Projet de Statistique Appliquée

Note de synthèse

Outil dynamique de suivi des prévisions de Trafic aérien

Claire He
Victor Huynh
Solène Blasco Lopez
Antonin Falher

Référent des Aéroports de Paris : Raphaël Boudra
Correspondant Ensae : Jean-Michel Zakoian

21 mai 2021

Introduction

Le Groupe ADP, anciennement Aéroports de Paris, s'occupe de la gestion aéroportuaire des principales plateformes en France. Il accueille et transporte chaque année des centaines de millions de passagers grâce à ses installations et ses services. Il est crucial pour une telle entreprise de pouvoir anticiper les besoins clients le plus précisément possible, en utilisant les données issues de l'analyse du trafic aérien et de la fréquentation des aéroports. Cependant, l'entreprise fait face à des contraintes opérationnelles. Elle doit donc pouvoir suivre les prévisions de manière dynamique selon différents indicateurs, différents filtres et différents niveaux d'agrégation. Nous avons donc travaillé autour de deux aspects : l'amélioration des performances de prédiction et l'efficacité opérationnelle à travers l'outil interactif Power BI.

Comment prédire au mieux le trafic aérien à différents horizons de prévision sur chacun des faisceaux desservis par les Aéroports de Paris ?

Afin de répondre à cette problématique, nous nous sommes intéressés à plusieurs modèles de prévision du nombre de passagers journaliers, dont nous avons comparé les performances à l'aide d'indicateurs. Les résultats ont été délivrés dans un outil Power BI, qui permet la visualisation dynamique et interactive des prévisions, sur différents faisceaux et horizons de prévision.

1 Présentation des données

Nous avons travaillé sur deux types de jeux de données fournis par le Groupe ADP : des données "historiques" contenant les informations concernant des vols réalisés entre 2008 et 2018, et des données de prévisions faites "à la main" par les Aéroports de Paris sur les vols de 2015 à 2017. Les deux jeux de données ont été agrégés aux mêmes niveaux de granularité. Ce sont les données "historiques" qui nous ont servi pour générer nos prédictions, à travers les différents modèles explicités ci-après. Notre but était de comparer la performance des prévisions de nos modèles avec celles calculées par les Aéroports de Paris.

Les historiques du trafic aérien recensent l'ensemble des vols réalisés et fournissent un certain nombre d'informations telles que :

- la date du vol
- le type de mouvement, soit une arrivée soit un départ
- le faisceau du vol qui correspond à la zone géographique associée au vol, parmi les 5 modalités *International*, *Schengen, National, Autre UE* et *Dom-Tom*
- le nombre de passagers pour chaque vol
- le nombre de sièges prévus sur chaque vol
- le coefficient de remplissage de chaque vol (un pourcentage estimé à partir d'un nombre de sièges moyen, et du nombre de passagers)

D'autres informations telles que le pays, la ville de provenance ou d'arrivée ou l'aérogare étaient aussi données dans la base initiale, mais n'ont pas été retenues puisque nous avons agrégé nos données au niveau "faisceau", par date (au jour) et selon les arrivées et les départs. D'autre part, afin de comparer la performance de nos modèles de prévisions à celles fournies par ADP, nous nous sommes restreints à l'aérogare d'Orly.

De plus, pour améliorer les prévisions de certains de nos modèles, nous avons utilisé une troisième base de données mise à disposition par les aéroports de Paris : il s'agit d'une base `Calendrier`, qui contient pour chaque jour des informations comme le jour de la semaine, le fait qu'il soit férié, ou qu'il fasse partie des vacances scolaires, ...

Nos modèles de prévisions génèrent alors les prédictions de trafic en terme de *nombre de passagers* par jour, à partir des historiques filtrés sur l'arrivée ou le départ, ainsi que le faisceau choisi. Ils peuvent ainsi être comparés facilement entre eux, ainsi qu'avec les prévisions fournies par les Aéroports de Paris sur le court terme.

2 Modèles implémentés

Différentes approches ont été envisagées pour la prédiction du trafic aérien à partir des données. Notre objectif a été de réaliser des prévisions capables de rendre compte de la tendance et des saisonnalités annuelle et hebdomadaire que présentent certains faisceaux. Nous avons également proposé le calcul d'intervalles de confiance pour chacun de nos modèles.

D'une part, en raison de la nature de nos données, qui sont des séries temporelles que nous avons converties en séries chronologiques par l'agrégation effectuée sur la date, nous avons implémenté des modèles de séries temporelles linéaires ARIMA et SARIMA - ce dernier étant une variante du premier qui tient compte des saisonnalités (motifs périodiques dans le temps). Ces modèles linéaires permettent d'effectuer des prédictions futures à partir de transformations linéaires des valeurs passées - ce qui explique que l'on parle d'un modèle AR ie autorégressif - et des erreurs, en supposant un certain nombre d'hypothèses. Ces modèles linéaires sont des modèles assez classiquement utilisés en statistique et en économétrie afin d'obtenir des prédictions temporelles.

D'autre part, nous avons été amenés à implémenter un modèle non paramétrique, dont le principe est basé sur l'observation des similarités entre certaines périodes de chaque année. Ces similarités peuvent être dues à des périodes de charge ou de décharge du trafic aérien lors de jours fériés ou de vacances. C'est pourquoi nous avons utilisé une base de données `Calendrier` contenant les dates correspondant à ces périodes particulières pour ce modèle. Le modèle non paramétrique pondère alors les motifs semblables dans le passé à l'aide de ce calendrier, afin de générer par une combinaison linéaire des futurs sur ces motifs la prévision future.

Enfin, nous avons travaillé sur un modèle de régression linéaire pénalisée Lasso, dans lequel nous avons cherché à prédire le trafic futur à l'aide d'indicatrices temporelles passées. L'idée est qu'en régressant sur les indicatrices temporelles passées, nous puissions "sélectionner" les périodes passées pertinentes à la prédiction. Cette méthode a l'avantage d'être parcimonieuse puisqu'elle sélectionne les meilleurs indicatrices explicatives lorsque le nombre de variables explicatives potentielles est grand.

Tous nos modèles sont implémentés sous Python, de manière à renvoyer en sortie la prévision à l'horizon demandé, mais aussi un intervalle de confiance à un seuil modulable par l'utilisateur. Les résultats sont explicités dans la section qui suit.

3 Résultats et livrable

Afin de satisfaire l'enjeu opérationnel de notre projet avec le Groupe ADP, les résultats ont été implémentés sous Python puis visualisés sous Power BI. L'outil est dynamique, ce qui permet d'actualiser les prévisions, de visualiser les résultats à long terme, à court terme ou à moyen terme. L'utilisateur peut aussi sélectionner le faisceau d'observation et le type de mouvement qui l'intéresse.

Sur la figure 1, nous avons choisi le faisceau *International*. Nous présentons des prévisions du nombre de passagers journalier sur 3 mois, concernant les départs de l'aérogare ORLY entre le 1er janvier 2016 et le 31 mars 2016. Sur ces prévisions spécifiques, on observe que pour les indicateurs de performance RMSE et MAPE, les meilleurs modèles sont le non paramétrique puis le Lasso. Toutefois, nos modèles ne parviennent pas à être meilleurs que la prévision ADP (appelée FQM), bien que les résultats obtenus avec le modèle non paramétrique soient proches.

De manière générale, en se référant à la figure 2, on constate que les modèles ARIMA et SARIMA, qui n'utilisent pas les données du calendrier, sont moins performants. Par ailleurs, leurs prévisions sont plutôt meilleures sur des petits horizons de prévisions que pour des prévisions de long terme, en raison d'une amplification de l'erreur de prévision. En revanche, ce sont des modèles relativement robustes face à des historiques anormaux, notamment sur le faisceau *Autre UE*, dont l'historique est anormalement plat. Les modèles ARIMA et SARIMA ont également l'avantage d'être bien interprétables et très documentés, car ce

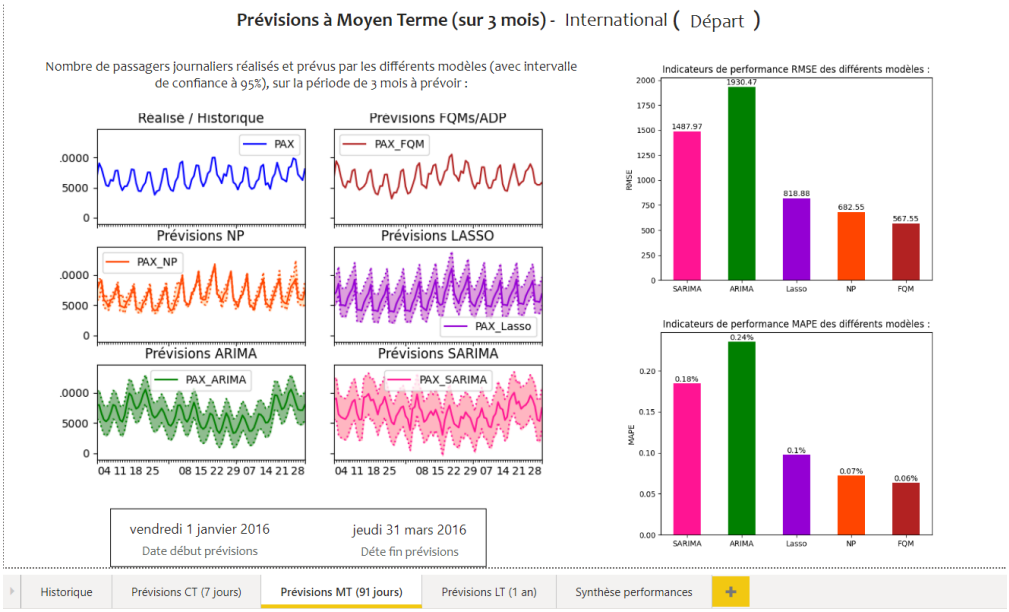

FIGURE 1 – Visualisation des prévisions à moyen terme sur Power BI

sont des modèles linéaires classiques.

Sur l'ensemble des prévisions, et sur l'ensemble des faisceaux, le modèle non paramétrique est très performant au niveau du RMSE et du MAPE et reste un modèle robuste. Ses principaux défauts sont ses intervalles de confiance peu pertinents, et sa difficile interprétabilité. D'autre part, le choix des différents paramètres utilisés a une influence sur les prédictions.

Le modèle Lasso est également plutôt performant, à l'exception du faisceau *Autre UE* sur lequel les prévisions sont mauvaises en raison d'un historique anormalement plat sur les premières années : le modèle ne parvient pas à capter les saisonnalités et variations. En revanche, c'est un modèle parcimonieux, puisqu'il permet de réduire les variables pertinentes, et facilement interprétable : on peut observer les indicatrices temporelles choisies par le modèle pour la prévision.

Conclusion

La problématique posée par le Groupe ADP dans ce sujet de statistique appliquée nous a permis de découvrir, d'implémenter et d'évaluer quatre modèles très différents mais adaptés aux données de séries temporelles. Les résultats obtenus pour l'ensemble de nos modèles ne sont malheureusement pas meilleurs que les prédictions réalisées par la méthode classique d'ADP à court terme (calcul à la semaine en reprenant la période équivalente des historiques précédents). En revanche, le modèle non paramétrique que nous avons implémenté semble constituer une méthode automatisée de prévisions satisfaisante pour répondre aux besoins des Aéroports de Paris : ce modèle permet d'obtenir des prévisions performantes à court comme à long terme, comme en témoigne la figure 2.

En revanche, aucun modèle ne semble être meilleur qu'un autre, si l'on regarde aussi la qualité des intervalles de confiance, ou encore la robustesse face à un historique dégénéré. Nous avons donc réalisé un tableau récapitulatif des avantages et inconvénients de chacun des modèles, présenté dans la table 1, afin d'aider les Aéroports de Paris ou tout autre utilisateur qui souhaiterait réaliser des prévisions à partir de nos modèles.

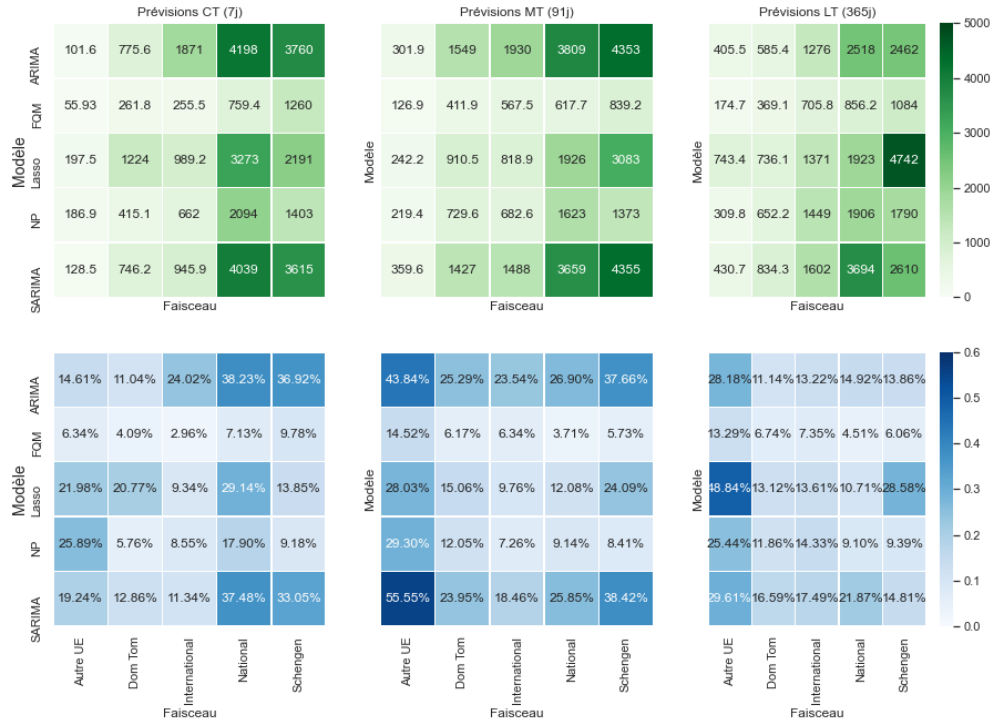

FIGURE 2 – Performance des différents modèles implémentés selon MAPE et RMSE

Critère \ Modèle	ARIMA	SARIMA	Lasso	Non-paramétrique
Précision à court et moyen terme	- -	-	+	++
Précision à long terme	-	- -	+	++
Qualité des intervalles de confiance	-	-	+	- -
Robustesse face à un historique qui s'aplatit dans le passé	+	-	- -	++
Interprétabilité du modèle	+	+	++	- -
Temps d'exécution	++	-	- - (avec IC) ++ (sans IC)	+

TABLE 1 – Tableau récapitulatif comparant la qualité des différents modèles, selon divers critères

Enfin, ce projet nous a également permis de réaliser un outil dynamique de suivi des prévisions, avec différents modèles que nous avons implémentés nous-mêmes sous langage Python. Cet outil de restitution des prévisions s'appuie sur le logiciel Power BI, qu'utilise le Groupe ADP afin de répondre à des contraintes opérationnelles.