Prédiction de l'affluence journalière dans un réseau de transports urbains

Apolline Louvet, Andony Arrieula, Jean Prost, Paul Freulon

Prédiction de l'affluence journalière dans un réseau de transports urbains

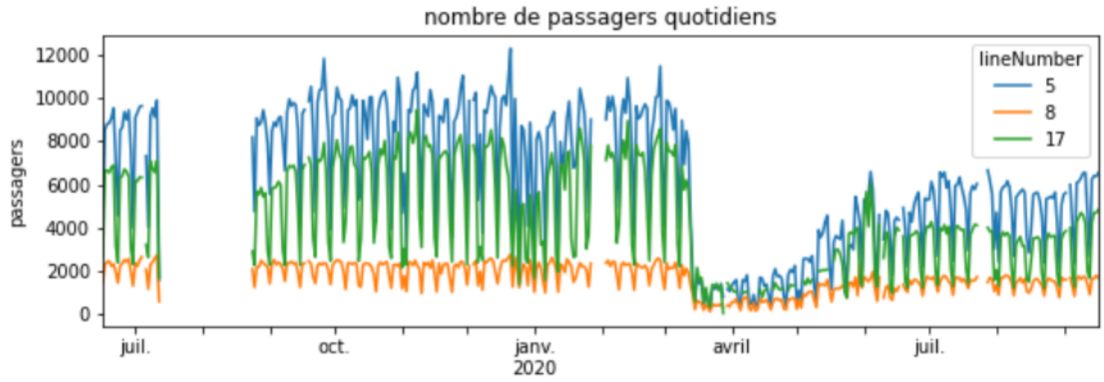
Apolline Louvet[1], Andony Arrieula[3,4,2], Jean Prost[2], Paul Freulon[2]

November 2020

[1] MAP5, UMR CNRS 8145, Université de Paris, 45 rue des Saints-Pères, 75270 Paris Cedex 06, France
[2] Université de Bordeaux, IMB, UMR 5251, Talence, France
[3] CARMEN Research Team, Inria Bordeaux Sud-Ouest, Talence, France
[4] IHU Liryc, fondation Bordeaux Université, Pessac, France

Table des matières

FIGURE 1 – Affluences journalières mesurées pour trois des lignes de bus présentes dans le jeu de données, de juin 2019 à septembre 2020.



1 Introduction

Ce rapport présente les résultats de travaux effectués durant la Semaine d'Etude Mathématiques et Entreprise (ou SEME) organisée conjointement par l'AMIES et l'Institut de Mathématiques de Bordeaux, du 26 au 30 octobre 2020. Durant cette semaine, nous avons travaillé sur un sujet proposé par l'entreprise Hupi, et portant sur la prédiction d'affluence sur des lignes de bus.

1.1 Présentation du problème

Pour une compagnie gérant un réseau de transports en commun, pouvoir prévoir à l'avance l'affluence sur différentes lignes de bus est une information importante. En effet, cela permet à l'entreprise de mettre des bus en priorité sur les lignes qui seront les plus fréquentées, et ainsi d'améliorer les conditions de transports des usagers tout en optimisant le nombre de bus en circulation. Cependant, beaucoup de facteurs peuvent influencer l'affluence. Citons entre autres la météo du jour, le jour de la semaine, ou le fait d'être un jour de vacances scolaires. Ces facteurs n'ont a priori pas le même effet selon la ligne de bus considérée : ainsi, une ligne de bus desservant des établissements scolaires sera plus fréquentée en semaine hors vacances scolaires, tandis qu'une ligne desservant une gare sera particulièrement fréquentée les jours de départ et retour de vacances. A l'effet de facteurs extérieurs se rajoute une potentielle corrélation entre les affluences sur une ligne donnée, d'un jour à l'autre.

La problématique sur laquelle nous avons travaillé a été la prédiction à 5 jours de l'affluence journalière sur une ligne de bus. Afin de pouvoir tester différentes approches, il a été mis à notre disposition un jeu de données rassemblant les affluences journalières sur différentes lignes de bus dans une même ville, sur une période allant de juin 2019 à septembre 2020, ainsi que les informations suivantes :

— jour de la semaine
— jour férié ou non
— météo du jour
— jour de vacances scolaires ou non
— événement sportif : match de football
— événement culturel.

1.2 Analyse préliminaire du jeu de données

Une analyse préliminaire du jeu de données a permis de mettre en évidence plusieurs éléments. Tout d'abord, nous avons pu observer une périodicité hebdomadaire marquée des données : l'affluence est plus élevée en semaine que le week-end. Les vacances scolaires ont elles aussi un effet visible, l'affluence étant généralement plus faible durant ces périodes. Cet effet est plus ou moins marqué selon la ligne de bus considérée. Enfin, l'effet des mesures prises à partir de mars 2020 pour lutter contre l'épidémie de COVID-19 est clairement visible. Il se caractérise par une chute drastique de l'affluence, suivi d'une augmentation progressive, sans jamais retrouver les niveaux pré-épidémie.
L'ensemble de ces observations sont visibles sur la figure 1, qui représente les affluences journalières mesurées pour trois des lignes de bus.

1.3 Approches considérées

Durant la SEME, nous avons adopté deux approches distinctes. Dans un premier temps, nous avons considéré que les variables explicatives identifiées permettaient d'expliquer complètement les variations d'affluence, et que conditionnellement à ces variables, les affluences en des jours distincts étaient indépendantes. Afin de simplifier le problème, nous avons supposé que les différentes variables explicatives étaient indépendantes, et que l'affluence observée pouvait s'exprimer comme une combinaison linéaire des différentes variables explicatives, en un sens que nous expliciterons plus loin. De plus, pour séparer le problème de prévision des affluences d'un problème de prévision de la météo, toutes les estimations ont été réalisées conditionnellement à la météo observée.

Ensuite, nous avons adopté le point de vue inverse, et supposé cette fois-ci que l'affluence un jour donné était entièrement expliquée par l'affluence des jours précédents, sans effet des variables explicatives. Ce point de vue a été motivé par la périodicité observée dans les affluences, et consiste à modéliser l'affluence journalière sur une ligne de bus donnée par une série temporelle. Ces deux approches correspondent en quelque sorte aux deux extrémités d'un spectre. En pratique, variables explicatives et affluences des jours précédents ont sans doute chacune un effet, et il pourrait être intéressant de développer des modèles combinant ces deux approches.

2 Approche par régression linéaire

2.1 Présentation

L'approche par régression linéaire consiste à modéliser le problème sous la forme d'un système linéaire. Autrement dit, pour une ligne de bus donnée, si $Y = (y_i)_{1 \leq i \leq n}$ est le vecteur des affluences mesurées sur cette ligne, et si $X = (x_{i,j})_{1 \leq i \leq n, 1 \leq j \leq m}$ est le vecteur des variables explicatives associées à chacun des jours d'observation, alors nous supposons l'existence de $\boldsymbol{\beta} = (\beta_j)_{0 \leq j \leq m}$ tel que

$$Y = X\boldsymbol{\beta}$$

auquel s'ajoute un terme d'erreur. Il est possible d'estimer $\boldsymbol{\beta}$ en cherchant à résoude le problème des moindres carrés

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \sum_{j=1}^{m} x_{i,j}\beta_j))^2$$

Dans le cas de notre problème, nous avons pris comme variables explicatives toutes celles présentées dans l'introduction. La régression linéaire est déjà implémentée en R (fonction *lm*)

ou en Python (bibliothèque *scikit-learn*). Dans la suite, nous avons utilisé l'un ou l'autre de ces deux langages de programmation selon les cas.

Plusieurs raisons peuvent pousser à préférer l'approche par régression linéaire à d'autres méthodes d'estimation. Déjà, le coût en calcul pour estimer $\boldsymbol{\beta}$ est très faible, même pour des jeux de données très conséquent. De plus, il est facile d'intégrer de nouvelles variables explicatives au modèle. Enfin, la méthode est robuste à la présence de jours d'observation manquants, car il n'est pas nécessaire d'avoir des jours d'observation consécutifs. Les inconvénients de la méthode sont surtout liés à la potentielle inadéquation du modèle à la situation réelle qu'il doit modéliser, entre autres dans le cas d'une variation progressive de l'affluence.

2.2 Résultats

Tout d'abord, nous nous sommes intéressés aux performances de la méthode d'estimation sur la période de temps Juin 2019 - Février 2020, qui correspond à la période pré Covid-19. Le jeu de données a été divisé en deux portions : 80% du jeu de données a été utilisé pour entraîner l'algorithme, et les 20% restants ont été utilisés pour tester la méthode d'estimation. Les résultats obtenus sur l'ensemble des lignes de bus sont indiqués dans la figure 2 en annexe. Nous pouvons remarquer que si le biais sur l'estimation est de l'ordre du passager, la variance est plus conséquente. L'erreur relative est quant à elle de 20%, que ce soit sur le jeu d'entraînement ou le jeu de test. On peut donc supposer qu'une meilleure calibration du modèle sur les données d'entraînement provoquera une meilleure précision dans les futures prédictions.

Nous avons ensuite testé l'estimateur sur la période de temps Juin 2019 - Septembre 2020, afin d'évaluer l'impact d'un changement brusque des affluences sur la qualité de l'estimation, en utilisant le même protocole que précédemment. Les résultats obtenus sont indiqués dans la figure 3 en annexe. Là encore, le biais sur l'estimation est de l'ordre du passager. Toutefois, cette fois-ci, la variance est deux fois plus grande que dans le cas précédent, et l'erreur relative est cette fois-ci de l'ordre de 40%. La méthode d'estimation n'est donc pas robuste à des changements brusques de l'affluence.

2.3 Variantes

Pour compléter l'approche par régression linéaire, nous avons aussi considéré plusieurs variantes permettant de prendre en compte des changements progressifs des affluences, et d'estimer l'incertitude sur les prévisions effectuées avec cette approche. Nous avons implémenté et testé l'une d'entre elles. Pour les autres, nous indiquons quelles méthodes et packages pourraient être utilisés pour les implémenter.

2.3.1 Estimation des variations d'affluences autour d'une valeur moyenne

Dans un premier temps, pour chaque ligne de bus, nous avons cherché à séparer affluence moyenne et variations d'affluence. De plus, nous avons fait l'hypothèse que les variables explicatives agissent sur l'écart relatif plutôt qu'absolu de l'affluence par rapport à l'affluence moyenne. Cela revient à dire que par exemple, l'affluence un dimanche est inférieure de 30% un dimanche, plutôt qu'inférieure de 200 personnes.
La différence entre écart relatif et écart absolu n'a pas d'intérêt si l'affluence un jour donné est comparée à l'affluence moyenne est calculée sur l'ensemble de la période d'observation. Cependant, la situation est très différente si cette affluence est comparée à l'affluence moyenne sur une période d'observation plus limitée et différente pour chaque jour considéré, par exemple

l'affluence sur l'année précédente.
L'intérêt de cette variante est qu'elle permet de prendre en compte une variation progressive de l'affluence moyenne. Toutefois, pour que les variations de l'affluence moyenne correspondent bien à une tendance réelle, il faut que les variables explicatives soient distribuées de la même façon quelle que soit la période d'observation. Ceci impose donc de prendre comme période d'observation un multiple d'une an, pour avoir le même nombre de jours de vacances scolaires et de jours fériés quelle que soit la période considérée.
Ne disposant que de moins d'une année d'observations hors période COVID, nous n'avons pas pu tester cette variante sur un jeu de données réelles. Cependant, dans la variante 2 présentée ci-dessous, nous avons travaillé avec l'écart relatif à la valeur moyenne sur l'ensemble de la période d'observation.

2.3.2 Pondération des données en fonction de leur âge

Afin de prendre en compte de possibles évolutions de l'affluence moyenne au cours du temps, une autre approche possible consiste à donner plus de poids aux observations les plus récentes par rapport aux observations plus anciennes. Formellement, si à chacune des observations y_i, $1 \leq i \leq n$ est associé un poids $w_i > 0$, alors il est possible d'intégrer les poids dans la régression linéaire en cherchant à minimiser

$$\sum_{i=1}^{n} w_i (y_i - \sum_{j=1}^{m} x_{i,j}\beta_j)^2$$

plutôt que (2.1).
En nous basant sur [1], qui cherche à prédire les résultats de matchs de football en fonction des résultats précédents, nous avons choisi une fonction poids de la forme

$$\exp(-\lambda \Delta t) \tag{1}$$

avec $\lambda > 0$, Δt la durée (en nombre de semaines) s'étant écoulée depuis l'observation. Il est possible d'agir sur λ pour donner plus ou moins de poids aux données les plus anciennes.
Nous avons testé cette variante sur quelques lignes de bus, en réalisant la régression linéaire pondérée sur les écarts relatifs par rapport à l'affluence moyenne, à partir des observations de septembre 2019 à janvier 2020, et en cherchant à prédire l'affluence en février 2020, conditionnellement à la météo. Nous avons pris comme variables explicatives la météo, le jour de la semaine, et si il s'agissait d'un jour de vacances scolaires ou non.
Malheureusement, nous n'avons pas réussi à trouver de valeur de λ montrant un effet positif sur l'estimation de l'ajout de cette pondération. Pour un λ faible, aucun effet n'était visible. Pour un λ plus grand, un effet était visible, mais l'estimation était plutôt globalement moins bonne (voir figure 4 en Annexe). Il est possible qu'avec plus de données, l'ajout d'une pondération des données les plus anciennes améliore l'estimation pour un λ bien choisi.
La régression linéaire pondérée est déjà implémentée en R et en Python, dans les modules utilisés pour la régression linéaire classique.

3 Approche par séries temporelles

3.1 Présentation

La deuxième approche que nous avons adopté consiste cette fois-ci à considérer que les variations de l'affluence peuvent être entièrement expliquées par l'affluence les jours précédents.

Formellement, si l'affluence journalière est donnée par $y_t, t \in \mathbb{N}$, nous considérerons qu'il existe $p \in \mathbb{N}^*$, et $\phi_0, \phi_1, ..., \phi_p \in \mathbb{R}$ tels que pour tout $t \in \mathbb{N}$,

$$y_{t+p} = \phi_0 + \sum_{i=0}^{p-1} \phi_i y_{t+i} + \epsilon_{t+p},$$

où ϵ_{t+p} est un bruit blanc gaussien. Ce type de modèle est appelé modèle autorégressif d'ordre p, et noté AR(p).
p peut être déterminé en utilisant un autocorrélogramme. Ici, du fait de la périodicité hebdomadaire observée lors de l'analyse préliminaire, nous avons considéré que $p = 7$. Du fait du nombre de jours d'observations limité du jeu de données, nous n'avons pas pu intégrer d'effet de saisonnalité, mais ceci est possible avec des jeux de données recouvrant plusieurs années d'observation.

Comme dans le cas de la régression linéaire, les paramètres $\phi_0, \phi_1, ..., \phi_p \in \mathbb{R}$ peuvent être estimés en utilisant la méthode des moindres carrés : si les observations d'affluence ont lieu jusqu'au temps T, alors la quantité à minimiser est

$$\sum_{t=p}^{T} (y_t - (\phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p}))^2.$$

Pour déterminer les paramètres du modèle autorégressif, nous avons utilisé le package python statsmodels [2]. Le principal obstacle qui s'est posé à l'implémentation de la méthode d'estimation est le fait qu'elle nécessite des observations consécutives. Ceci nous a amené à n'entraîner d'abord le jeu de données que sur la période de Septembre 2019 à Janvier 2020, puis à réentraîner le modèle régulièrement (tous les 1 à 5 jours) au fur et à mesure que de nouvelles observations étaient faites. Il serait toutefois possible d'adapter la méthode pour intégrer la présence de données manquantes.

3.2 Résultats

Dans un premier temps, nous nous sommes intéressés à la qualité des estimations en l'absence de changements brusques dans l'affluence. Pour cela, nous avons cherché à estimer les affluences sur la période de février 2020 pour chaque ligne de bus, en entraînant le modèle sur les données de septembre 2019 à janvier 2020, et en réentraînant le modèle chaque jour. Les estimations obtenues pour deux des lignes de bus figurent en annexe (figure 5). Nous pouvons observer que les estimations réalisées une semaine donnée sont en fait généralement très proches de ce qui a été observé la semaine précédente.

Dans un second temps, nous nous sommes intéressés à la capacité de la méthode d'estimation à suivre des changements brusques dans l'affluence. Pour cela, nous avons cherché à estimer les affluences de février à avril 2020 pour chaque ligne de bus, en entraînant le modèle sur les données de septembre 2019 à janvier 2020, et en réentraînant le modèle tous les 5 jours. Du fait du COVID-19, l'affluence dans les transports en commun a fortement chuté à partir de mi-mars. La méthode d'estimation arrive effectivement à capter cette diminution, et après un retard d'une semaine les affluences estimées sont du même ordre de grandeur que les affluences mesurées. Les résultats obtenus pour deux des lignes de bus figurent en annexe (figure 6).

4 Conclusion et perspectives

Les deux approches que nous avons présenté dans ce rapport sont en fait complémentaires. L'approche par régression linéaire permet de prendre en compte l'effet de variables explicatives,

telles que la météo ou le fait d'être un jour de vacances scolaires. Toutefois, elle ne permet pas de détecter de variation progressive de l'affluence, et n'est pas adaptée à des changements brusques d'affluence, comme cela a été le cas à partir de Mars 2020 du fait du COVID-19.
A l'inverse, l'approche par séries temporelles, en ne prenant en compte que les corrélations entre l'affluence un jour donné et l'affluence les jours précédents, permet d'intégrer les variations progressives comme les changements abrupts d'affluence. Cependant, l'absence de l'intégration de variables explicatives telles que la météo fait qu'une partie de la variation de l'affluence reste inexpliquée. Une piste intéressante serait de chercher à combiner ces deux approches, par exemple en exprimant l'affluence sur une ligne un jour donné comme une combinaison linéaire des variables explicatives *et* des affluences les jours précédents.

Remerciements

Nous tenions à remercier l'AMIES, l'Institut de Mathématiques de Bordeaux et Edoardo Provenzi pour l'organisation de la SEME, et l'entreprise Hupi pour nous avoir proposé ce sujet. Nous remercions tout particulièrement Kattin Dassance, qui a représenté Hupi à la SEME, pour son aide durant cette semaine.

Références

[1] Mark J Dixon and Stuart G Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 46(2) :265–280, 1997.

[2] Skipper Seabold and Josef Perktold. statsmodels : Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

5 Annexes

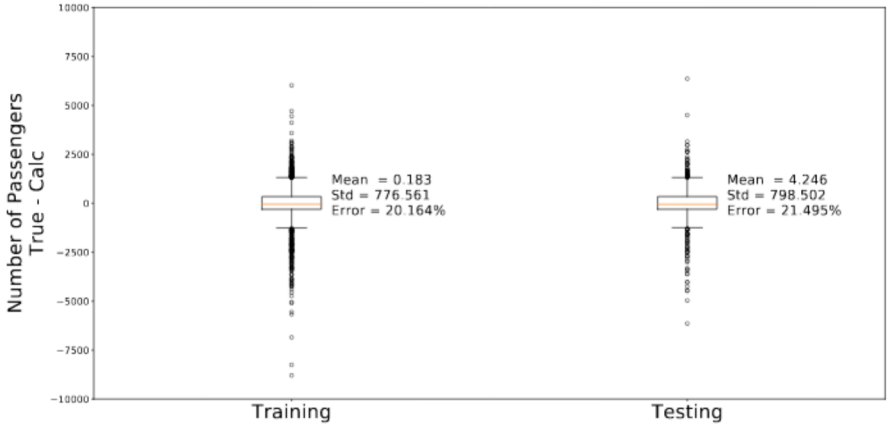


FIGURE 2 – Différences entre les valeurs enregistrées et les valeurs estimées par régression linéaire, pour un jeu de données allant du 16/06/2019 au 29/02/2020 (période pré-Covid).

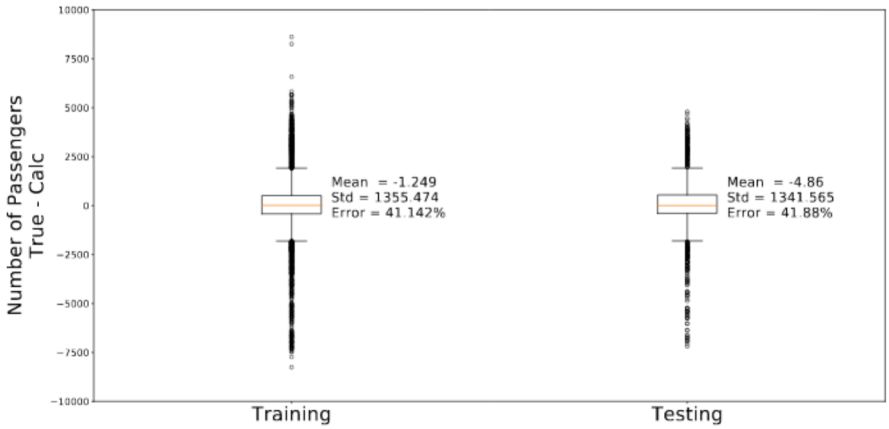


FIGURE 3 – Différences entre les valeurs enregistrées et les valeurs estimées par régression linéaire, pour un jeu de données allant du 16/06/2019 au 17/09/2020.

FIGURE 4 – Affluences observées et estimées avec ou sans pondération des observations en fonction de l'âge pour deux des lignes de bus. Les courbes bleues correspondent aux affluences réellement observées. Les courbes rouges correspondent aux affluences estimées sans pondération, et les courbes jaunes aux affluences estimées en pondérant les observations en fonction de leur âge par 1, avec $\lambda = 0.1$.

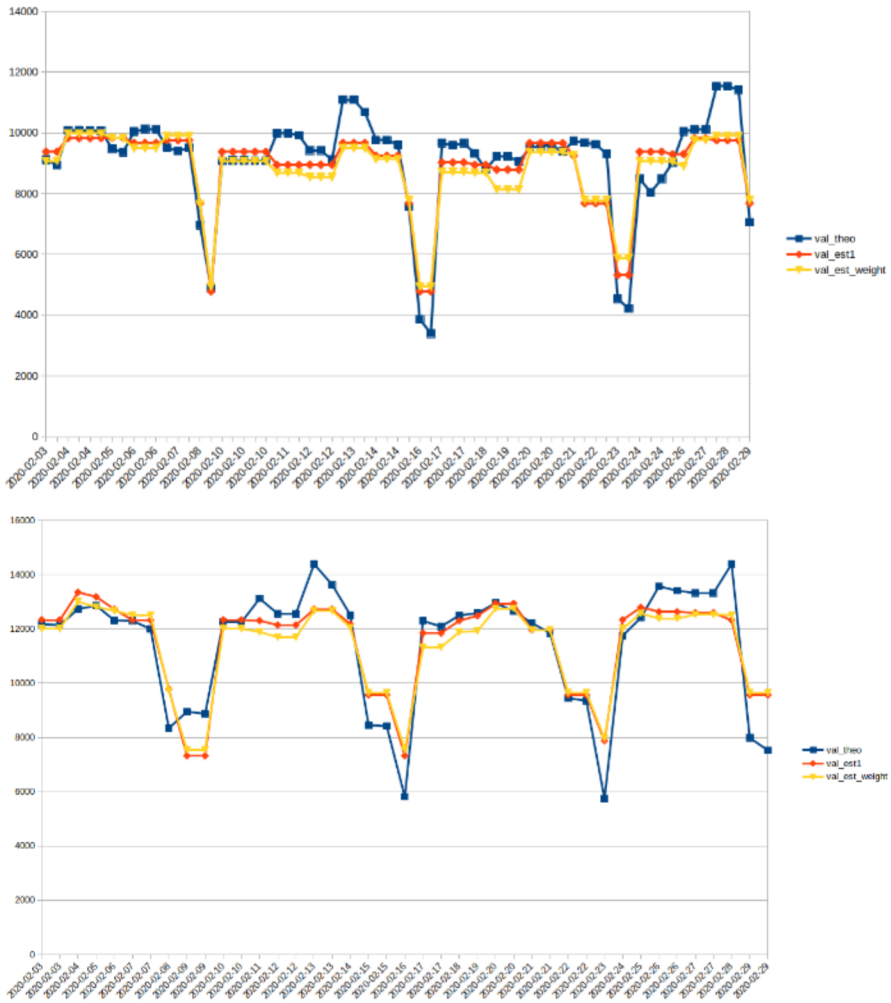


FIGURE 5 – Affluences observées et estimées en février 2020 pour deux lignes de bus, en utilisant un modèle autorégressif d'ordre 7. Les courbes bleues correspondent aux affluences mesurées, et les courbes oranges aux affluences estimées. L'algorithme a été réentraîné tous les jours en intégrant à chaque fois les nouvelles données.

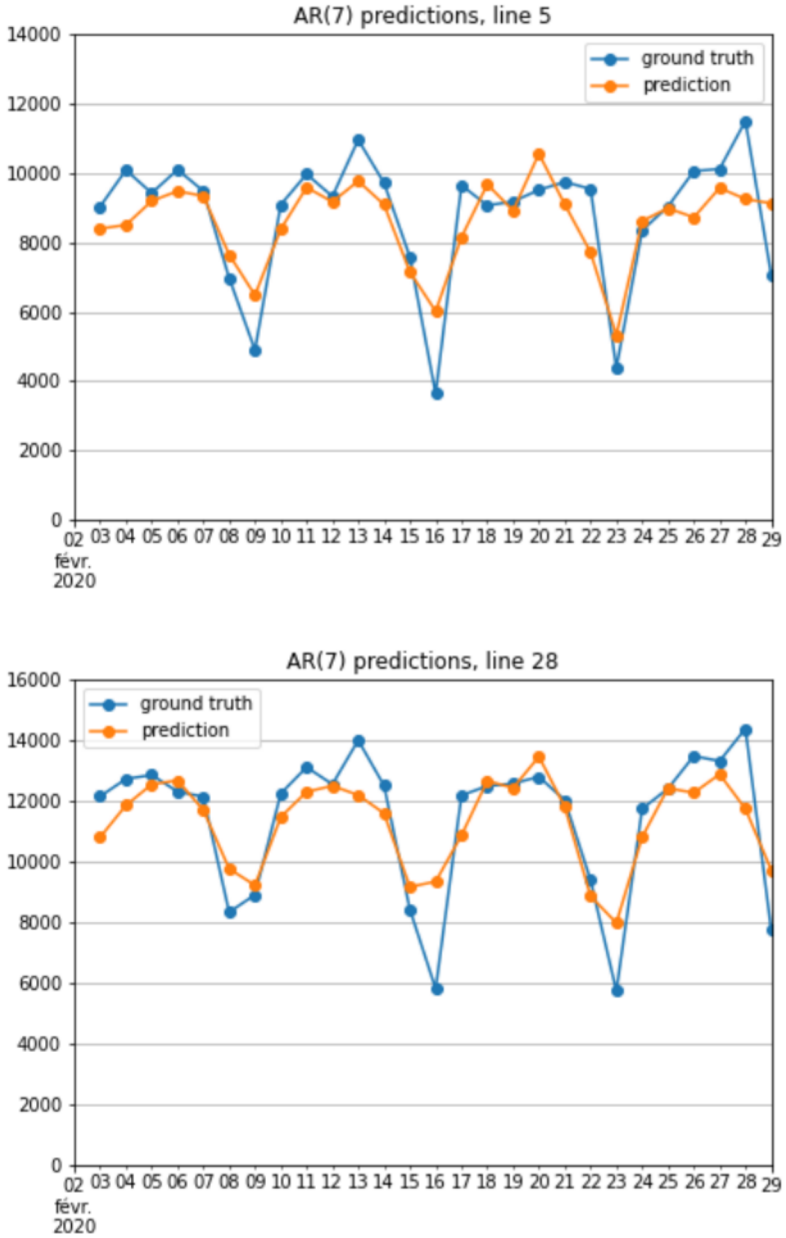


FIGURE 6 – Affluences observées et estimées de février 2020 à avril 2020 pour deux lignes de bus, en utilisant un modèle autorégressif d'ordre 7. Les courbes bleues correspondent aux affluences mesurées, et les courbes oranges aux affluences estimées. L'algorithme a été réentraîné tous les 5 jours en intégrant à chaque fois les nouvelles données. Les mesures de lutte contre le COVID-19 prises à partir de mi-mars ont entraîné une chute des affluences mesurées.