Quels modèles pour le temps de stationnement des trains en Île de France ?

Rémi Coulaud, Christine Keribin, Gilles Stoltz

Quels modèles pour le temps de stationnement des trains en Île de France ?

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

[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*

Résumé. Nous modélisons le temps de stationnement des trains à l'arrêt à partir de données ferroviaires et de fréquentation issues des trains connectés de la ligne H. Le plan de transport structure l'exploitation ferroviaire ; nous modélisons le temps de stationnement différemment en fonction de l'avance ou du retard du train par rapport à l'heure d'arrivée théorique. Nous montrons que les variables pertinentes dépendent de la place de l'arrêt dans le jalonnement ainsi que de l'heure d'arrivée puis nous comparons les performances de nos modèles avec ceux obtenus par d'autres modèles de la littérature.

Mots-clés. Choix de modèles, données ferroviaires, étude de cas, forêts aléatoires

Abstract. We model train dwell times at train stops with railway data and passengers flow data coming from connected trains of the Île-de-France line H. The timetable constrains railway exploitation, which led us to set up different models of dwell time depending on arrival compared to planed hour. We show that variables selected depends on timetable accordance and train run then we compare our models results to the litterature best models.

Keywords. Models selection, railway data, case study, random forests

1 Introduction

Le temps de stationnement, différence entre l'heure de départ d et l'heure d'arrivée a, d'un train k à un arrêt s représente en Île-de-France en moyenne 30 % du temps de trajet entre une origine et une destination : la capacité d'une ligne ferroviaire et le temps de parcours des voyageurs sont ainsi directement impactés par le temps de stationnement. Sa prévision reste un problème largement ouvert comme en attestent les travaux de Kecman et Goverde (2015), Li et al. (2016), Cornet et al. (2019) et Palmqvist et al. (2019). L'étude des données de temps de stationnement ainsi que les intuitions de Kecman et Goverde (2015) et Pedersen et al. (2018) nous amènent à proposer trois régimes pour la modélisation : les trains en avance, à l'heure ou en retard par rapport à l'heure d'arrivée théorique. Cette segmentation permet d'isoler, pour chaque régime, les principaux facteurs influençant le temps de stationnement, pré-requis à la gestion du temps de stationnement en

opérationnel. Après la présentation des données, nous expliciterons le découpage en trois régimes. Nous sélectionnerons ensuite les variables importantes et nous finirons par la comparaison des performances des modèles induits avec certains de ceux proposés dans la littérature.

2 Présentation du jeu de données

L'étude porte sur les arrêts en zone dense des trains de la ligne H Transilien de septembre 2017 à septembre 2019. Cette ligne étant peu fréquentée, il convient de s'intéresser principalement à des périodes et à des branches où le nombre de trains et de passagers est conséquent : par exemple, les trains qui circulent de Paris-Nord à Pontoise, voir figure 1, en heures de pointe du soir (de 17h à 20h) pendant les jours de semaine hors vacances et jours fériés.

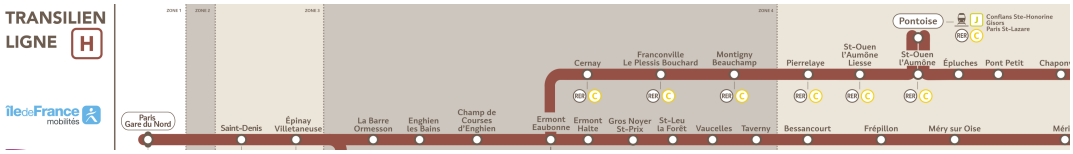

Figure 1 : Plan de la ligne Paris-Nord-Pontoise, passant par Ermont-Eaubonne.

Les données proviennent de deux systèmes embarqués des rames Z50000. Le premier système, ATESS[1], mesure notamment l'heure d'arrivée et de départ d'un train en gare. Le second système, CAVE[2], compte le nombre de montées et de descentes[3] à chaque arrêt à l'aide de caméras infrarouges au-dessus des portes. Ces fichiers donnent aussi des informations sur l'heure de création et d'envoi de l'événement de comptage au moment où le train repart, ces informations permettent de mesurer incidemment le temps de stationnement. Le jeu de données d'étude est constitué par le rapprochement de ces deux bases de données à partir du triplet (train k, gare s, jour t). Le taux de données manquantes (temps de stationnement ou comptage) est de l'ordre de 10 % pour ATESS, 20 % pour CAVE et de 25 % pour le jeu de données d'étude. En les mettant de côté, il reste un nombre total de 21 841 observations pour les onze gares de l'axe Paris-Nord-Pontoise. Nous nous focalisons sur les trois premiers arrêts : Saint-Denis, Epinay-Villetaneuse et Enghien-les-Bains. Le nombre d'observations pour chaque gare est 2 233, 2 243, 2 238. Ce sont trois gares importantes en terme de fréquentation pour la ligne H dont la distribution des montées et des descentes par heure est différente. Les deux premières gares sont des gares avec des double-flux (à chaque heure de la journée il y a autant de montées que de descentes) tandis que la gare d'Enghien-les-Bains est une gare pendulaire (à chaque heure

[1] Acquisition et Traitement des Enregistrements de Sécurité Statique.
[2] Comptage Automatique Voyageur Embarqué.
[3] Le nombre de voyageurs qui montent et qui descendent.

de la journée il y a soit principalement des montées, soit principalement des descentes). Ces différences ont été isolées à partir de méthodes de classification non supervisée.

2.1 Redressement des temps de stationnement

Comme souligné précédemment, nous avons accès à deux sources de mesures des heures d'arrivée et de départ d'un train à un arrêt. Les informations internes à la SNCF ainsi qu'une enquête terrain ont permis de montrer que le temps de stationnement issu d'ATESS était fiable mais il n'est disponible qu'à J+8. Les données issues du système CAVE sont accessibles quasiment en temps réel, mais présentent une différence de mesures des temps de stationnement, que l'on peut raisonnablement représenter par une relation affine, avec une erreur moyenne de l'ordre de 4,5 s. Ceci est acceptable pour une modélisation opérationnelle compte tenu de la précision des outils et des comportements des conducteurs. Nous utilisons, par la suite, les mesures du temps de stationnement d'ATESS.

2.2 Variables explicatives

Les variables susceptibles d'expliquer le temps de stationnement ($Y_{k,s,t}$) pour le train k, l'arrêt s, le jour t, se scindent en deux groupes, les variables de fréquentation utilisées notamment par Cornet et al. (2019) et Palmqvist et al. (2019) et les variables ferroviaires essentiellement utilisées par Kecman et Goverde (2015) et Li et al. (2016). Un des apports de notre travail est d'utiliser à la fois des variables ferroviaires et de fréquentation.

Logiquement, les variables de fréquentation à considérer sont : le nombre de montées ($M_{k,s,t}$), le nombre de descentes ($D_{k,s,t}$) et la charge à bord à l'arrivée ($C_{k,s,t}$).

Pour introduire les variables ferroviaires, nous distinguons les heures d'arrivée et de départ réalisées (a^{real}, d^{real}) des heures d'arrivée et de départ théoriques (a^{theo}, d^{theo}) prévues dans le plan de transport. Ainsi ces variables sont : l'écart à l'heure d'arrivée théorique ($a^{\text{theo}}_{k,s,t} - a^{\text{real}}_{k,s,t}$), l'espacement entre deux trains $d^{\text{real}}_{k-1,s,t} - a^{\text{real}}_{k,s,t}$ ($E_{k,s,t}$).

Suivant les travaux de Chandesris (2014), nous ajoutons le temps de stationnement aux arrêts précédents ($Y_{k,s-1,t}, \ldots, Y_{k,1,t}$) qui apporte une information spatio-temporelle ainsi que le temps de stationnement au même arrêt pour le train précédent ($Y_{k-1,s,t}$). Remarquons que pour l'arrêt Saint-Denis qui est le premier arrêt après l'origine (Paris-Nord) les variables de retard n'existent pas, voir tables 1 et 2.

3 Trois régimes du temps de stationnement

Un train est en avance si $a^{\text{real}} \leq a^{\text{theo}}$, en retard si $a^{\text{real}} \geq d^{\text{theo}}$ et est à l'heure sinon. Cette séparation se justifie car un conducteur d'un train en avance doit normalement attendre l'heure de départ théorique avant de repartir. Inversement, le conducteur d'un train en retard doit normalement repartir immédiatement après que l'échange voyageur est terminé afin de rattraper son retard. Ainsi, un train en avance aura généralement

tendance à stationner plus longtemps, comme constaté sur la figure 2. Pedersen et al. (2018) considèrent les temps de stationnement pour les trains en retard comme des temps de stationnement minimaux. Cependant, il faut être prudent avec cette notion de temps de stationnement minimal car, comme le notent Cornet et al. (2019), rien ne garantit en pratique que le train puisse partir dès que l'échange est terminé ni que le conducteur respecte scrupuleusement la règle en cas de retard. Cette segmentation prend tout son sens pour la gare d'Enghien-les-Bains, voir figure 2, où le temps de stationnement décroît linéairement avec l'écart à l'heure d'arrivée pour les trains en avance. Nous modélisons le temps de stationnement pour ces trois régimes sauf pour les trains en avance à l'arrêt Saint-Denis où il y a moins de 30 observations, d'où la première colonne grisée des tables 1 et 2.

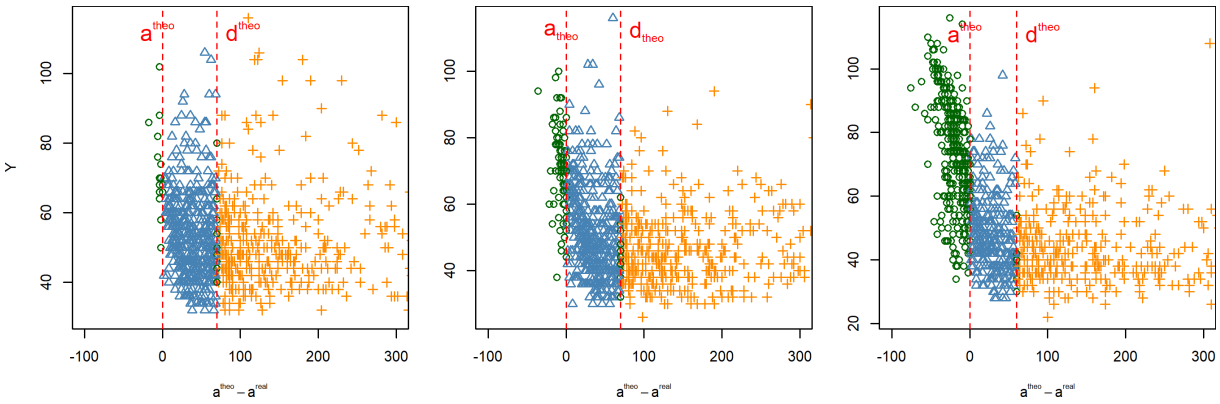

Figure 2 : Temps de stationnement (Y) en fonction de l'écart à l'heure d'arrivée théorique ($a^{\text{theo}} - a^{\text{real}}$) : une décomposition en trois régimes des arrêts en gare de Saint-Denis (à gauche), d'Epinay-Villetaneuse (au centre) et d'Enghien-les-Bains (à droite) ; nous distinguons les trains en avance (○), les trains à l'heure (△) et les trains en retard (+).

4 Modélisation du temps de stationnement

Nous nous intéressons à deux types de modèles, la régression linéaire, modèle paramétrique le plus simple, et les forêts aléatoires. Pour évaluer la qualité de la modélisation et sélectionner les variables intéressantes, nous divisons le jeu de données en : un jeu d'entraînement du 01/09/2017 - 31/08/2018 et un jeu de test du 01/09/2018 - 31/08/2019. La modélisation du temps de stationnement passe par la sélection des variables les plus pertinentes par régime et par gare parmi l'ensemble des variables présentées en section 2.2. Une fois ces variables identifiées par régime, nous comparons les performances de nos modèles à celles de méthodes élémentaires et de méthodes connues dans la littérature.

4.1 Choix de variables en régression linéaire et forêts aléatoires

Nous utilisons une régression linéaire et nous sélectionnons les variables par une recherche *backward* avec le critère BIC, voir Schwarz (1978). Nous utilisons aussi un modèle de forêts aléatoires, voir Breiman (2001), entraîné avec le package `randomForest` (le nombre d'arbres ainsi que le nombre de variables sélectionnées à chaque nœud sont laissés par défaut, 500 et 3 au plus, respectivement). Nous décidons de retenir les variables dont l'importance est supérieure à la moyenne de l'importance de toutes les variables.

	Arrêts en avance			Arrêts à l'heure			Arrêts en retard		
Variables \ Gares	G1	G2	G3	G1	G2	G3	G1	G2	G3
$M_{k,s,t}$				X			X	X	
$D_{k,s,t}$			X	X		X	X		X
$C_{k,s,t}$									
$a^{\text{theo}}_{k,s,t} - a^{\text{real}}_{k,s,t}$		X	X	X	X	X			
$E_{k,s,t}$									
$Y_{k-1,s,t}$			X						
$Y_{k,s-1,t}$			X		X				X
$Y_{k,s-2,t}$			X			X			X

Table 1: Résultats de la sélection de variables par la recherche *backward* avec le critère BIC avec la régression linéaire pour les trois régimes et les trois premières gares du parcours ; les gares de Saint-Denis (G1), Epinay-Villetaneuse (G2) et Enghien-les-Bains (G3).

	Arrêts en avance			Arrêts à l'heure			Arrêts en retard		
Variables \ Gares	G1	G2	G3	G1	G2	G3	G1	G2	G3
$M_{k,s,t}$				X			X	X	X
$D_{k,s,t}$				X		X	X	X	X
$C_{k,s,t}$				X		X	X	X	X
$a^{\text{theo}}_{k,s,t} - a^{\text{real}}_{k,s,t}$		X	X		X	X			
$E_{k,s,t}$		X				X	X	X	
$Y_{k-1,s,t}$									
$Y_{k,s-1,t}$									
$Y_{k,s-2,t}$									X

Table 2: Sélection de variables par l'importance des forêts aléatoires pour trois régimes et trois gares.

Nous remarquons pour les deux modèles que les variables de fréquentation, voir les tables 1 et 2, ne sont que très rarement sélectionnées pour modéliser le temps de sta-

tionnement des trains en avance tandis qu'elles le sont presque systématiquement pour celui des trains en retard. La variable écart à l'heure d'arrivée théorique ($a^{\text{theo}} - a^{\text{real}}$) est sélectionnée uniquement dans le cas où les trains ne sont pas en retard, ceci vient confirmer l'intuition des trois régimes. Nous constatons enfin que les variables de retards ne sont sélectionnées que dans le cas de la régression linéaire.

4.2 Performances des modèles pour les différents régimes

À notre connaissance aucun article ne compare sur un même jeu de données les performances des modèles statistiques existant. Les modèles mis en compétition par régime sont : la régression linéaire et la forêt aléatoire avec les variables sélectionnées dans la section 4.1 par régime ; des modèles de la littérature dont Kecman et al. (2015) qui utilisent des forêts aléatoires avec l'écart à l'heure d'arrivée théorique, Li et al. (2016) qui utilisent une régression linéaire avec $Y_{k-1,s,t}$, $Y_{k,s-1,t}$ et $Y_{k,s-2,t}$ et Palmqvist et al. (2019) qui utilisent une régression linéaire avec toutes les variables de fréquentation ; un modèle élémentaire (la moyenne du temps de stationnement passé pour un train) ; le temps de stationnement théorique prévu par le plan de transport de la SNCF (PdT SNCF). Les moyennes des différences absolues (MAE) et absolues relatives (MAPE) sont calculées pour les différents modèles sur le jeu de données test.

	Trains en avance		Trains à l'heure		Trains en retard	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
Régression linéaire	**11.01**	**0.17**	9.09	**0.18**	**8.57**	**0.19**
Forêt aléatoire	11.42	0.18	**9.05**	**0.18**	8.86	0.20
Kecman et al.	11.42	0.18	9.44	0.19	9.74	0.22
Li et al.	13.73	0.21	9.21	**0.18**	9.15	0.20
Palmqvist et al.	15.36	0.22	9.56	0.19	8.72	**0.19**
Moyenne	31.38	0.38	9.94	0.19	9.35	0.21
PdT SNCF	19.78	0.25	14.35	0.33	17.71	0.46

Table 3: Performances des modèles sur le jeu de données test pour la gare d'Enghien-les-Bains.

Nos modèles font mieux ou aussi bien que les modèles de la littérature. Pour les trains en avance la différence de performances entre le modèle élémentaire et nos modèles est importante. Pour les trains en retard l'utilisation de variables explicatives ne permet pas une réelle amélioration des performances par rapport au modèle élémentaire.

4.3 Perspectives

Au vu des résultats précédents, il conviendra de prendre en compte de façon plus fine les dépendances spatiales et temporelles des temps de stationnement afin d'améliorer leur

estimation pour les trains en retard. Cette étape de modélisation focalisée sur le temps de stationnement nous permettra de passer plus facilement de la modélisation à la prévision.

Les trains connectés comptent non seulement le nombre de montées et de descentes à l'échelle du train mais aussi à l'échelle de la porte ce qui nous permettra de quantifier l'influence de la répartition à quai des échanges sur le temps de stationnement.

Enfin, la modélisation suppose connues les montées et descentes lors du temps de stationnement. C'est une première étape, qui devra s'enrichir par la prévision de ces informations.

Bibliographie

Breiman, L. (2001), *Random forests*, Machine learning, 45(1), 5-32.

Chandesris, M. (2014), *Approche non-paramétrique pour la prédiction d'heure d'arrivée dans les transports*, 46 ème journées de Statistique de la SFDS, Rennes.

Cornet, S., Buisson, C., Ramond, F., Bouvarel, P. et Rodriguez, J. (2019), *Methods for quantitative assessment of passenger flow influence on train dwell time in dense traffic areas*, Transportation Research Part C : Emerging Technologies, vol. 106, pp. 345-359

Kecman, P. et Goverde, R. M. P. (2015), *Predictive modelling of running and dwell times in railway traffic*, Public Transport 7, 295–319.

Li, D., Daamen, W. et Goverde, R. M. P. (2016), *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, 877–896.

Palmqvist, C-W., Tomii, N., et Ochiai, Y. (2019), *Dwell Time Delays for Commuter Trains in Stockholm and Tokyo*, Paper presented at RailNorrköping, Norrköping, Suède.

Pedersen, T., Nygreen, T. et Lindfeldt, A. (2018), *Analysis of temporal factors influencing minimum dwell time distributions*, WIT Transactions on The Built Environment. 181, 447–458

Schwarz, G. (1978), *Estimating the dimension of a model*, The annals of statistics, 6(2), 461-464.