	17	20	20	12	28	41	26	21	29	29	42	81	51	1
2	123400	06/12/2018	49	42	31	34	31	22	38	41	28	30	24	29	34	35	20	22	0
2	123402	15/05/2019	28	32	39	43	26	23	20	21	34	27	23	25	26	27	8	6	0
2	123424	18/03/2019	90	63	45	57	47	45	44	54	52	39	53	57	39	60	33	28	0
2	123490	28/12/2018	93	69	53	58	50	47	27	39	27	25	16	17	26	34	14	17	0

C

Share of Strategic Alighting Passengers combining Automatic Passenger Counting and OpenStreeMap

Cet article est un travail complémentaire au Chapitre 5. Nous proposons une mesure objective de la proportion de voyageurs stratégiques au moment de descendre du train et de sortir de la gare, appelés stratégiques en sortie. Ces voyageurs stratégiques en sortie cherchent à minimiser leur distance de marche à la gare de destination. Il a été soumis et sera présenté à la conférence : Conference on Advanced Systems in Public Transport (CASPT) en novembre 2022.

Coulaud, Rémi, Mazon, Valentine, Sanchis, Laura, et Cats, Oded. Share of strategic alighting passengers combining automatic passenger counting and OpenStreeMap. In *Conference on Advanced Systems in Public Transport (CASPT)*, 2022.

CASPT 2022

Share of Strategic Alighting Passengers combining Automatic Passenger Counting and OpenStreeMap

Extended Abstract

Rémi Coulaud · Valentine Mazon · Laura Sanchis · Oded Cats

Abstract Understanding passengers' distribution on-board trains and along public transport platforms is crucial for improving service's performance and ensure passengers' comfort. We propose a revealed preference measure of passengers alighting behaviours using automatic passenger counting (APC) data. Our findings revealed that the share of strategic alighting passengers per station is influenced by its layout and the overall passengers volume at this given station.

Keywords Passenger counts, strategic alighting passenger, passenger behaviour, passenger distribution, empirical study

1 Introduction

In public transports, most passengers board and alight neither randomly, nor uniformly, leading to very heterogeneous crowding inside trains and along public transport platforms. Yet, one critical overcrowded zone of a platform or a train can have great impact on many aspects of the service (e.g., dwell time, service punctuality, passengers' comfort). Hence, crowding has long been recognised both as an indicator of public transport performance as well as an important measure of passenger satisfaction with the service (Szplett and

Rémi Coulaud
Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay
Orsay, France
E-mail: remi.coulaud@sncf.fr

Valentine Mazon and Laura Sanchis
SNCF Voyageurs – Transilien
Saint-Denis, France

Oded Cats
Department of Transport and Planning, Delft University of Technology
Delft, The Netherlands

Wirasinghe, 1984; Kim et al, 2014; Börjesson and Rubensson, 2019). A deeper understanding of passengers' flow and distribution in public transports became all the more crucial with the COVID pandemic, since these flows were rapidly and constantly changing, as a consequence of evolving governmental restrictions undertaken in all countries. To tackle this issue, many studies have focused on understanding the underlying reasons that give rise to the emergence of passengers' uneven distribution along platforms and how crowding valuation can influence their route choice (Drabicki et al, 2021). Most of these studies, reported in Table 1, rely either on stated preferences collected through field surveys, or revealed preference obtained from passenger count data.

Among the few studies exploring passengers' positioning choice through stated preferences is the one reported in Kim et al (2014). Results revealed that 53% of passengers intentionally choose a specific car with the aim of minimising walking distance at their destination station. These findings were later confirmed by Elleuch (2019) who found a very similar share (54%) for Paris region. We then refer to these passengers as strategic alighting passengers (SAP). Finally, Szplett and Wirasinghe (1984), Krstanoski (2014) analysed alighting and boarding distribution using revealed preferences data only, namely manual counting measures. They noticed that passengers' distribution is significantly influenced by the station layout. Our principal contribution lies within the proposed methodology – quantifying the proportion of strategic alighting passengers, using exclusively automatic passenger counting (APC) and OpenStreetMap mapping data, thus enabling a larger temporal and spatial study scope.

Table 1: Literature review on passenger's behaviour on-board and at the train-station interface

Authors	**Study interest**	**Data collection**
Kim et al (2014)	Boarding	Survey
Elleuch (2019)	Boarding	Survey
Szplett and Wirasinghe (1984)	Boarding/Alighting	Manual counting
Krstanoski (2014)	Boarding/Alighting	Manual counting
Drabicki et al (2021)	Boarding/Alighting	
Our study	**Alighting**	**APC**

2 Measuring willingness to minimise walking distance at station

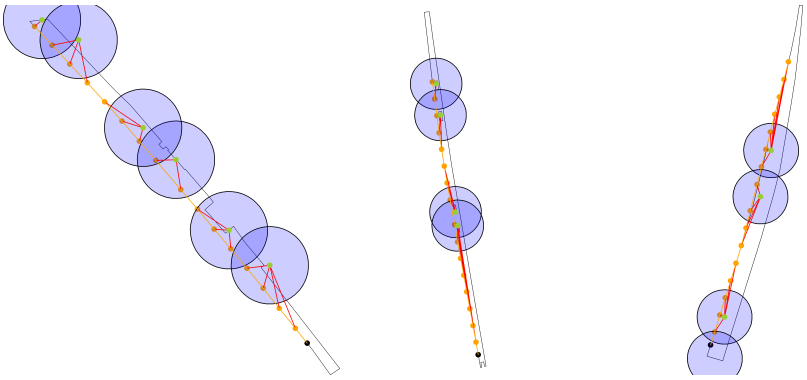
We build a method to compute SAP based on revealed preference through APC counts per door. The quality of APC data was confirmed by a field survey revealing a 95% precision for alighting $a^i_{k,s,d}$ and boarding $b^i_{k,s,d}$ passengers measure per door $i \in \{1, \ldots, I\}$ for each stop defined by a train number k, a station s and a day d.

In order to analyse SAP, we link each door identifier to its precise location along the platform and consequently, its distance to all platforms exit points. We use OpenStreetMap (OpenStreetMap contributors, 2017) to retrieve the platform layout: borders, exit and entrance points. Finally, we use stop signal location and rolling stock characteristics to deduce doors' location on platform as illustrated in Figure 1. For each station, we define doors V_s^i and exits E_s^j coordinates. We then compute for each door i and each station s, the distances to all platform exits j: $d(V_s^i, E_s^j)$ and identify, for each door of a train, the distance to the nearest platform exit noted $d_{s,min}^i$.

$$d_{s,min}^i = \min_{j=1,\dots,J} d(V_s^i, E_s^j)$$

The distance used is the great circle distance based on the spherical reference of earth WSG 84, displayed in red in Figure 1.

We then search the number of passengers choosing to minimise their walking distance once arrived i.e the share of alighting passengers near platform exits. We define the platform exit attractiveness as a circle area of radius r (meters) centred around the platform exit location, the blue circles in Figure 1. We assume that all platform exits have the same attractiveness.



(a) Epinay-Villetaneuse (b) Saint-Denis (c) Paris gare du Nord

Fig. 1: Platform graph with a 20 meters exit attractiveness for three different platforms used by trains running from the suburbs to Paris. We display several components: • Exit location; ○ 20 meters exit attractiveness; • Stop signal location; • Door location; - Distance to the nearest platform exit per door.

Once these areas are set, we categorise doors into strategic or not. A door $i \in \{1, \dots, I$ is strategic if it belongs to an exit attractiveness area; that is, a door located within r meters of an exit: $d_{s,min}^i \leq r$. Finally, we derive the share of SAP with respect to a given exit attractiveness r:

$$SAP_r = \frac{\sum_{i \in \mathcal{I}} a_{k,s,d}^i}{a_{k,s,d}}$$

where $a_{k,s,d} = \sum_{i=1}^{I} a^i_{k,s,d}$ is the total number of alighting passengers for one stop. SAP_r will always be an upper bound of SAP because we do not control for boarding position as we do not know origin of the alighting passengers and rolling stocks are car communicant (i.e, passengers can move within the same consist).

3 Share of strategic alighting passengers and its variability in space and time

We compute the share of SAP for the first 6 stations of Line H, see Figure 2, using APC data per door from the 1^{st} of April to the 30^{th} of June 2019. We discard one-unit trains because their stop signal position may vary from two units' trains. In total, the dataset contains observations from 31,000 train stops going from the suburbs to Paris and 31,300 train stops going from Paris to the suburbs. We first determine the right exit attractiveness radius to compute the share of SAP.

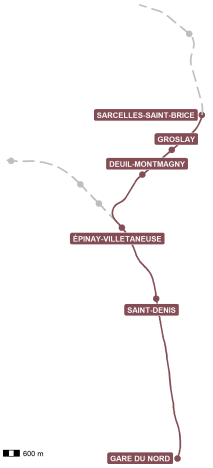


Fig. 2: Spatial perimeter of the study on line H

In Figure 3, we present the variability of SAP per stations as a function of exit attractiveness. The observed differences are mainly due to the varying exits number by platform and their location. For instance, as shown in Figure 1, Epinay-Villetaneuse platform for trains running to Paris has many exits, which are well distributed along the platform such that the share of SAP increases rapidly while it is not the case for Saint-Denis or Gare du Nord. An exit attractiveness of 20 meters seems just enough to be consistent with previous studies exploring this phenomenon using stated preferences which found a SAP share of 54%.

The proportion of SAP is not only influenced by the platform layout, but also fluctuates during the day. From Figure 4, we see that for trains running to Paris, the proportion of SAP is the highest during morning rush hours for Saint-Denis but not for Gare du Nord. The same result was observed for trains running to the suburbs, as the SAP for Groslay is the greatest during evening rush hours but not for Sarcelles-Saint-Brice. We believe this result for Gare du Nord and Sarcelles-Satin-Brice is mostly due to a large increase in alighting passengers' volumes, which prevents passengers from intentionally choosing a specific car to alight, due to on board crowding conditions. Indeed, in Figure 5, we see a clear effect of alighting passengers' volume on the proportion of SAP, which decreases by 10-20 points when comparing situations with few and many alighting passengers.

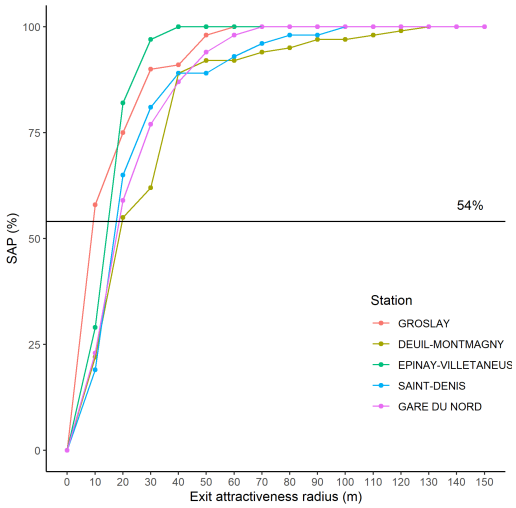
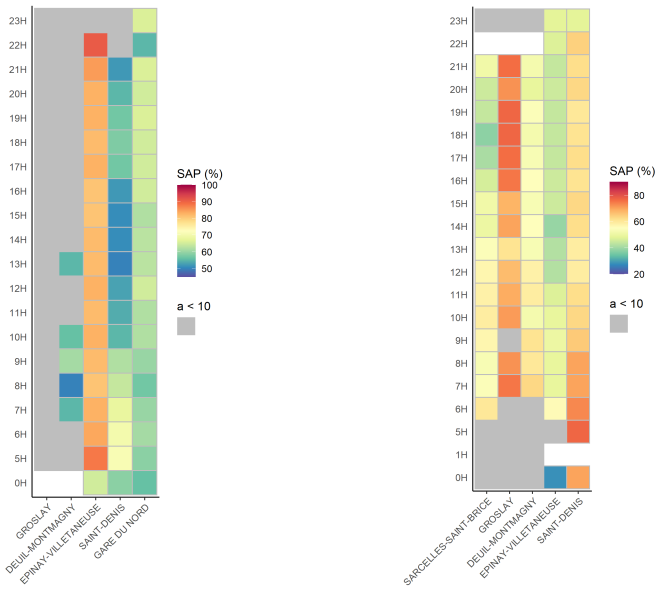


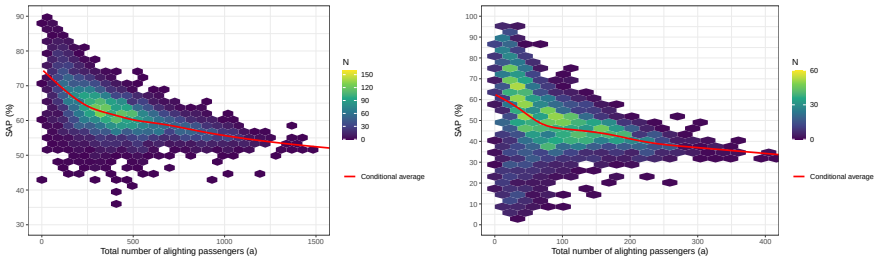
Fig. 3: Share of SAP per station with respect to a uniform 10m increase of exit attractiveness. The studied rolling stock has doors that are 13.2m apart so 10m is almost equivalent to adding a door. The 54% line represents Kim et al (2014) previous findings regarding the share of SAP.



(a) Suburbs to Paris

(b) Paris to the suburbs

Fig. 4: Share of SAP_{20m} on 6 stations during working days for trains running from the suburbs to Paris (left) and from Paris to the suburbs (right). Grey periods have an average number of alighting passengers by stop below 10.



(a) Paris Gare du Nord for trains running from the suburbs to Paris

(b) Sarcelles-Saint-Brice for trains running from Paris to the suburbs

Fig. 5: Share of SAP with an exit attractiveness of 20m with respect to the total number of alighting passenger for two selected platforms. The conditional average is depicted in red and is computed through generalised additive models.

4 Outlooks

In this work we propose an intuitive way of computing the share of SAP using APC data. We see three directions to go further: (i) we want to improve the SAP indicator taking into account the platform layout where alighting passengers board; (ii) we want to model the effect of volume and time of the day on SAP; (iii) we will design a method to locate exits using only on APC data.

References

Börjesson M, Rubensson I (2019) Satisfaction with crowding and other attributes in public transport. Transport policy 79:213–222

Drabicki A, Kucharski R, Cats O, Szarata A (2021) Modelling the effects of real-time crowding information in urban public transport systems. Transportmetrica A: Transport Science 17(4):675–713

Elleuch F (2019) Transférabilité d'une modélisation-simulation multi-agents: le comportement inter-gares des voyageurs de la sncf lors des échanges quai-train. PhD thesis, Conservatoire national des arts et metiers-CNAM

Kim H, Kwon S, Wu SK, Sohn K (2014) Why do passengers choose a specific car of a metro train during the morning peak hours? Transportation research part A: policy and practice 61:249–258

Krstanoski N (2014) Modelling passenger distribution on metro station platform. International Journal for Traffic & Transport Engineering 4(4)

OpenStreetMap contributors (2017) Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org

Szplett D, Wirasinghe S (1984) An investigation of passenger interchange and train standing time at lrt stations. Journal of advanced transportation 18(1):1–12

D

Modélisation de l'impact des flux voyageurs sur les temps d'échange pour la simulation des marges d'exploitation : une application à la ligne N de Transilien

Cet article est un travail complémentaire au Chapitre 3. Il aborde la question du calcul *a posteriori* de marges des temps de stationnement théoriques. Ce travail est un premier pas vers l'optimisation des temps de stationnement théorique en fonction des flux de voyageurs en phase tactique. Il a été soumis et présenté aux Rencontres Francophones Transport Mobilité (RFTM) en juin 2022.

Coulaud, Rémi et Grangé, Martine. Modélisation de l'impact des flux voyageurs sur les temps d'échange pour la simulation des marges d'exploitation : une application à la ligne N de transilien. In *4èmes Rencontres Francophones Transport Mobilité (RFTM)*, 2022.

Modélisation de l'impact des flux voyageurs sur les temps d'échange pour la simulation des marges d'exploitation : une application à la ligne N de Transilien

Rémi Coulaud[a], Martine Grangé[a]

a - Transilien, SNCF Voyageurs, 12 rue Jean Philippe Rameau, 93220, Saint-Denis, France

Introduction

Le temps passé à l'arrêt sur le réseau Transilien représente 20 à 30% du temps de trajet total d'un voyageur. On identifie trois composantes du temps d'arrêt (Figure 1) : un temps technique (ouverture et fermeture des portes), un temps d'échange (entre le premier et dernier échange voyageurs) et des marges (le temps restant). La marge au train est le temps entre la fin de l'échange à la porte critique* et l'initialisation de la fermeture des portes. La marge à la porte est le temps entre la fin de l'échange à la porte considérée et la fin de l'échange à la porte critique. Par ailleurs, on sait grâce à Wiggenraad (2001), Szplett & Wirasinghe (1984) et Coulaud *et al.* (2022) que les échanges voyageurs ne se répartissent pas uniformément le long du train. Les temps d'échange à la porte critique* peuvent être significativement plus longs qu'aux autres portes. L'objectif de ce travail est de quantifier la marge au train et le déséquilibre de marge entre les portes.

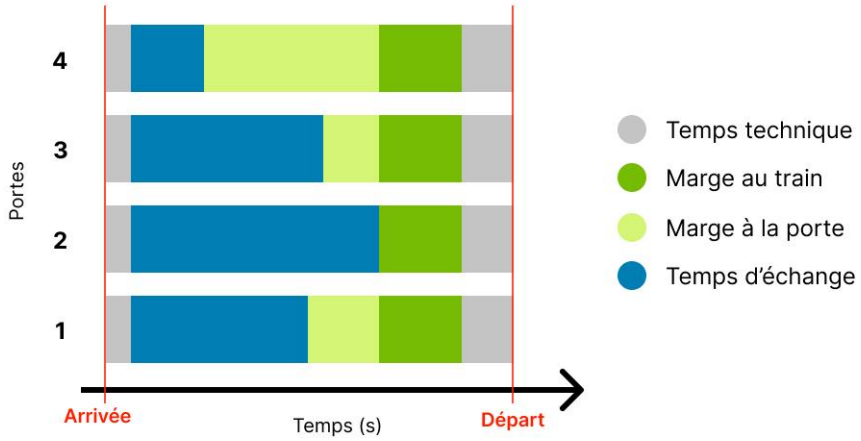


Figure 1 : Schéma de décomposition du temps de stationnement (temps technique + temps d'échange + marges) pour 4 portes d'un train

Pour répondre à cet objectif dans un contexte ferroviaire contraint, il est nécessaire : (i) de mesurer le nombre de passagers montant/descendant par train et/ou par porte (ii) de mesurer les temps d'échange. Dans ce travail, nous mesurons de façon automatique et par porte le nombre de montées et descentes comme Wiggenraad (2001) et Buchmüller *et al.* (2008).

Buchmüller *et al.* (2008), Medeossi & Nash (2020), Coulaud *et al.* (2022) ne mesurent pas le temps d'échange mais le substituent par les temps de stationnement des trains

en retard où la marge à la porte critique est supposée nulle. Dans cette étude, nous mesurons le temps d'échange à partir d'une série de mesures horodatées enregistrées à chaque porte lors de l'échange.

Nos contributions :

1. Calcul du temps d'échange à la porte critique. Nous confirmons que les marges pour les trains en avance sont plus grandes que pour les trains en retard ;
2. Calcul du temps d'échange pour toutes les portes afin d'observer la somme des marges par porte donc la perte de temps dû à la mauvaise répartition des passagers le long du quai ;
3. Simulation des temps d'échange pour optimiser les marges opérationnelles.

Des événements de comptage à la mesure du temps d'échange

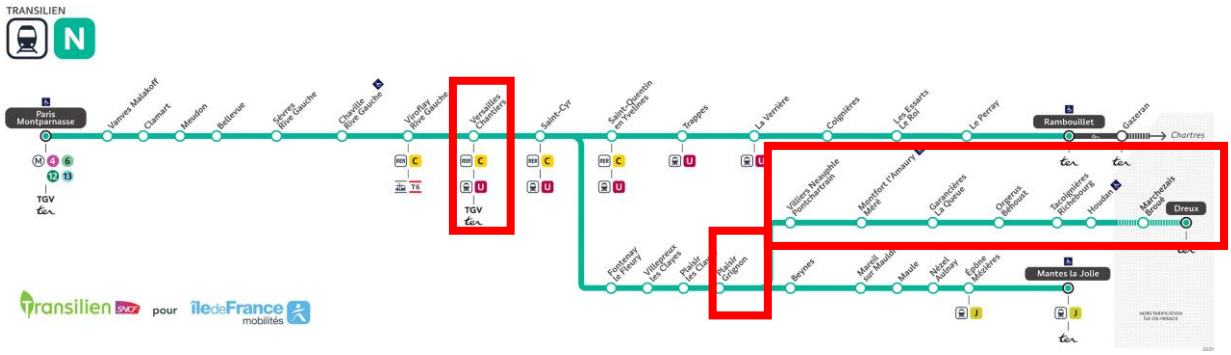


Figure 2 : Plan de la ligne N avec en rouge le périmètre d'étude

Nous utilisons des données de comptage infra-rouge issues des trains Regio2N composés de 8 portes de 160cm de large (Figure 3) qui circulent sur la branche Montparnasse-Dreux de la ligne N (Figure 2). Un événement de comptage représente le nombre de voyageurs passés sous les capteurs dans les quelques secondes précédant l'enregistrement.

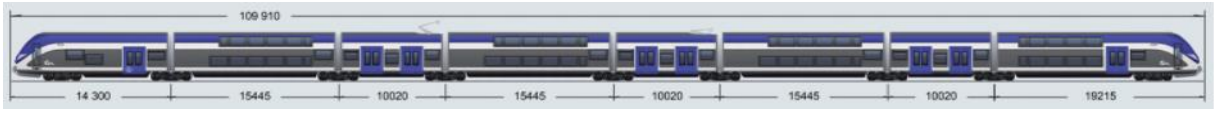


Figure 3 : Schéma d'une rame Regio2N circulant sur la branche Montparnasse - Dreux

Pour modéliser les temps d'échange, il faut convertir ces événements de comptage en temps d'échange exploitables. Nous définissons la notion de cluster de voyageurs, inspirée de Wiggenraad (2001), comme une file de voyageurs continue telle que deux voyageurs de cette file sont séparés par moins de deux secondes (Figure 4). Si plusieurs clusters sont observés à la même porte, nous définissons le temps d'échange comme l'écart de temps entre le début du premier cluster et la fin du dernier cluster. Par exemple sur la Figure 4 (en bas), le premier cluster de la porte 6 (P6)

commence 2s après l'arrivée en gare et se termine à 25s ce qui correspond à un temps d'échange de 23s.

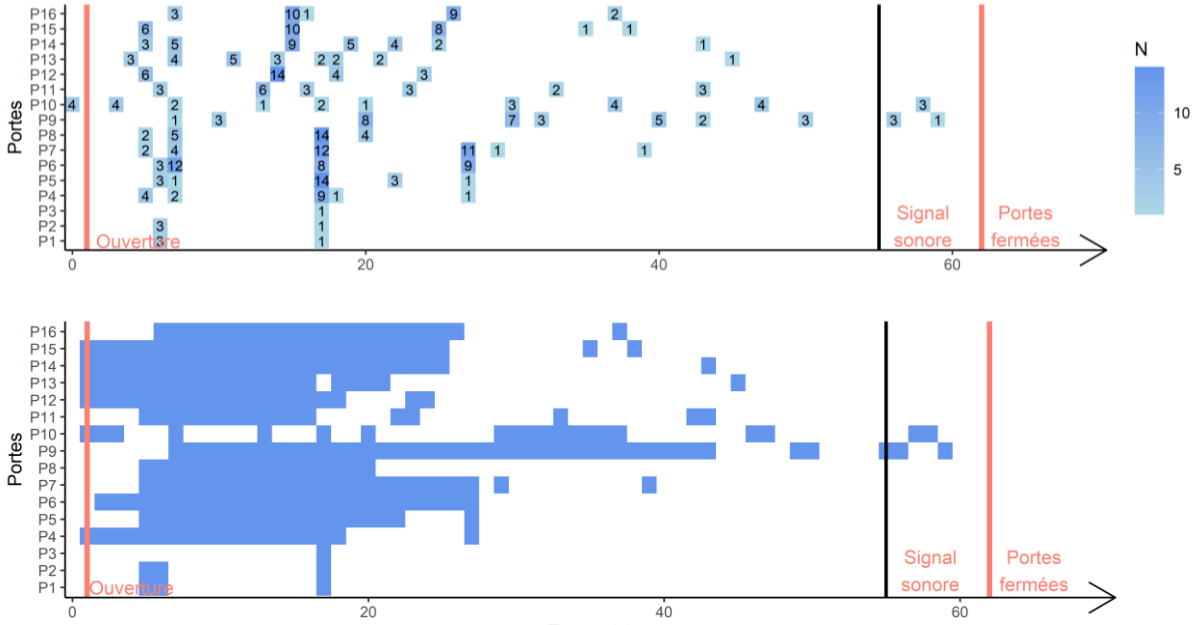


Figure 4 : Illustration pour un arrêt d'un train avec 16 portes de la méthode de conversion des événements de comptage (en haut) au temps d'échange en passant par les clusters de passagers (en bas). En haut, les événements de comptages mesurés par les capteurs. En bas, les zones bleues représentent le lissage des événements de comptage en clusters de voyageurs.

2500 arrêts caractérisés par un numéro de train k, une gare s et une date d composent les données. Ces observations proviennent de trains ayant circulé dans les deux sens de circulation de septembre à octobre 2021 dans 9 gares de la branche Montparnasse-Dreux de la ligne N (Figure 1).

Pour modéliser les temps d'échange, nous utilisons des variables, décrites dans la Table 2, spécifiques aux flux voyageurs (nombre de montées, nombre de descentes et le taux d'occupation) et spécifiques aux gares (largeur des quais, espacement vertical et horizontal à l'interface quai-train, nombre d'entrées/sorties). Nous n'avons pas de variables spécifiques au matériel, contrairement à Harris *et al.* (2022), car les données proviennent d'un matériel unique.

Table 2 : Variables disponibles pour la modélisation du temps d'échange : variables spécifiques aux flux voyageurs (en haut) et variables spécifiques aux quais (en bas).

Variables	*Domaine*	*Echelle*	*Notation*
Temps d'échange	[0, 150]	Secondes	$Y^i_{k,s,d}$
Variables spécifiques aux flux voyageurs			
Nombre de montées par porte	[0, 1, …, 86]	Voyageurs	$B^i_{k,s,d}$
Nombre de descentes par porte	[0, 1, …, 136]	Voyageurs	$A^i_{k,s,d}$
Nombre de voyageurs par porte	[0, 1, …, 145]	Voyageurs	$N^i_{k,s,d} = B^i_{k,s,d} + A^i_{k,s,d}$
Charge à bord	[0, 1, …, 2058]	Voyageurs	$L^i_{k,s,d}$
Taux d'occupation	[0, 2]		$O^i_{k,s,d} = L^i_{k,s,d}/Cap$
Variables spécifiques aux quais			
Largeur des quais	Étroit, moyen, large		$Q_{s,w}$
Espacement vertical à l'interface quai-train	[0, 1, …, 35]	cm	$H_{s,w}$
Espacement horizontal à l'interface quai-train	[0, 1, …, 26]	cm	$V_{s,w}$
Nombre d'entrées/sorties	[1, 2, …, 5]		$E_{s,w}$

Modélisation du temps d'échange

a. Quelles variables pour l'estimation des temps d'échange ?

Le temps d'échange (Y) par porte permet de réutiliser facilement dans le ferroviaire des stratégies de modélisation déjà éprouvées pour la modélisation des temps de stationnement des métros (Harris *et al.* 2022). La sélection des meilleures variables se fait habituellement en deux étapes :

1. Ajout des variables de flux voyageurs transformées : $X^2, \sqrt{X}, log(X)$ et des interactions deux à deux entre toutes les variables ;
2. Sélection des meilleures variables au sens du Bayesian Information Criterium (BIC) sous l'hypothèse gaussienne par ajout/retrait successif de variables.

Cet ensemble de variables final (X) permet d'estimer sur le jeu de données d'entrainement (70% des données) un modèle pour l'estimation des temps d'échange.

b. Quel modèle pour l'estimation des temps d'échange ?

Notre objectif est double : estimer correctement les temps d'échange par porte à partir d'un modèle unique pour toutes les portes, gares, trains et jours ; simuler avec un certain niveau de confiance (50%, 90%, 99%) les temps d'échange. Les auteurs précédents (Harris *et al.* 2022) considèrent que les temps de stationnement et *a fortiori* les temps d'échange sont gaussiens. Nous testons trois modèles :

a. Un modèle linéaire gaussien :

$$Y = \beta X + \varepsilon, \text{ avec } \varepsilon \sim N(\mu, \sigma^2)$$

b. Un modèle linéaire généralisé avec une distribution Gamma tel que $Y \sim \Gamma(k, \theta)$:

$$\log(k\theta) = \beta X$$

c. Une approximation de ce modèle en utilisant le log(Y) tel que :

$$log(Y) = \beta X + \varepsilon, \text{ avec } \varepsilon \sim N(\mu, \sigma^2)$$

c. Simulation des marges au train et à la porte

Les modèles probabilistes permettent de simuler, une fois les paramètres estimés $\hat{\beta}$, plusieurs temps d'échange (Y_α), plus ou moins tendus, pour absorber un volume de voyageur donné (N) en fonction de la marge α définie préalablement :

$$P\left((Y|N) \leq Y_\alpha, \hat{\beta}\right) = \alpha$$

Si $\alpha = 99\%$, cela signifie que, dans 99% des cas, le temps d'échange sera suffisant pour laisser monter et descendre les voyageurs.

Résultats préliminaires

a. Le calcul de marge opérationnelle à l'échelle du train en fonction du régime de ponctualité du train

La Table 3 permet de confirmer que lorsqu'un train est en retard, sa marge moyenne est plus faible (11 secondes), que lorsqu'il est en avance (39 secondes).

Table 3 : Marge moyenne en fonction du régime de ponctualité : en avance (heure d'arrivée observée < heure d'arrivée théorique), en retard (heure d'arrivée observée > heure de départ théorique), à l'heure (entre les deux).

	en avance	*à l'heure*	*en retard*
Moyenne (s)	39	16	10.7

b. Le calcul de marge opérationnelle à l'échelle de la porte

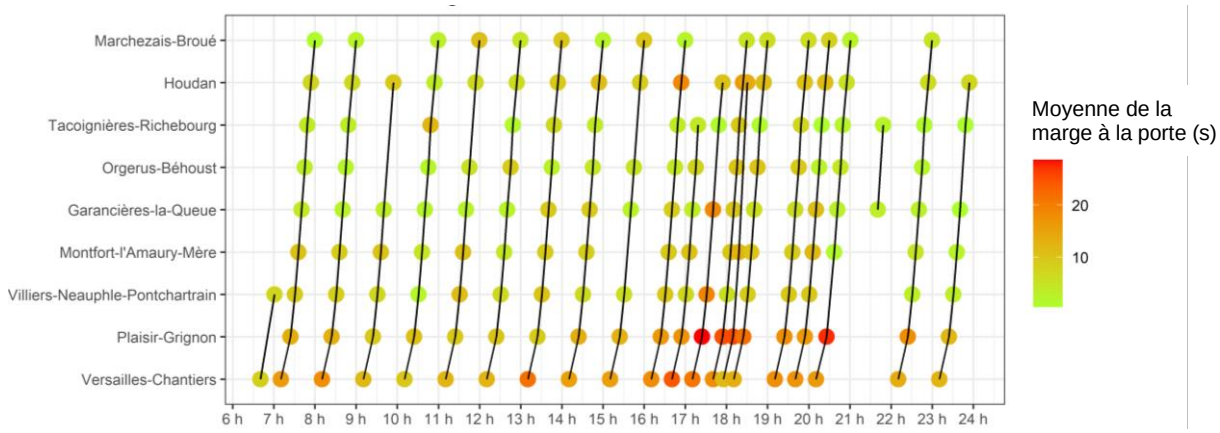


Figure 5 : Moyenne de la marge à la porte en fonction de la gare et de l'heure de l'arrêt

Premièrement, d'après la Figures 5, la moyenne de la marge à la porte est plus forte à Versailles-Chantiers. Cela signifie que les passagers sont moins bien répartis le long du quai à Versailles que dans les autres gares. Deuxièmement, en heure de pointe, la répartition des passagers est plus hétérogène. Ces résultats nous amènent donc à penser que la répartition des passagers est moins homogène quand il y a plus de monde et que cette répartition varie de gare à gare.

.

c. L'estimation d'un modèle probabiliste pour le temps d'échange

Nous estimons sur notre jeu de données d'entrainement les trois modèles présentés dans la section 3. Pour choisir un des trois modèles (i) nous calculons des métriques standards sur notre jeu de données test : RMSE, MAE, MAPE (Table 4) (ii) nous comparons les fonctions de répartitions des prédictions du temps d'échange (Figure 6).

Table 4 : Calcul des erreurs de prévisions du temps d'échange des trois modèles sur notre jeu de données test

	RMSE (s)	*MAE (s)*	*MAPE (%)*
Normal	6,1	3.7	44%
Gamma	6,9	4.3	57%
Log-Normal	6,5	4.0	40%

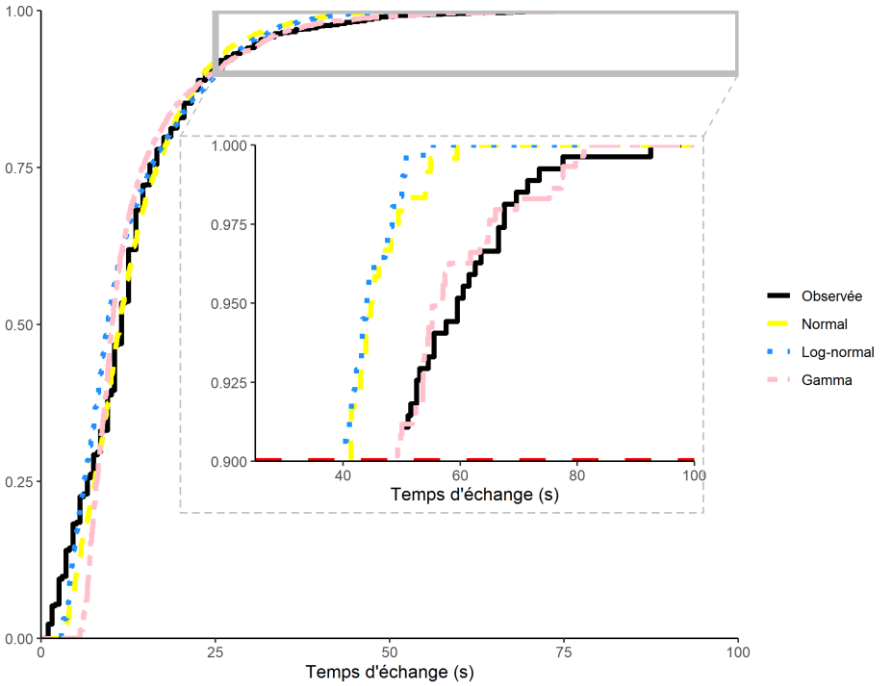


Figure 6 : Fonction de répartition empirique du temps d'échange estimée sur le jeu de données test pour les trois modèles proposés comparés à l'observée. Zoom sur la queue de distribution dans l'encadré en pointillé.

D'après la Table 4 et la Figure 6, nous choisissons le modèle linéaire gaussien pour la suite car il permet de mieux estimer le temps d'échange en moyenne. Il reste que le modèle Gamma semble plus adapté pour capturer la queue lourde du temps d'échange, voir zoom de la Figure 6.

d. La simulation de temps d'échange à partir de ce modèle pour des gains de performance

A partir du choix du modèle, une estimation de la marge au train dans un intervalle de confiance donné est réalisée. Ainsi, deux minutes peuvent être gagnées par trajet, avec une probabilité de 99% de permettre à tous les passagers de monter dans le train. Ce gain de temps est décliné par gare et représente jusqu'à 30 secondes à Marchezais-Broué. Toutefois, il est clair que d'autres facteurs que les flux de voyageurs peuvent influencer la construction de temps de stationnement théorique.

Bibliographie

Buchmüller, S., Weidmann, U., & Nash, A. (2008) Development of a dwell time calculation model for timetable planning. *WIT Transactions on The Built Environment,* 103, 525-534.

Coulaud R., Mazon V., Sanchis L., & Cats O. (2022) Share of Strategic Alighting Passengers combining Automatic Passenger Counting and OpenStreeMap. Preprint

Coulaud R., Keribin C., & Stoltz G., (2022) Modeling dwell time in a data-rich railway environment: with operations and passenger flows data, Preprint

Daamen W., Lee Y., & Wiggenraad P. (2008) Boarding and alighting experiments, *Transportation Research Record: Journal of the Transportation Research Board*, 2042, 71-81.

Harris, N. G., de Simone, F., & Condry, B. (2022). A Comprehensive Analysis of Passenger Alighting and Boarding Rates. *Urban Rail Transit*, 8(1), 67-98.

Li D., Goverde R.M.P., Daamen W., & He H. (2014) Train dwell time distributions at short stop stations, *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC).*

Medeossi, G., & Nash, A. (2020) Reducing delays on high-density railway lines: London – Shenfield case study. *Transportation Research Record*, 2674(7), 193-205.

Szplett D. & Wirasinghe S. (1984) An investigation of passenger interchange and train standing time at lrt stations*. Journal of advanced transportation,* 18(1), 1-12

Wiggenraad P. (2001). *Alighting and boarding times of passengers at Dutch railways stations*, Trail Delft University

Mots clés

Temps d'échange ; Modèle linéaire généralisé ; Marge opérationnelle ; Données de comptage ; Simulation

Sessions visées

Par ordre décroissant de préférence

1. Session n°SG 5 (Analyse des comportements de mobilité et des activités, pratiques spatiales, pratiques sociales et représentations de la mobilité, impacts des TIC sur les comportements, mobilité partagée, mobilités actives)
2. Session n°SG 2 (Gestion du trafic, systèmes de transport intelligent, management des infrastructures et des réseaux)
3. Session n°ST 18 ()