Analyse de la prédiction de la fréquentation annuelle en gare en fonction du comptage cumulé sur les quais transiliens,

Aurélien Henriques

Source des données: - https://ressources.data.sncf.com/explore/dataset/comptage-voyageurs-trains-transilien/information/?sort=year

- https://ressources.data.sncf.com/explore/dataset/frequentation-gares/table/?disjunctive.nom_gare&disjunctive.code_postal&sort=total_voyageurs_2022

La prédiction du nombre de voyageurs annuels fréquentant une gare est un indicateur important de la productivité des infrastructures. Elle permet notamment d'évaluer l'impact d'un changement d'offre sur une ligne de train ainsi que la croissance à long-terme de la gare. Entre autres, cela permet aussi d'évaluer les mouvements au sein du réseau et de projeter les flux potentiels générés par un prolongement ou une nouvelle ligne.

Dans ce contexte, son estimation est un paramètre qu'il est important d'évaluer de manière précise. Dans ce rapport, je propose d'évaluer l'estimation de fréquentations des gares SNCF sur les données disponibles sur les sources. En particulier, à partir de ces jeux de données « fréquentations en gare 2022 SNCF » et « comptage-voyageurs-trains-transilien », je tente d'évaluer l'importance du comptage dans le modèle de prédiction de fréquentation. On sait que c'est une variable qui entre en jeu dans le calcul mais on ne dispose pas d'information quant à l'importance de la variable, et donc du biais qu'elle implique dans le nombre final de voyageurs annuels. J'utilise les données issues de l'année 2019 dans le jeu de données de fréquentations car c'est la dernière année pré-covid où les conditions sont similaires et je cumule les comptages sur tout le jeu de données, c'est-à-dire sur les 6 années de 2014 à 2019, quel que soit le jour du comptage. En recoupant les noms équivalents dans les deux jeux de données, on obtient n=162 gares à partir desquelles j'effectue une régression yn=f(xn) ou yn est la fréquentation de la n-ème gare et xn son comptage total cumulé. Les résultats obtenus sont renseignés dans la figure 1 :

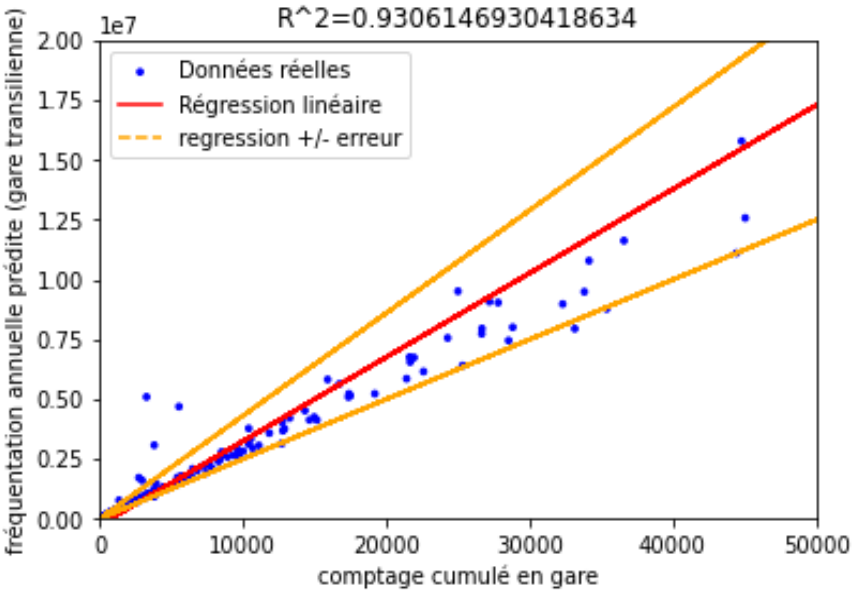


figure 1 : Régression linéaire entre la fréquentation annuelle et le comptage cumulé sur les quais transiliens de la gare. « Données réelles » correspond aux métriques proposées par la SNCF pour chaque gare.

On trouve donc un coefficient de Régression de 0.93, ce qui indique une relation quasi-linéaire entre les deux grandeurs. En particulier, la droite reste cohérente en changeant d'ordre de grandeur, malgré plusieurs points à noter :

-premièrement, on voit que certaines gares à faible comptage sont totalement hors des plages d'incertitudes. On s'intéressera aux raisons de cet écart ensuite.

-deuxièmement, les grandes gares parisiennes, notamment les 3 premières (gare de lyon, saint-lazare et gare du nord) sont clairement dans la plage d'incertitude basse, alors que les fréquentations de ces gares prennent en compte le réseau grande ligne et que les comptages effectués spécifient n'avoir décompter que les voies transiliennes (sans le réseau grande ligne). Ceci est une aberration puisque dans ce cas, ces gares parisiennes devraient se situer au-dessus de la régression, correspondant aux voyageurs de surface en excédent. On comprend donc que la linéarité entre les deux variables n'est plus valable dans le cas de grandes gares avec d'autres voyageurs, mais aussi qu'il existe un biais.

-troisièmement, on s'aperçoit que les gares à très faibles fréquentations (figure 2) augmentent le coefficient de régression par rapport à celles à grandes valeurs : à partir de 25000 comptages, plus aucune gare ne se situe au-delà de la courbe. On peut attribuer ces changements au fait qu'avec un nombre grandissant de passagers, on augmente aussi le nombre de passagers non-comptés dans les foules et donc on sous-estime le comptage réel.

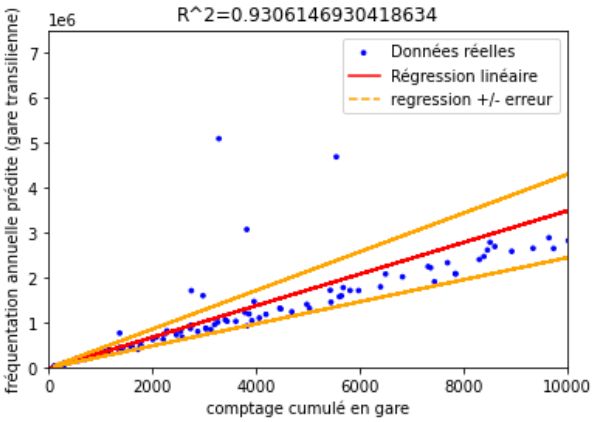


Figure 2 : Zoom de la précédente figure sur les faibles fréquentations/comptages. Pour les très faibles comptages (<1500), les points sont au-dessus de la régression, tandis que les points avec un comptage plus élevés sont au-dessous.

Pour compenser ces différentes aberrations, on peut entreprendre d'une part de changer les poids dans la régression, quitte à diminuer notre R**2 afin d'obtenir une meilleure mesure sur les gares à fortes fréquentations. On peut aussi s'intéresser aux raisons de l'incohérence des valeurs basses au-dessus de la courbe (points qui ne sont pas dans la marge d'erreur) :

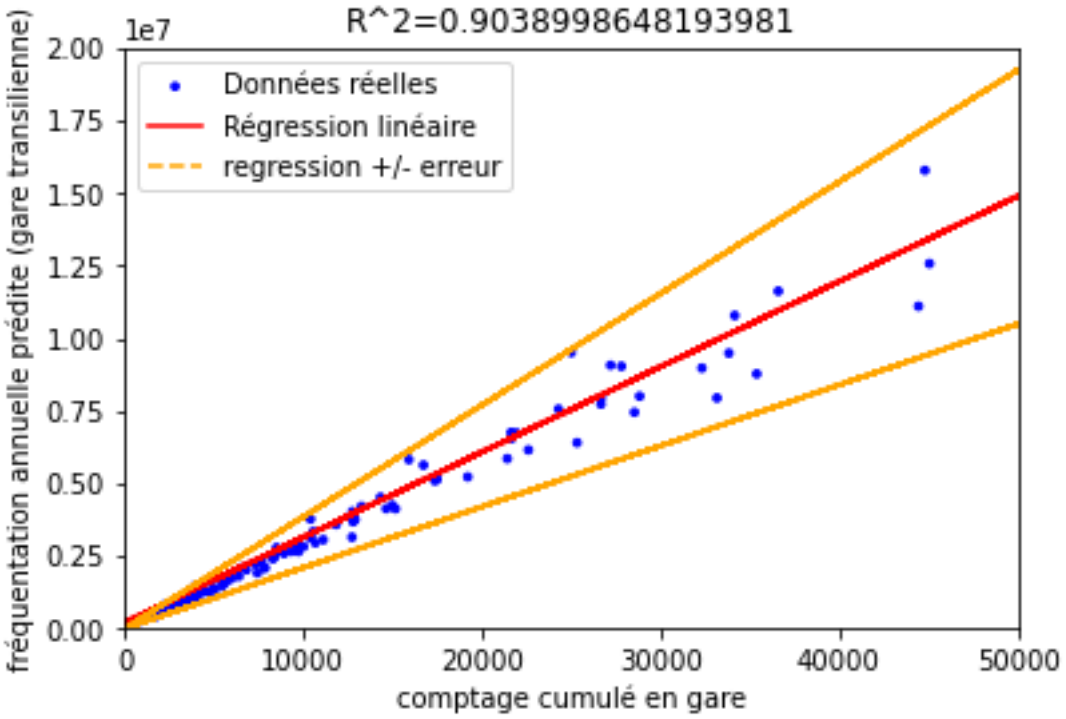
En les identifiant (gares de Montereau, Creil, Rambouillet, Gisors, Dreux, Montargis), on constate qu'il s'agit de gares situées loin du centre de Paris et représentant des pôles importants dans leur proximité ferroviaire (terminus transilien, aiguillage, …). Cela signifie qu'elles sont toutes desservies par des TER/Intercités qui offrent une meilleure performance pour rejoindre le centre de Paris que les transiliens, ce qui implique que le comptage sera d'une part amoindrie par le manque de personnes empruntant le transilien dans ces gares, mais sous-estime aussi le nombre réel de voyageurs présent dans ces gares qui emprunte d'autres trains. La surestimation de la fréquentation par rapport au comptage transilien dans ce cas est donc justifiée.

On peut supprimer ces gares de notre régression pour obtenir une meilleure prédiction sur les gares transiliens pures (gain de 0.1% sur R**2). Entre autres, les gares de Château-Thierry, Etampes, Dourdan, Crépy-en-Valois, Vernon représentent un cas de figure similaire. Les gares de Longueville, Mantes-la-Jolie et Meaux sont des cas particuliers qu'il est difficile de classer dans la catégorie ou

non, puisqu'il n'est pas possible avec ces données de prédire l'utilisation du service transilien ou TER du fait de la complexité de leurs services.

Fort heureusement, les gares d'Etampes, Château-Thierry et Vernon n'étaient déjà pas considérées dans le comptage de la base de données. De plus, la suppression des gares restantes ne permet pas d'améliorer la régression. Deux hypothèses permettent de l'expliquer : soit l'offre transilien dans ces gares est telle que l'alternative ter n'est pas valable pour les passagers, ce qui implique que la partie utilisant le ter soit négligeable et donc n'influe pas sur la relation fréquentations/comptage transilien, soit le biais d'augmentation du coefficient par les gares à faible fréquentations compense la différence entre fréquentation totale et fréquentation transilien, ce qui se traduit par une changement nul de la régression par ces gares. En vue du graphique, il est semblable que cette deuxième option soit la véritable cause puisque à partir d'un seuil (environ 600k passagers et 2000 personnes comptées) plus aucun des points ne se situent au-dessus de la courbe.

Pour compenser ce biais, on peut ajuster les poids de la régression pour que les hautes valeurs soient prises en compte de manière plus importantes. Par exemple, en prenant l'inverse des valeurs de fréquentations comme poids, on peut réajuster le coefficient pour que les plus grandes comptent plus (on abaisse le coefficient de la régression de 352 passagers annuels/comptage à 317). Une autre méthode est la régression de Huber qui attribue à la fonction de coût (erreur entre la sortie yn prédite par la régression et la fréquentation de la base de données) un poids quadratique proche de 0 et un poids linéaire pour les grandes valeurs, ce qui implique que les données qui sortent du régime linéaire de la courbe, comme c'est le cas pour les grandes gares, seront mieux traités par la régression. En l'occurrence, il s'agit seulement d'un moyen de rendre la droite plus cohérente par rapport au graphique, tout en réajustant le coefficient à 293 voyageurs/comptage.



En conclusion, bien que le comptage semble être la mesure majoritaire pour établir les modèles de prévision de Traffic sur les gares transilien (forte dépendance linéaire entre les deux), il apparaît que cette variable n'est pas suffisante pour décrire la fréquentation, notamment dans les gares à faible fréquentation où l'offre ter est plus compétitive que l'offre transilien et donc les passagers sont moins

nombreux sur les quais transiliens. En outre, il semble aussi y avoir un ajustement entre les fréquentations des grandes gares parisiennes, où les voyageurs du réseau grande ligne peuvent être comparables à ceux du transilien, et le comptage, puisque celui-ci décrit bien la fréquentation alors que le comptage n'est pas considéré sur le réseau grande ligne, qui est quant à lui bien pris en compte dans la fréquentation de la gare. On en déduit que le comptage est particulièrement utilisé sur les gares Transilien Pur pour la fréquentation mais que la fréquentation profite d'un ajustement notable pour les gares dont le Traffic se dispatche notablement entre Transilien et offre de trainq externe. En particulier, l'ajout de variables extérieures (comptage automatique dans les trains, sur les quais, utilisation des nombres de validations Navigo, du nombre de billet transilien vendus…) permettrait d'améliorer la précision de la mesure, et amoindrir l'erreur potentielle qu'induirait une personne non-comptée (actuellement de l'ordre de 300 voyageurs annuels en moins dans la gare).