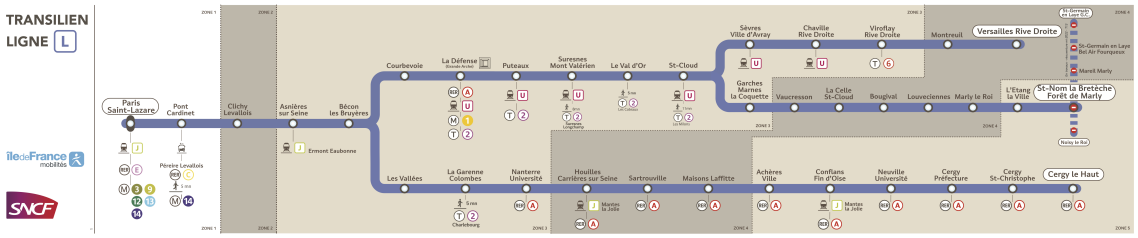
Estimation de matrices Origine-Destination

Application à la ligne L du réseau Transilien



Lab' Mass Transit Academy
SNCF - Transilien
Auteur : Joshua Wolff
Supervision : Rémi Coulaud, Mathilde Vimont
Août 2021

Table des matières

Notation	Nom	Domaine	Description
G	Nombre de gares	$\mathbb{N}^*$	Nombre de gares sur la ligne étudiée
T	Seuil temporel	$\mathbb{R}_+^*$	Seuil temporel arbitraire limitant la durée d'un trajet entre deux gares sur la ligne considérée (60-90 minutes)
M	Nombre de coefficients OD	$\mathbb{N}$	Nombre de coefficients dans les matrices OD estimées
k	Indice voyageur	$\mathbb{N}^*$	Indice désignant un voyageur unique qui a validé au moins une fois sur la ligne considérée le jour considéré
N_k	Nombre de validations de k	$\mathbb{N}^*$	Nombre de validations de l'utilisateur k sur la ligne considérée pour le jour considéré
g_i^k	Gare de validation de k	$-$	Gare où a eu lieu la i-ème validation de la journée de k sur la ligne considérée
h_i^k	Heure de validation de k	$-$	Heure à laquelle k a validé pour la i-ème fois de la journée sur la ligne considérée
t_i^k	Créneau de validation de k	$-$	Tranche temporelle pendant laquelle k a validé pour la i-ème fois de la journée sur la ligne considérée
v_i^k	Type de validation de k	$-$	Type de la i-ème validation de la journée de k sur la ligne considérée (entrée = 1 ou sortie = 2)
n_{ij}^t	Coefficient OD	$\mathbb{R}_+$	Nombre de passagers allant de la gare i à la gare j avec un départ dans la tranche temporelle t
B_i^t	Volume d'entrants	$\mathbb{N}$	Nombre de validations d'entrée en gare i dans la tranche temporelle t
r_i	Ratio gare-ligne	$[0, 1]$	Proportion moyenne des voyageurs allant sur la ligne considérée en gare i
$\sim$		-	Symbole utilisé pour designer les grandeurs de référence
$\bullet$		-	Symbole remplaçant un indice pour désigner la sommation sur cet indice

TABLE 1 – Table des notations

1 Introduction

Les matrices Origine-Destination caractérisent les déplacements au sein d'un réseau de transport aussi bien aérien, routier que ferré. Elles sont à distinguer des comptages (nombre de validations, nombre de vehicules) qui permettent d'estimer un volume sans mesurer ni le sens de circulation ni la distance parcourue. Dans notre cas, les matrices Origine-Destination (OD) consistent à établir le nombre de passagers (ou la proportion de passagers) allant d'une gare à une autre du réseau.

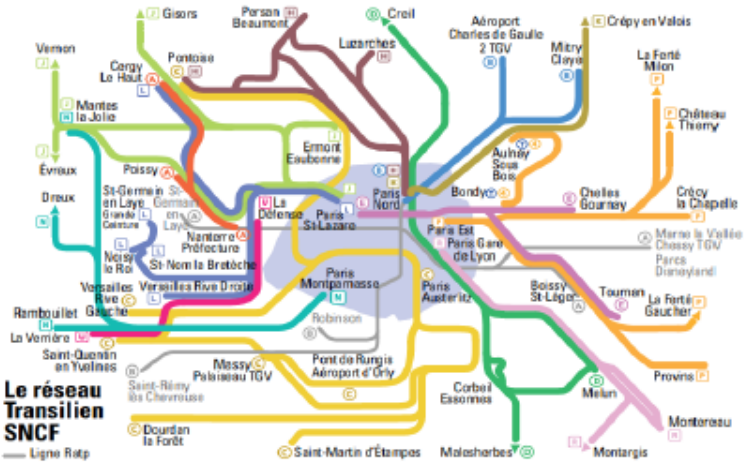


FIGURE 1 – Plan du réseau Transilien (SNCF)

Transilien et IdFM cherchent tous les deux à mieux comprendre les habitudes de déplacement des Franciliens pour mieux adapter la desserte des trains. Ces matrices OD sont aussi utilisées a posteriori pour calculer le retard voyageur, à partir de l'écart entre le temps de trajet théorique et celui réalisé : il s'agit de connaître la quantité de personnes ayant effectué ce trajet.
Aujourd'hui, les matrices OD sont issues d'enquêtes massives auprès des clients effectuées tous les 4 ans. Cette information pourrait être améliorée en utilisant une autre source d'information : la télé-billetique. En effet, cette dernière permettra de plus en plus de connaître où le voyageur a validé en entrée et où il a validé en sortie. Ce mouvement sera amplifié par l'introduction du pay as you go qui consiste à faire payer le voyageur en fonction du trajet qu'il a effectué. Ce système est utilisé à Londres à travers l'Oyster Card où les voyageurs sont obligés de valider en sortie pour terminer leur trajet. S'ils ne le font pas, ils payent le trajet le plus cher.

Dans ce travail, on définit un *trajet* comme le transport d'un voyageur sans correspondances (nécessairement unimodal), et un *déplacement* comme une succession de trajets dans un laps de temps donné, éventuellement multimodal. La première question que l'on se pose est technique : comment reconstruire des trajets à l'aide de la télé-billettique ? Une littérature assez riche existe à ce sujet et IdFM dispose d'un algorithme qui permet

de faire cette reconstruction. La deuxième question que l'on se pose est : dans quelle mesure ces trajets reconstruits peuvent nous être utiles à la fois pour optimiser l'exploitation ferroviaire et la qualité de service ?

Les cas d'usages auxquels on souhaite répondre sont les suivants :

— Reconstruction a posteriori de la charge à bord des trains sans dispositifs de comptage automatique en croisant les informations d'OD et le nombre de validations avec le plan de transport. La reconstruction en temps réel est possible uniquement si les validations sont accessibles en temps réel.
— Optimisation du plan de transport pour prendre en compte les fluctuations de la demande à des échelles de temps fines.
— Identification des événements de saturation au niveau d'OD ou de points de correspondance.

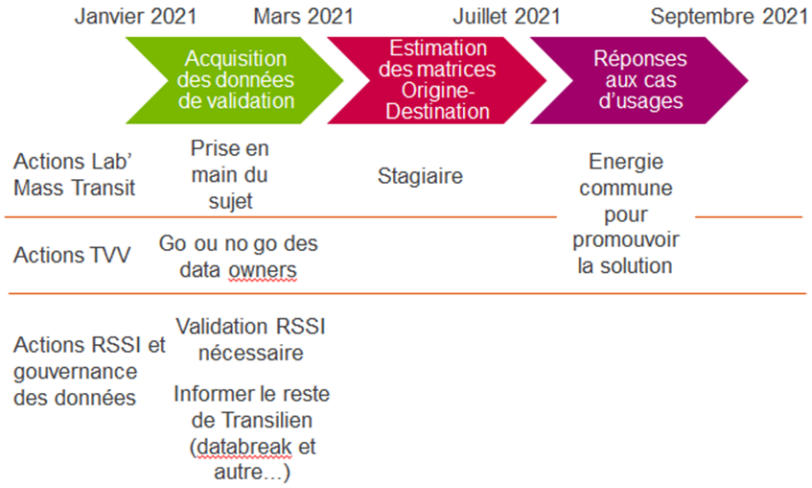


FIGURE 2 – Les différents jalons du projet OD

Notons qu'il n'est a priori pas question de faire de la prédiction, mais plutôt d'estimer à postériori ces matrices. Nous choisissons par ailleurs de nous focaliser à l'échelle d'une ligne car nous cherchons *in fine* à allouer des passagers à des trains pour en estimer la charge à bord.

On dispose de données massives de télé-billetique pour répondre aux cas d'usages précédemment cités. La littérature traitant de l'estimation de matrices OD fait état de trois grandes catégories de modèles : les modèles probabilistes, les modèles de poursuite d'utilisateur, ainsi que les modèles de mise à jour de matrices OD. La première catégorie de modèles a pour intérêt principal de fournir un cadre mathématique solide (maximisation d'entropie / de vraisemblance, chaînes de Markov, filtres de Kalman ...) tandis que la seconde englobe des modèles davantage orientés "métier", avec des règles facilement interprétables. On ne s'attardera pas sur les modèles de mise à jour de matrices OD (aussi

appelés *growth factors models*) car ils supposent une connaissance d'une matrice OD initiale, là où notre intérêt se porte davantage sur de la modélisation à partir des données de télé-billetique uniquement. Une expérimentation de ce type a cependant déjà été faite chez Transilien dans le cadre du projet IVA [IVA, 2020] qui s'appuie sur les données de comptage automatique des passagers. Par ailleurs, Florian Toqué se penche dans sa thèse soutenue en 2019 sur l'utilisation des réseaux de neurones pour l'estimation de matrices OD avec des applications au réseau de transport de Rennes ainsi qu'à une partie du réseau Transilien [Toque, 2019]. Ces modèles sont cependant limités en terme d'interprétabilité. Souhaitant avoir des modèles explicables, on s'intéresse particulièrement aux deux classes de modèles restantes.
Le modèle le plus naturel consiste à suivre (quand c'est possible) le parcours de chaque utilisateur grâce aux différentes validations pendant son parcours. Une approche de ce type est décrite et appliquée à un réseau de bus dans Trépanier and Tranchant [2007], et est facilement généralisable à notre cas d'étude. Des modèles probabilistes liés à la maximisation d'entropie / vraisemblance, souvent appelés "modèles gravitationnels", sont décrits dans Wilson [1969]. Une application à un réseau de transport ferré de ce type de modèle est proposée dans Ait Ali and Eliasson [2019], permettant ainsi de proposer une alternative aux modèles de poursuite d'utilisateur. L'affectation gravitationnelle comme la poursuite d'utilisateur présentent l'avantage de ne se baser que sur les données télé-billetiques pour estimer les coefficients des matrices Origine-Destination. Les modèles basés sur des concepts théoriques plus subtils sont souvent appliqués à des réseaux "jouets" et impliquent un formalisme relativement lourd. On peut néanmoins se référer à Abareshi et al. [2019] (chaînes de Markov) et Cho et al. [2008] (filtres de Kalman) si l'on souhaite se pencher sur cette classe de modèles.

Ce document se découpera de la manière suivante : nous présentons dans la section 2 les différents modèles d'estimation de matrices OD qui n'utilisent que les données de télébillétique, la section 3 est l'occasion de présenter plus en détail les données à disposition, la référence et le périmètre d'étude pour l'obtention des résultats, la section 4 est dédiée à la présentation des résultats.

2 Modélisation

Cette section présente trois modèles visant à estimer a posteriori des matrices OD à l'échelle d'une ligne à partir des données de télé-billetique. Notons néanmoins qu'une fois ces matrices estimées, il est possible de les appliquer en temps réel aux volumes de validation.

Le premier modèle repose essentiellement sur la notion de poursuite d'utilisateur. Certaines règles ont été supprimées / ajoutées par rapport à l'article original [Trépanier and Tranchant, 2007] pour s'adapter à notre cas d'étude. Le second modèle est partiellement issu de Ait Ali and Eliasson [2019]. Il s'agit d'un modèle d'affectation gravitationnelle, dont les fondements reposent sur un principe de maximisation d'entropie qui est justifié dans Ait Ali and Eliasson [2019] et Xie et al. [2011]. Le troisième et dernier modèle est quant à lui une combinaison des deux approches précédentes.

La table 1 présente les différentes notations employées dans la déscription des modèles

2.1 Poursuite d'utilisateur

On définit la i-ème validation de l'utilisateur k comme étant le quadruplet $(g_i^k, h_i^k, t_i^k, v_i^k)$. Rappelons qu'il s'agit uniquement des validations effectuées dans des gares traversées par la ligne considérée. On parle de suite de validations d'un utilisateur k pour désigner $((g_i^k, h_i^k, t_i^k, v_i^k))_{1 \leq i \leq N_k}$.

Dans l'imaginaire d'un réseau férré parfaitement hermétique (validation en gare d'origine (en entrée) et validation en gare de destination (en sortie)), il suffirait de "lire" la suite de validations correspondant à un utilisateur pour en déduire chacun de ses trajets de la journée. Il serait de ce fait évident d'en déduire des matrices Origine-Destination. L'un des défis majeurs est de formuler des règles de reconstruction des trajets dans le cas où certaines étapes sont manquantes. Il est par exemple possible d'observer une succession de données de validation en entrée, sans qu'il y ait de données de validation en sortie, car la majorité des gares ne sont pas équipées de contrôle automatique de billet (CAB) en sortie de gare. Il n'y a en effet aucune obligation pour le moment de valider en sortie sur le réseau Transilien à part si il y a des lignes de CAB imperméables. La procédure de poursuite d'utilisateur est la suivante :

Algorithme 1 : Poursuite d'utilisateur

pour chaque k **faire**
  si $N_k > 1$ **alors**
    pour chaque $i \in [\![1, N_k - 1]\!]$ **faire**
      si $(v_i^k, v_{i+1}^k) = (1,2)$ ***ET*** $g_i^k \neq g_{i+1}^k$ ***ET*** $h_{i+1}^k - h_i^k \leq T$ **alors**
        # motif "entrée-sortie"
        $n_{g_i^k g_{i+1}^k}^{t_i^k} \leftarrow n_{g_i^k g_{i+1}^k}^{t_i^k} + 1$
      fin
      si $(v_i^k, v_{i+1}^k) = (1,1)$ ***ET*** $g_i^k \neq g_{i+1}^k$ **alors**
        # motif "entrée-entrée"
        $n_{g_i^k g_{i+1}^k}^{t_i^k} \leftarrow n_{g_i^k g_{i+1}^k}^{t_i^k} + 1$
      fin
    fin
    si $(v_1^k, v_{N_k}^k) = (1,1)$ ***ET*** $g_1^k \neq g_{N_k}^k$ **alors**
      # motif "retour au domicile"
      $n_{g_{N_k}^k g_1^k}^{t_{N_k}^k} \leftarrow n_{g_{N_k}^k g_1^k}^{t_{N_k}^k} + 1$
    fin
  fin
fin

Cette méthode se résume donc en deux points principaux :

- soit on constate que l'utilisateur entre dans une gare puis sort dans une autre gare après un temps inférieur à T (motif *entrée-sortie*) ;
- soit qu'il est entré dans une gare puis entré dans une gare différente sans qu'il y ait de sortie entre ces deux validations (motif *entrée-entrée*).

Ce second cas correspond à un premier trajet dont la gare d'arrivée n'est pas équipée par des CAB en sortie, il faut donc "attendre" dans l'espoir que le voyageur fasse sa validation suivante dans la gare par laquelle il est précédemment sortie. Les deux schémas suivant illustrent les motifs *entrée-entrée* et *entrée-sortie*.

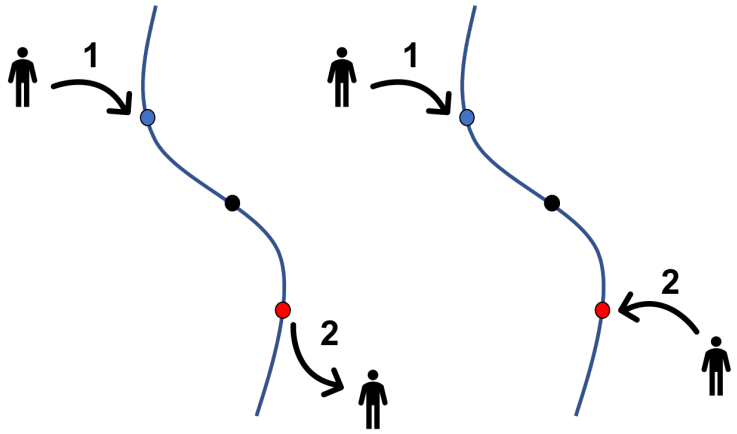


FIGURE 3 – Motif *entrée-sortie* (gauche) et motif *entrée-entrée* (droite)

Dans le cas où la dernière validation (sur la ligne considérée) de la journée d'un utilisateur est une validation en entrée, nous identifions un motif dit de *retour au domicile*. Si la première validation était une entrée dans une gare différente de la gare où a eu lieu la dernière validation, nous en déduisons que le voyageur "rentre" de la dernière gare de validation vers la première gare de validation.

Le principal défaut de la poursuite d'utilisateur est de passer sous silence certaines successions de validations qui ne rentrent pas dans les motifs précédemment cités. Par exemple, une validation en entrée unique sur une journée ne sera pas comptabilisée par notre algorithme. Bien sûr, si la gare de validation est traversée par une autre ligne, il est possible que le voyageur ait simplement emprunté cette dernière, mais le doute reste entier.

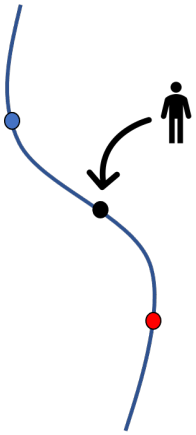


FIGURE 4 – Exemple de motif de validation non interprétable

On s'intéresse maintenant au modèle probabiliste d'affectation gravitationnelle, qui présente l'avantage de prendre en compte toute les validations indiquant qu'un voyageur entre sur la ligne considérée.

2.2 Affectation gravitationnelle

Ce modèle est issu de l'article de Ait Ali and Eliasson [2019], lui même basé sur celui de Wilson [1969], qui propose un modèle visant à estimer des matrices OD à l'échelle d'une ligne à partir d'un problème de "maximisation d'entropie" à coût linéaire. On se dispense ici de présenter tout le contexte lié à la notion d'entropie et à la résolution pratique du problème, tout étant décrit dans [Ait Ali and Eliasson, 2019]. L'objectif de ce paragraphe est d'interpréter la solution de ce problème d'optimisation et d'en déceler les limites.

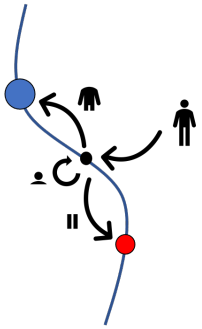


FIGURE 5 – Schéma d'une affectation gravitationnelle

Le schéma ci-dessus explique l'effet produit par l'affectation gravitationnelle : sachant qu'un utilisateur est rentré dans une gare de la ligne considérée, il est partiellement affecté à toutes les autres gares de la ligne en fonction de leurs attractivités respectives. La question principale est donc de savoir comment estimer cette notion d'*attractivité* afin d'en déduire des matrices OD. La formulation théorique du problème est la suivante :

$$\min_{n_{ij}^t} \sum_{i,j\neq i,t} n_{ij}^t \log(n_{ij}^t) - n_{ij}^t$$

$$\text{s.c. } (1) \sum_{j\neq i} n_{ij}^t = r_i B_i^t \text{ et } (2) \sum_{i\neq j,t} n_{ij}^t = \sum_t r_j B_j^t$$

Remarquons qu'il s'agit bien d'une minimisation et non d'une maximisation contrairement à ce qui est noté dans [Ait Ali and Eliasson, 2019] (il s'agit d'une faute de frappe). L'article de Xie et al. [2011] reprend quant à lui la bonne notation. La résolution de ce problème d'optimisation revient à déterminer quelle est la façon "la moins arbitraire" de déplacer les voyageurs sur la ligne compte tenu des observations des données télébilletiques. La contrainte (1) correspond à une contrainte de conservation du flux de voyageurs sortant de la gare i pendant la tranche temporelle t et la contrainte (2) correspond à une hypothèse de symétrie : sur une journée complète le nombre de montants dans une gare est égale au nombre de descendants dans cette même gare.

Un ajout fait dans notre modèle par rapport au modèle original est l'introduction des ratios gare-ligne r_i. Ils permettent de prendre en compte le cas des gares traversées par différentes lignes du réseau Transilien. En effet, un utilisateur validant en entrée dans une

gare *multiligne* (*ie* qui est traversée par plusieurs lignes) peut tout à fait se rendre sur une autre ligne que celle que nous considérons. Ces ratios indiquent donc la proportion moyenne de voyageurs empruntant la ligne considérée pour chaque gare de cette ligne. Notons que l'estimation de ces ratios n'est absolument pas abordée dans cette étude, ils sont considérées comme étant des données du problème.

On obtient la solution en introduisant des multiplicateurs de Lagrange (encore une fois, voir Ait Ali and Eliasson [2019]) :

$$n_{ij}^t = \begin{cases} r_i B_i^t \frac{\sum_t r_j B_j^t}{\sum_{i' \neq i} r_{i'} B_{i'}^t} = r_i B_i^t \tau_{ij} & \text{si } i \neq j \\ 0 & \text{sinon.} \end{cases}$$

Le modèle obtenu coïncide avec le formalisme des modèles gravitationnels dans la littérature du transport, de la logistique et de la dynamique des populations (voir Wilson [1969]). Il faut interpréter le terme $\tau_{ij} = \frac{\sum_t r_j B_j^t}{\sum_{i' \neq i} r_{i'} B_{i'}^t}$ comme le taux d'attractivité sur la ligne considérée de la gare j sachant que l'on est en partance de la gare i. Notons qu'il n'y a pas de passagers faisant des trajets de la gare i vers la gare i, τ_{ii} est donc systématiquement nul, ce qui entraîne la nullité des termes n_{ii}^t. La limite principale de ce modèle est que ce taux ne dépend pas du temps alors même que l'attractivité des gares varie au cours de la journée (on pourrait penser à une gare comme la Défense Grande Arche, très attractive le matin et beaucoup moins le soir).

$$\underbrace{r_i B_i^t}_{\text{dépendant du temps}} \quad \times \quad \underbrace{\tau_{ij}}_{\text{indépendant du temps}}$$

En définitive, ce modèle est intéressant pour plusieurs raisons :

- il possède un fond théorique solide ;
- il demande très peu de calculs du point de vue numérique ;
- il est facilement interprétable comme un ensemble de taux d'attractivité.

Cependant, le fait que les taux d'attractivité soient constants quelle que soit l'heure de la journée est un défaut majeur de ce modèle qui entraînera sûrement des erreurs conséquentes.

2.3 Poursuite d'utilisateur étendue

Ce troisième et dernier modèle n'amène rien de nouveau par rapport aux parties précédentes : il en fait la synthèse. La poursuite d'utilisateur, a priori sans faille dans l'imaginaire d'un réseau parfaitement hermétique, possède en pratique un défaut par son incapacité à prendre en compte les validations différentes des motifs connus. L'idée de la poursuite d'utilisateur étendue est d'utiliser l'affectation gravitationnelle lorsque l'on rencontre un motif dans les validations qui n'est pas interprétable par la poursuite d'utilisateur.

Dans ce dernier modèle, une règle supplémentaire est employée. Il s'agit d'un garde fou pour éviter d'utiliser l'affectation gravitionnelle dans des cas inopportuns. Imaginons qu'une validation unique, en entrée, ait lieu dans une gare multiligne de la ligne considérée. Nous sommes donc dans un cas non interprétable au sens de la poursuite d'utilisateur : on ne sait a priori pas où est allé le voyageur. Il serait bienvenu d'utiliser l'affectation gravitationnelle pour répartir notre utilisateur sur la ligne. Mais avant de faire cette opération, nous allons vérifier que notre voyageur n'effectue pas une validation sur une des autres lignes passant par notre gare multiligne. Si tel est le cas, nous n'affectons pas le voyageur sur la ligne considérée, puisqu'il a en réalité emprunté une des autres lignes passant par sa gare d'entrée.

L'approche consiste donc à faire de la poursuite d'utilisateur (section 2.1.) secondée d'une phase d'affectation gravitationnelle (section 2.2.) lorsque un motif non interprétable est rencontré (sous réserve de passer le garde fou des gares multilignes).

3 Données et périmètre d'étude

On présente dans cette section la nature de la source de données employée dans nos modèles ainsi que le périmètre d'étude sur lequel nous avons obtenu nos résultats.

3.1 Données de validation (AFC)

Ces données correspondent aux informations issues de la validation des voyageurs sur les bornes / portiques du réseau Transilien. Les champs utilisés dans notre étude sont :

- L'identifiant utilisateur
- L'heure de validation
- L'identifiant de l'équipement de validation (associé à une gare)
- Le type de validation (entrée / sortie)

À partir de ces éléments nous pouvons reconstituer les suites de validation pour une journée $((g_i^k, h_i^k, t_i^k, v_i^k))_{1 \leq i \leq N_k}$ pour chaque utilisateur k, ainsi que calculer les volumes d'entrants B_i^t. Il est alors possible de d'estimer des matrices OD en utilisant nos modèles. Les données de validation sont disponibles à J+2 voir J+3, ce qui permet en pratique d'évaluer les modèles régulièrement.

FIGURE 6 – CAB-MT Transilien en gare de Saint-Lazare

Il est important de souligner qu'à l'échelle du réseau Transilien, seulement 25% des gares sont fermées en sortie mais qu'en pratique près de 40% des validations sont des validations en sortie contre 50% dans un réseau parfaitement hermétique. Cela est dû au fait que les gares les plus importantes, avec les plus gros volumes de voyageurs, sont souvent fermées en sortie.
Notons par ailleurs que certaines gares du réseau Transilien ne sont actuellement pas fermées en entrée mais disposent de bornes de validation. Bien qu'un utilisateur en règle a tout intérêt à valider sur une borne avant de monter dans son train, 8% à 40% des voyageurs dans ce type de gare ne le font pas.

3.2 Périmètre d'étude

Les données sur lesquelles nous avons testé nos modèles correspondent à des validations sur la ligne L du réseau Transilien pendant la journée du mardi 5 novembre 2019. Notre choix s'est porté sur cette ligne car elle possède de nombreuses gares en correspondance avec d'autres lignes Transilien (A, J, U, C) et RATP, ce qui représente un des défis majeurs pour nos modèles.

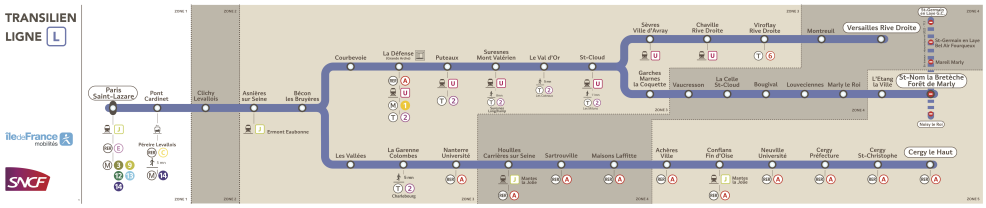


FIGURE 7 – Plan de la ligne L

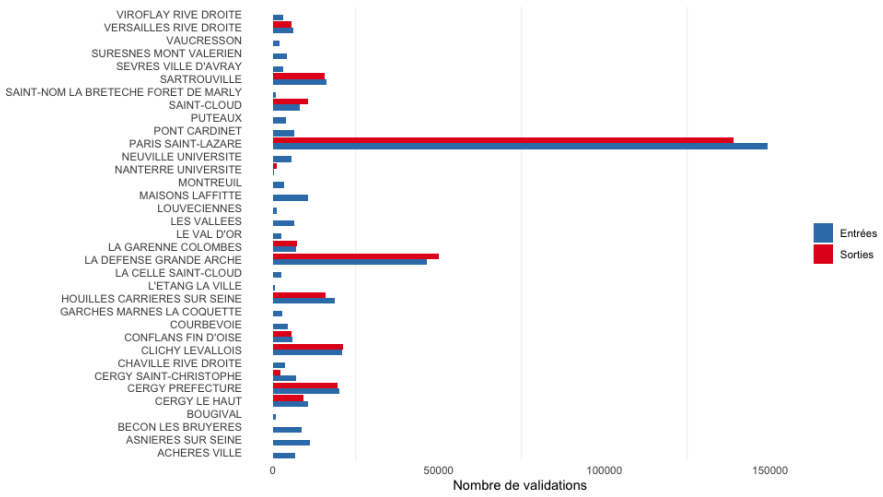


FIGURE 8 – Nombre d'entrées / sorties par gare (ligne L, 05/11/2019)

On observe sur la figure 8 que pour les gares avec des CAB en sortie le nombre de validations en entrée est environ égal au nombre de validations en sortie. Ceci conforte l'idée que sur une journée complète, le nombre d'entrants dans une gare est environ égal au nombre de sortants de cette gare. Notons par ailleurs que seulement 37% des gares sont fermées en sortie sur la ligne L, mais que 42% des validations sont des validations en sortie pour la journée étudiée (contre 50% si toutes les gares étaient équipées de CAB en sortie). Cela indique que la majeure partie des voyageurs passent par des gares fermées en sortie.

3.3 Ratios gare-ligne

L'obtention des ratios gare-ligne n'est pas spécifiquement étudiée dans ce document. Comme expliqué dans la section portant sur l'affectation gravitationnelle, les ratios gare-ligne correspondent à la proportion moyenne des voyageurs empruntant la ligne considérée pour chaque gare de cette ligne. Notons cependant que ces ratios peuvent être estimés à partir d'enquêtes terrain et/ou en couplant les données de validation et les données de comptage automatique de montants / descendants (données CAVE) pour les lignes équipées par ce type de dispositifs. Pour notre part, nous avons utilisé des ratios établis par Jean Lagrange qui sont obtenus à partir des comptages manuels et automatiques.
Ces ratios sont **uniquement** utilisés dans les modèles d'affectation gravitionnelle et de poursuite d'utilisateur étendue. En effet le modèle de poursuite d'utilisateur se base seulement sur les données de validation.

3.4 Matrices OD de référence

Deux références sont utilisées pour pour comparer nos modèles : d'une part les **OD manuelles** de Transilien et d'autre part les **OD IdFM** fournies à Transilien par Île-de-France Mobilités via le programme SIDV.

Les OD manuelles présentent l'avantage d'être une référence terrain, basées sur des enquêtes statistiques fiables et sont disponibles sur tout le périmètre Transilien. Cette référence est mise à jour tous les 4 ans uniquement du fait de la complexité logistique et du coût de ces enquêtes terrain. Ces matrices OD sont par ailleurs seulement disponibles avec un pas de temps d'une heure.

Les OD IdFM sont quant à elles estimées à partir des données de validation détenues par Île-de-France Mobilités sur les réseaux Transilien et RATP. La méthode d'obtention de cette référence est très similaire au modèle de poursuite d'utilisateur. Notons qu'il s'agit d'OD **multimodales** à l'échelle **réseau** tandis que les matrices OD que nous estimons sont **unimodales** (trains Transilien) et à l'échelle d'une **ligne**. Les OD IdFM sont disponibles à des pas de temps variables, ce qui nous permettra de les comparer à nos estimations à des échelles temporelles fines.

Dans les deux cas ces références ne sont pas absolues, mais permettent de constater si les modèles proposés sont cohérents avec les références actuellement employées au sein de Transilien.

4 Résultats

4.1 Métriques

On introduit dans cette partie deux métriques afin de comparer quantitativement nos estimations aux deux références précédemment décrites. Les notations utilisées sont décrites dans la table 1.

Métrique	En volume	En proportion
MaxAE (Maximum Absolute Error)	$\max_{i,j,t} \lvert\tilde{n}^t_{ij} - \frac{\tilde{n}^{\bullet\bullet\bullet}_{\bullet\bullet\bullet}}{n^{\bullet\bullet\bullet}_{\bullet\bullet\bullet}} n^t_{ij}\rvert$	
MeanAE (Mean Absolute Error)	$\frac{1}{M}\sum_{i,j,t} \lvert\tilde{n}^t_{ij} - \frac{\tilde{n}^{\bullet\bullet\bullet}_{\bullet\bullet\bullet}}{n^{\bullet\bullet\bullet}_{\bullet\bullet\bullet}} n^t_{ij}\rvert$	$\frac{1}{M}\sum_{i,j,t} \lvert\frac{\tilde{n}^t_{ij}}{\tilde{n}^t_{i\bullet}} - \frac{n^t_{ij}}{n^t_{i\bullet}}\rvert$

Table 2 – Table des métriques

La métrique MaxAE quantifie à quel point l'estimation est différente de la référence dans le pire des cas. La métrique MeanAE permet de quantifier la différence moyenne avec la référence. Notons que l'on se compare aux références aussi bien en volume qu'en proportion de passagers. Il n'est cependant pas pertinent d'observer l'erreur maximale en proportion car dans le cas des gares avec peu de validations, une faible erreur en volume peut facilement entraîner une grande erreur en proportion.

4.2 Comparaison quantitative des modèles

Les tables ci-dessous récapitulent les différences entre les références et nos estimations. Le terme "Couplage" désigne la poursuite d'utilisateur étendue, car elle couple la méthode de poursuite d'utilisateur avec celle d'affectation gravitationnelle.

Métrique	Aff. grav.	Poursuite	Couplage	OD IdFM
MaxAE (volume)	3382	**1329**	1746	1587
MeanAE (volume)	12.2 (0.4)	**9.8 (0.3)**	9.9 (0.3)	10.5 (0.3)
MeanAE (proportion)	3.2%	**2.3%**	2.4%	2.4%

TABLE 3 – Table de comparaison des modèles - Référence OD manuelles

Métrique	Aff. grav.	Poursuite	Couplage	OD Manuelles
MaxAE (volume)	1991	**821**	867	1145
MeanAE (volume)	8.6 (0.3)	**4.1 (0.1)**	4.9 (0.2)	7.6 (0.2)
MeanAE (proportion)	3.1%	**1.5%**	1.7%	2.4%

TABLE 4 – Table de comparaison des modèles - Référence OD IdFM

On remarque que le modèle de poursuite d'utilisateur est systématiquement plus proche de la référence, aussi bien pour les OD manuelles que pour les OD IdFM. Ces résultats sont à nuancer avec les métriques MeanAE en volume et proportion, où les modèles de poursuite d'utilisateur étendue et de poursuite d'utilisateur sont très proches. Rappelons par ailleurs que ces résultats sont obtenus sur une seule journée. Le modèle d'affectation gravitationnelle est quant à lui nettement différent des références relativement aux deux autres modèles proposés.

4.3 Quelques observations qualitatives

Compte tenu des résultats précédents, nous allons nous pencher plus spécifiquement sur le modèle de poursuite d'utilisateur et observer certains coefficients OD au cours de la journée étudiée pour mieux comprendre l'origine des différences observées.

La figure 9 ci-dessous montre la différence entre les coefficients en proportion estimés par poursuite d'utilisateur et ceux issus des OD manuelles. Elle a été moyennée sur les différentes tranches horaires de la journée.

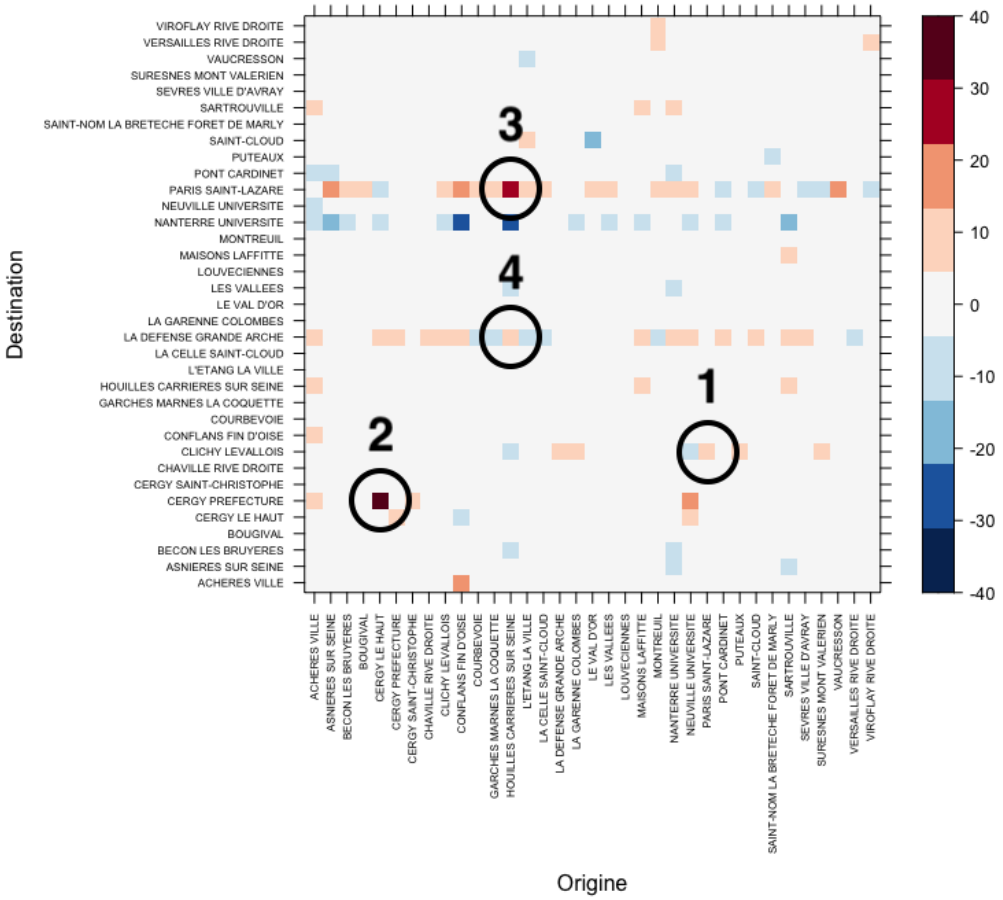


FIGURE 9 – Différence entre matrice OD poursuite d'utilisateur et matrice OD manuelles, en proportion moyenne sur une journée (échelle en %)

On observe une tendance du modèle de poursuite d'utilisateur à surestimer la proportion de personne allant vers Paris Saint-Lazare et La Défense Grande Grande Arche par rapport aux OD manuelles. On constate aussi une légère tendance à sous-estimer la proportion de voyageurs à destination de Nanterre Université. On s'intéresse plus particulièrement aux quatre coefficients entourés qui correspondent aux couples Origine → Destination suivants :

1. **Paris Saint-Lazare → Clichy-Levallois** : paire OD avec le plus gros volume journalier ;
2. **Cergy le Haut → Cergy Préfecture** : paire OD avec le plus gros écart en proportion par rapport aux OD manuelles ;
3. **Houilles Carrières sur Seine → Paris Saint-Lazare** : paire OD avec un écart important en proportion et en volume par rapport aux deux références ;

4. **Houilles Carrières sur Seine → La Défense Grande Arche** : paire OD avec le plus gros écart en volume par rapport aux OD IdFM.

Ces quatre coefficient particuliers sont étudiés plus en détail dans les quatre sous-sections suivantes.

4.3.1 Paris Saint-Lazare - Clichy-Levallois

La figure 10 ci-dessous montre l'évolution temporelle du coefficient correspondant au trajet entre Paris Saint-Lazare et Clichy-Levallois :

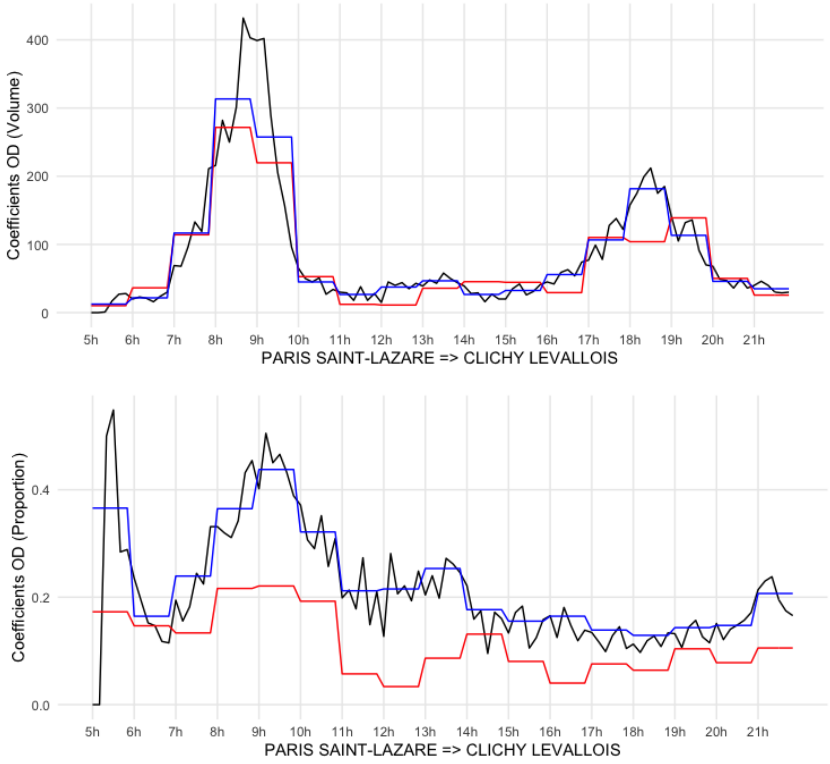


FIGURE 10 – En volume (Haut), en proportion (Bas)
OD poursuite en **noir**, OD manuelle en **rouge**, OD IdFM en **bleu**
Pas de temps de 10 minutes pour poursuite, 1 heure pour les autres

On remarque bien sur le graphe en volume l'intérêt d'avoir un pas de temps plus fin. En effet, avec un pas de temps d'une heure, on estime un volume quasiment constant entre 8h et 10h. L'utilisation d'un pas de temps de 10 minutes permet d'observer que les volumes de voyageurs varient en fait du simple au double sur cette plage horaire. Remarquons que les OD IdFM permettent aussi d'avoir un pas de temps fin, il s'agit simplement d'un choix de notre part de les représenter avec un pas de temps d'une heure pour les comparer aux OD manuelles.

Ce couple Origine-Destination est particulièrement intéressant car la gare de Clichy-Levallois est équipée avec des CAB en sortie et aucune autre ligne ne permet d'effectuer ce trajet directement. La méthode de poursuite d'utilisateur est donc censée donner des résultats très réalistes pour cette paire OD.

4.3.2 Cergy le Haut - Cergy Préfecture

La figure 11 ci-dessous montre l'évolution temporelle du coefficient correspondant au trajet entre Cergy le Haut et Cergy Préfecture :

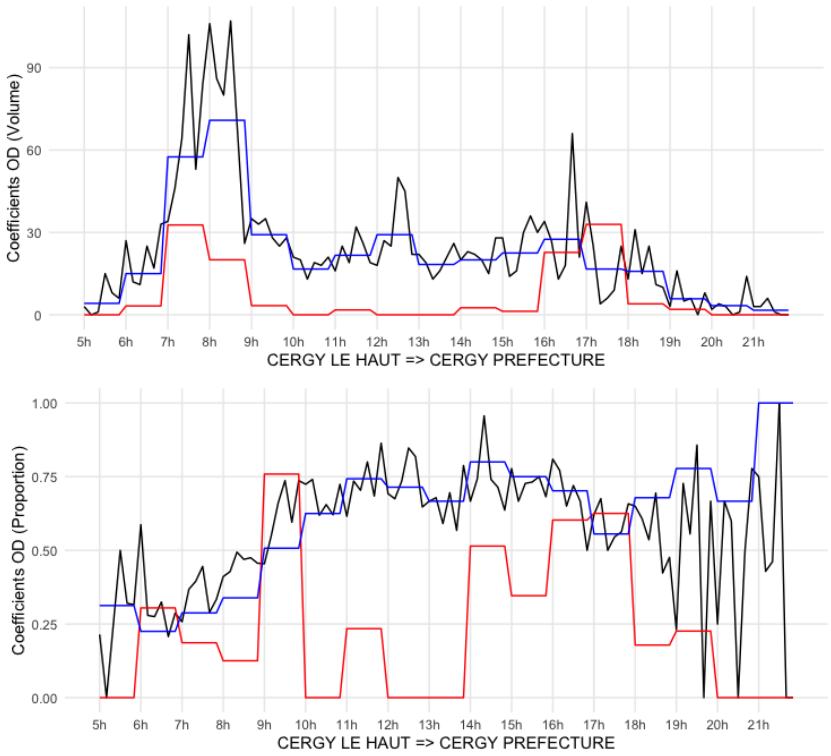


FIGURE 11 – En volume (Haut), en proportion (Bas)
OD poursuite en **noir**, OD manuelle en **rouge**, OD IDFM en **bleu**

La gare de destination de cette nouvelle paire OD est elle aussi équipée de CAB en sortie. Rappelons que nous avons décidé d'étudier ce coefficient OD car le modèle de poursuite d'utilisateur en donne une estimation très différent de ce qui est observé dans les OD manuelles. On remarque cependant que notre modèle est très proche de ce que l'on obtient avec les OD IdFM aussi bien en volume qu'en proportion.

4.3.3 Houilles Carrières sur Seine - Paris Saint-Lazare

La figure 12 ci-dessous montre l'évolution temporelle du coefficient correspondant au trajet entre Houilles Carrières sur Seine et Paris Saint-Lazare :

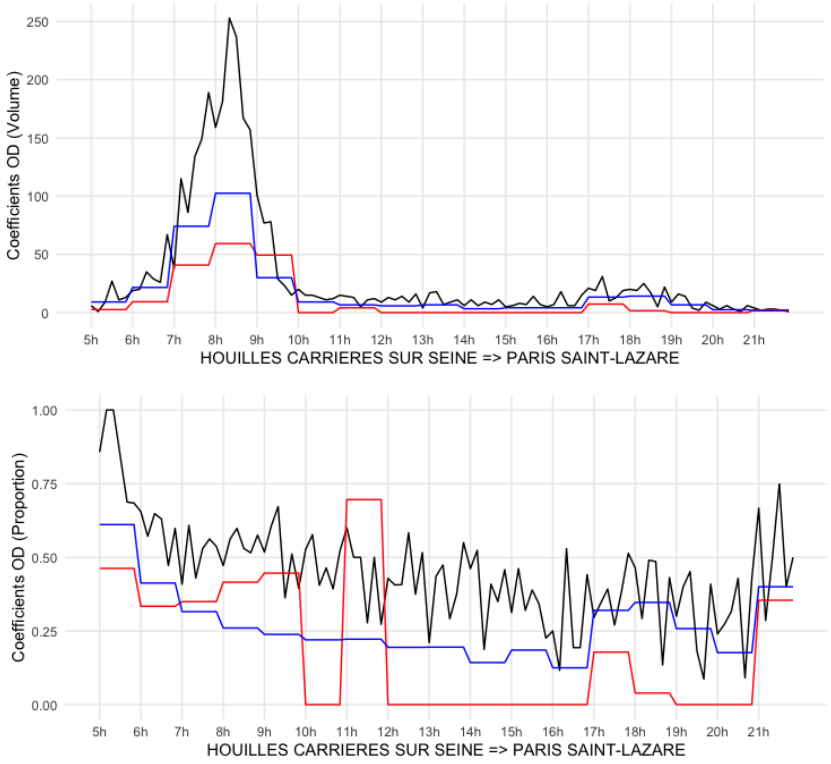


FIGURE 12 – En volume (Haut), en proportion (Bas)
OD poursuite en **noir**, OD manuelle en **rouge**, OD IDFM en **bleu**

Les gares de Houilles Carrières sur Seine et de Paris Saint-Lazare sont reliées directement par deux lignes Transilien : la ligne L et la ligne J. Or, il est beaucoup plus rapide de relier ces deux gares en utilisant la ligne J, ce qui expliquerait pourquoi nous surestimons autant le nombre de voyageurs effectuant ce trajet. Il faut néanmoins remarquer que la méthodologie employée par IdFM semble prendre en compte ce type de singularité.

4.3.4 Houilles Carrières sur Seine - La Défense Grande Arche

La figure 13 ci-dessous montre l'évolution temporelle du coefficient correspondant au trajet entre Houilles Carrières sur Seine et La Défense Grande Arche :

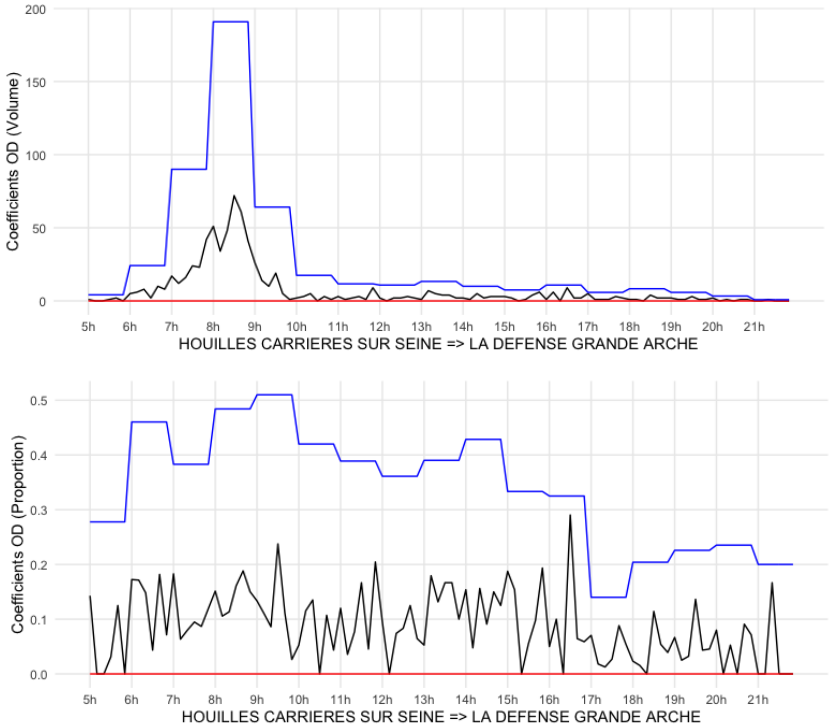


FIGURE 13 – En volume (Haut), en proportion (Bas)
OD poursuite en **noir**, OD manuelle en **rouge**, OD IDFM en **bleu**

Les gares de Houilles Carrières sur Seine et de Paris Saint-Lazare sont reliées par deux lignes : la ligne L (Transilien) et le RER A (portion RATP). Néanmoins la liaison par la ligne L est assez invraisemblable car les deux gares sont sur des branches distinctes, tandis que la liaison par le RER A s'effectue directement. Bien que le coefficient OD issu de la méthodologie IdFM prenne des valeurs importantes il ne s'agit pas d'une erreur car leur objectif est de recontruire des OD à l'échelle réseau et non pas à l'échelle ligne. Cependant il est rassurant d'observer que les volumes et proportions de voyageurs effectuant ce trajet restent faibles avec la méthode de poursuite d'utilisateur et même nuls avec la référence manuelle car il est très probable que la grande majorité des utilisateurs effectuant ce trajet empruntent le RER A.

5 Pistes de recherche

Trois axes de recherche principaux se dégagent : l'amélioration du modèle de poursuite d'utilisateur, l'élargissement des résultats et enfin l'application des matrices estimées.

La principale amélioration à apporter au modèle de poursuite d'utilisateur porte sur la prise en compte des itinéraires alternatifs possibles entre certaines gares. Il s'agit du problème identifié dans la sous-section 4.3.3 et l'utilisation de ratios gare-ligne pour

les couples de gares posant ce problème semble être une solution possible. La question de l'estimation de ces ratios gare-ligne reste néanmoins ouverte. Il serait par ailleurs intéressant d'appliquer le modèle de poursuite d'utilisateur sur d'autres journées et sur d'autres lignes du réseau Transilien afin d'identifier d'éventuels nouveaux problèmes.

Enfin, une application concrète des matrices OD estimées serait de les croiser avec le plan de transport pour en déduire la charge à bord des trains. Les matrices OD seraient estimées en proportion en amont et utilisées pour affecter les voyageurs à partir des données télébilletiques. Ces volumes de voyageurs seraient affectés dans le train correspondant au plan de transport. Les lignes équipées de dispositifs de comptage automatique des montants/descendants disposeraient d'une référence de la charge à bord à comparer avec les estimations du modèle proposé.

6 Conclusion

Ce document présente et compare trois modèles d'estimation de matrices OD à l'échelle d'une ligne à partir des données télé-billetiques.

Les matrices sont estimées sur une ligne du réseau Transilien pour un jour fixé et comparées à des matrices OD issues de deux autres méthodologies de référence. Le modèle le plus proche des deux références, dit de poursuite d'utilisateur, est étudié plus en détail. On observe que ce modèle donne des résultats cohérents dans la grande majorité des cas et en particulier pour les trajets avec des volumes importants. L'étude approfondie de certains coefficients OD nous a permis d'identifier des situations problématiques dans lequel le modèle de poursuite d'utilisateur est défaillant.

Des pistes de réflexions sont finalement proposées pour pallier à ces problèmes d'une part, mais aussi pour poursuivre les recherches en appliquant nos résultats dans une démarche d'estimation de la charge à bord des trains.

Références

Maryam Abareshi, Mehdi Zaferanieh, and Mohammad Reza Safi. Origin-destination matrix estimation problem in a markov chain approach, 2019.

Abderrahman Ait Ali and Jonas Eliasson. Dynamic origin-destination estimation using smart card data : An entropy maximisation approach., 2019.

Hsun-Jung Cho, Yow-Jen Jou, and Chien-Lun Lan. Time dependent origin-destination estimation from traffic count without prior information, 2008.

Projet IVA. Redressement et affectation des od télé-billetiques, 2020.

Florian Toque. Prevision et visualisation de l'affluence dans les transports en commun a l'aide de methodes d'apprentissage automatique, 2019.

Martin Trépanier and Robert Tranchant, Nicolas Chapleau. Individual trip destination estimation in a transit smart card automated fare collection system, 2007.

A.G. Wilson. The use of entropy maximizing models in the theory of trip distribution, mode split and route split, 1969.

Chie Xie, Kara M. Kockelman, and S. Travis Waller. A maximum entropy-least squares estimator for elastic origindestination trip matrix estimation, 2011.