

Projet d'expertise en statistiques et probabilités

Apprentissage statistique dans un réseau de capteurs et application à la reconstruction de la dynamique temporelle de stations de vélos en libre service

Théo Lavandier - Alexandre Leys - Mathilde Tissandier

Année 2023-2024
CMI ISI

Contents

1	Introduction	3
2	Présentation du sujet	3
2.1	Contextualisation	3
2.2	Présentation des données	4
2.3	Présentation de Notre Application Dash	5
2.3.1	Interface Utilisateur de l'Application	5
2.3.2	Sections de l'Application	6
3	Analyse statistique des données	6
3.1	La distribution des stations de vélos à Toulouse	7
3.2	Statistiques descriptives	8
3.3	Analyse des corrélations	11
3.3.1	Corrélation de Pearson	11
3.3.2	Analyse des Corrélations entre les Stations	12
3.4	Analyse en Composantes Principales (ACP)	13
3.4.1	Préparation des Données	13
3.4.2	Analyse des Résultats de l'ACP	13
3.4.3	Interprétation des Composantes Principales	14
3.4.4	Visualisation des Coefficients des Composantes	14
3.4.5	Reconstruction des Courbes Moyennes	16
3.4.6	Implications pour les Prédictions Futures	16
4	Prédiction des Activités des Stations de Vélo	17
4.1	Objectifs de Prédiction	17
4.2	Méthode d'entraînement des modèles	17
4.2.1	Extraction de Caractéristiques Temporelles	17
4.2.2	Objectifs de la Méthode	18
4.2.3	Avantages de la Méthode	18
4.2.4	Inconvénients de la Méthode	18
4.2.5	Exemple Pratique	18
4.3	Métriques d'Évaluation des Modèles	19
4.3.1	Erreur Moyenne Absolue (MAE)	19
4.3.2	Erreur Quadratique Moyenne (MSE)	19
4.4	Comparaison des Performances	20
4.4.1	Comparaisons entre Modèles	20
4.4.2	Comparaison Géographique des Performances	21
4.5	Gestion des interpolations	22
4.6	Description des Modèles de Prédiction	23
4.7	Modèle sur la moyenne par jours et par heures	23

4.7.1	Description du modèle	23
4.7.2	Avantages du modèle	24
4.7.3	Inconvénients du modèle	24
4.7.4	Exemple de prédiction	24
4.7.5	Intérêt et Utilisation	24
4.8	Modèle via reconstruction de courbe avec l'ACP	25
4.8.1	Description du modèle	25
4.8.2	Avantages du modèle	25
4.8.3	Inconvénients du modèle	25
4.8.4	Exemple de prédiction	25
4.8.5	Intérêt et Utilisation	26
4.9	Modèle via Régression Linéaire Multiple	26
4.9.1	Description du modèle	26
4.9.2	Avantages du modèle	27
4.9.3	Inconvénients du modèle	27
4.9.4	Exemple de prédiction	27
4.9.5	Intérêt et Utilisation	28
4.10	Modèle via Forêts Aléatoires	28
4.10.1	Description du modèle	28
4.10.2	Avantages du modèle	29
4.10.3	Inconvénients du modèle	29
4.10.4	Exemple de prédiction	29
4.10.5	Intérêt et Utilisation	30
4.11	Modèle XGBoost	30
4.11.1	Description du modèle	30
4.11.2	Avantages du modèle	31
4.11.3	Inconvénients du modèle	31
4.11.4	Exemple de prédiction	32
4.11.5	Intérêt et Utilisation	32
4.12	Modèle XGBoost avec ACP	32
4.12.1	Description du modèle	32
4.12.2	Avantages du modèle	33
4.12.3	Inconvénients du modèle	33
4.12.4	Exemple de prédiction	33
4.12.5	Intérêt et Utilisation	34
5	Comparaison des modèles	34
5.1	Performances globales des modèles	34
5.1.1	Analyse détaillée des modèles	34
5.1.2	Interprétation des résultats	35
5.1.3	Analyses locales	36
5.2	Analyse Géographique des métriques	39
6	Conclusion du projet et observations	40
6.1	Première partie : Analyse statistique des données	40
6.2	Deuxième partie : Prédictions et comparaison des modèles	40
6.3	Réflexions et perspectives d'amélioration	40
7	Méthodes envisagées pour aller plus loin	41
7.1	Méthode d'entraînement des modèles	41
7.1.1	La Méthode Rolling Window	41
7.1.2	La Méthode de Lagging	41
7.2	Conclusion	42

1 Introduction

Dans le cadre de notre projet d'expertise en statistiques et probabilités de Master 1, nous avons décidé de nous concentrer sur le sujet suivant : Apprentissage statistique dans un réseau de capteurs et application à la reconstruction de la dynamique temporelle de stations de vélos en libre service. Dans notre étude, nous nous concentrerons sur les données de Toulouse du 1er avril 2016 au 27 septembre 2016. Chaque heure, une observation est effectuée, nous permettant ainsi de suivre en temps réel la répartition des vélos dans les stations.

Pour effectuer notre analyse de données et créer nos modèles de prédition, nous avons utilisé le langage de programmation Python. Tous nos résultats sont présentés sous la forme d'une application Dash [5], qui est entièrement interactive.

L'application est conçue pour suivre la même structure que ce rapport, avec une section dédiée à l'analyse des données suivie d'une section de prédition basée sur différents modèles prédictifs. Nous vous recommandons fortement d'utiliser l'application en parallèle de la lecture de ce rapport afin de mieux comprendre et visualiser les résultats de notre étude.

Dash offre la possibilité d'explorer de manière interactive les données, proposant des visualisations dynamiques et la possibilité d'analyser en profondeur les résultats de nos analyses et prédictions. En utilisant cette application, vous pourrez :

- Visualiser les données de base sur les stations de vélos, y compris leur répartition géographique et la disponibilité des vélos.
- Explorer les matrices de distance entre les stations pour comprendre les relations spatiales.
- Analyser les observations horaires de la disponibilité des vélos sur une période de six mois.
- Interagir avec les modèles prédictifs pour voir comment différents facteurs influencent la disponibilité des vélos.

Nous avons choisi Dash pour sa capacité à créer des applications web interactives et faciles à utiliser, ce qui facilite la compréhension et l'analyse des données complexes. Pour l'accès à l'application, nous l'hébergeons sur Google Cloud afin de garantir son accès à tout moment. Nous espérons que cette application vous aidera à mieux appréhender les dynamiques de la disponibilité des vélos dans les stations de Toulouse.

Vous pouvez retrouver notre application sur le site web suivant: <https://tlavandierter.nw.r.appspot.com/>

2 Présentation du sujet

2.1 Contextualisation

VélôToulouse est un système de vélos en libre-service situé à Toulouse, en France. [9] Lancé en 2007, il est devenu un élément fondamental de la mobilité dans la ville. Les habitants et les visiteurs peuvent profiter de la commodité, de l'accessibilité et de l'abordabilité des vélos pour se déplacer dans toute la ville.

VélôToulouse a été créée par le conseil de la ville de Toulouse [1], en collaboration avec l'opérateur JCDecaux, qui est en charge des vélos en libre-service. Le système est en constante croissance et en évolution pour s'adapter à la demande.

Le programme VélôToulouse fonctionne comme suit:

- **Inscription** : L'inscription des utilisateurs à VélôToulouse se fait via le site web, l'application mobile ou les bornes d'agences situées près des stations.
- **Location de vélos** : Après l'inscription, tous les utilisateurs peuvent louer un vélo dans n'importe quelle station de vélos en présentant une carte d'abonnement ou en suivant les instructions de l'application.
- **Paiement** : Les prix du programme du vélo comprennent diverses offres : un abonnement court terme ou long terme et un taux horaire des utilisateurs ad hoc.

- **Utilisation de vélos :** En outre, tous les utilisateurs peuvent utiliser les vélos à leur disposition pour des promenades à court ou à long terme dans la ville et les déposer dans n'importe quelle station VéloToulouse après avoir terminé.

VéloToulouse a connu une croissance remarquable depuis ses débuts modestes en 2007, avec 600 vélos répartis dans 60 stations du centre-ville. Le service s'est étendu en janvier 2008 avec 1 470 vélos dans 135 stations, proposant ainsi une solution de transport pratique et respectueuse de l'environnement aux habitants et aux visiteurs de la ville rose.

L'importance de ce système de vélos en libre-service a été rapidement saisi par la mairie de Toulouse, qui a annoncé des projets ambitieux pour son développement. Dans le but ambitieux de créer 10 nouvelles stations chaque semaine, le réseau a connu une expansion significative, atteignant désormais 283 stations et plus de 2 400 vélos.

Au fil du temps, le service a connu des améliorations afin de mieux satisfaire les attentes des utilisateurs. En juillet 2011, on a étendu les horaires de location afin de garantir une disponibilité 24h/24, ce qui offre une plus grande souplesse aux utilisateurs.

En 2013, 30 nouvelles stations et 300 vélos ont été ajoutés, ce qui a permis d'avoir un total de 283 stations et 2 600 vélos. Cependant, certains vélos sont inservibles en raison de l'usure naturelle et parfois de l'intrusion. Les utilisateurs ont mis en place un système informel pour signaler ces vélos manquants en retournant la selle, ce qui facilite leur identification.

Au départ, le service était restreint à la commune de Toulouse, mais des projets ambitieux d'expansion ont été annoncés pour s'étendre à d'autres communes voisines comme Blagnac, Colomiers, Tournefeuille, L'Union et Ramonville-Saint-Agne à partir de 2024. Toutefois, la municipalité de Balma a choisi de mettre en place son propre réseau, soulignant ainsi la variété des approches locales en matière de mobilité durable.

En quête d'une mobilité toujours plus respectueuse de l'environnement, la municipalité de Toulouse prévoit également de transformer la moitié du parc de vélos en vélos électriques à partir de 2024, proposant ainsi une solution encore plus respectueuse de l'environnement et efficace pour se déplacer dans la ville. En résumé, VéloToulouse demeure un acteur clé dans la vie quotidienne des habitants de Toulouse, tout en représentant un exemple de succès en matière de mobilité urbaine durable.

Le vélo occupe une part croissante dans les modes de déplacements urbains. Au cœur des villes, les dispositifs de vélo en libre-service contribuent au développement des mobilités douces et de l'intermodalité. La satisfaction des utilisateurs est au rendez-vous comme le montrent les résultats des enquêtes menées récemment à Toulouse. En moyenne, chaque vélo est utilisé 6 fois par jour en semaine. Selon les données, 94% des personnes interrogées ont l'intention de se réabonner à la fin de leur abonnement et 95% aimeraient recommander le service à un proche.

2.2 Présentation des données

Afin de réaliser notre analyse de données et d'effectuer des prédictions sur la présence de vélos dans les stations de vélo de Toulouse, nous disposons de plusieurs jeux de données. On divise ces informations en trois fichiers CSV :

- **coordinates_toulouse.csv** : Contient les noms des stations de Toulouse ainsi que leurs coordonnées géographiques.
- **distance_toulouse.csv** : Matrice des distances (à vol d'oiseau) entre les différentes stations de vélos.
- **X_hour_toulouse.csv** : Observations de la proportion de vélos disponibles dans les différentes stations de vélos. Chaque ligne représente une observation horaire pendant environ six mois (du 1er avril 2016 au 27 septembre 2016).

Vous pouvez aussi retrouver l'explication des fichiers dans ce tableau récapitulatif :

Fichier	Description
coordinates_toulouse.csv	Contient les noms des stations de vélo ainsi que leurs coordonnées géographiques. <ul style="list-style-type: none"> • <i>station_name</i> : Nom de la station de vélo. • <i>latitude</i> : Latitude de la station. • <i>longitude</i> : Longitude de la station.
distance_toulouse.csv	Représente une matrice 185×185 où chaque cellule (i, j) indique la distance à vol d'oiseau entre la station i et la station j .
X_hour_toulouse.csv	Regroupe les observations de la proportion de vélos disponibles dans les différentes stations de vélos. Chaque ligne représente une observation horaire. <ul style="list-style-type: none"> • Chaque ligne : Une observation horaire. • Les colonnes : Identifiants des stations de vélos.

Nos données couvrent un total de 185 stations de vélos.

Ces données nous permettront de modéliser et de prédire la disponibilité des vélos dans les stations de Toulouse en fonction de différents facteurs temporels et géographiques. Nous commencerons par une exploration descriptive des données avant de passer à des modèles prédictifs plus sophistiqués.

2.3 Présentation de Notre Application Dash

Dans le cadre de notre projet, nous avons développé une application interactive utilisant Dash, conçue pour permettre une exploration approfondie et intuitive des données des stations de vélo de la ville de Toulouse. Cette application sert de plateforme centrale pour visualiser et interagir avec les résultats de nos analyses statistiques et prédictives.

2.3.1 Interface Utilisateur de l'Application

Lorsque vous lancez l'application, vous êtes accueillis par la page d'accueil, qui se présente comme suit :

Page d'accueil de l'application

L'interface utilisateur de notre application est divisée en deux grandes parties, accessibles via la barre de navigation située en haut à droite de l'écran.

2.3.2 Sections de l'Application

Les deux sections principales de l'application sont :

- **Statistiques Descriptives** : Cette section est subdivisée en plusieurs sous-sections qui fournissent des analyses détaillées des données :

1. *Statistiques Générales* : Affichage des statistiques de 'base' telles que la moyenne, la médiane, les quantiles,... affichés dans des graphiques que nous avons analysés et qui donnent un aperçu initial des tendances générales dans les données.
2. *Analyse des Corrélations* : Exploration des relations entre différentes variables pour identifier les facteurs qui influencent le plus sur l'utilisation des vélos.
3. *ACP (Analyse en Composantes Principales)* : Visualisation des principaux facteurs influençant la variabilité des données à travers une réduction dimensionnelle.

- **Prédictions** : Cette section inclut des outils pour prédire et analyser l'activité future des stations de vélo :

1. *Prédictions* : Offre des modèles de prévision de l'activité des stations pour le jour suivant ou la semaine à venir en fonction de votre sélection sur l'application.
2. *Comparaison des Modèles* : Permet de comparer différents modèles de prédiction selon leur performance, en utilisant des métriques telles que le MAE et le MSE.
3. *Comparaisons Géographiques* : Visualise les résultats des modèles pour chaque station sur une carte, mettant en évidence les zones où les prédictions sont les plus précises pour nos divers modèles de prédiction.

Nous vous invitons à naviguer dans notre application pendant que vous lisez ce rapport afin de mieux comprendre nos analyses sur les stations de vélos de la ville de Toulouse. La conception de chaque partie de l'application vise à être à la fois informative et conviviale, assurant ainsi une expérience utilisateur enrichissante et éducative.

3 Analyse statistique des données

Avant de procéder à des prédictions sur nos données, il est nécessaire de réaliser une analyse statistique complète de tout notre jeu de données. Cette opération, qui représente une première étape fondamentale, est nécessaire pour comprendre les propriétés propres au jeu de données, pour détecter les tendances et les relations qui les caractérisent et qui, à leur tour, pourront influencer nos futures prédictions.

Afin de réaliser cela, nous commençons par explorer le jeu de données en s'assurant qu'il n'y a pas de valeurs manquantes, de valeurs aberrantes, et d'avoir une idée générale de la structure des données. Cette étape nous donne les données initiales dont on a besoin pour pouvoir rentrer dedans et faire les requêtes.

De plus, nous effectuerons des statistiques descriptives relatives aux stations de vélo. Il s'agira d'un certain nombre d'analyses descriptives, telles que les histogrammes ou les boîtes à moustaches. Ces types de visualisation nous permettront de mieux comprendre la distribution des données, les tendances, la variabilité et les anomalies. Par exemple, un histogramme peut être utilisé pour montrer la distribution de vélos disponibles à différents moments de la journée, et une boîte à moustaches pour déterminer si les stations diffèrent sensiblement en termes de vélos disponibles.

Par la suite, nous réaliserons une analyse de corrélation entre les stations. L'objectif de cette analyse est de déterminer s'il existe des relations linéaires entre les disponibilités des différentes stations. En calculant les coefficients de corrélation de Pearson, nous allons chercher à déterminer les paires de stations qui présentent des comportements similaires ou différents. Cette étape est importante pour détecter des patterns potentiels qui peuvent être utilisés pour améliorer les modèles de prédiction.

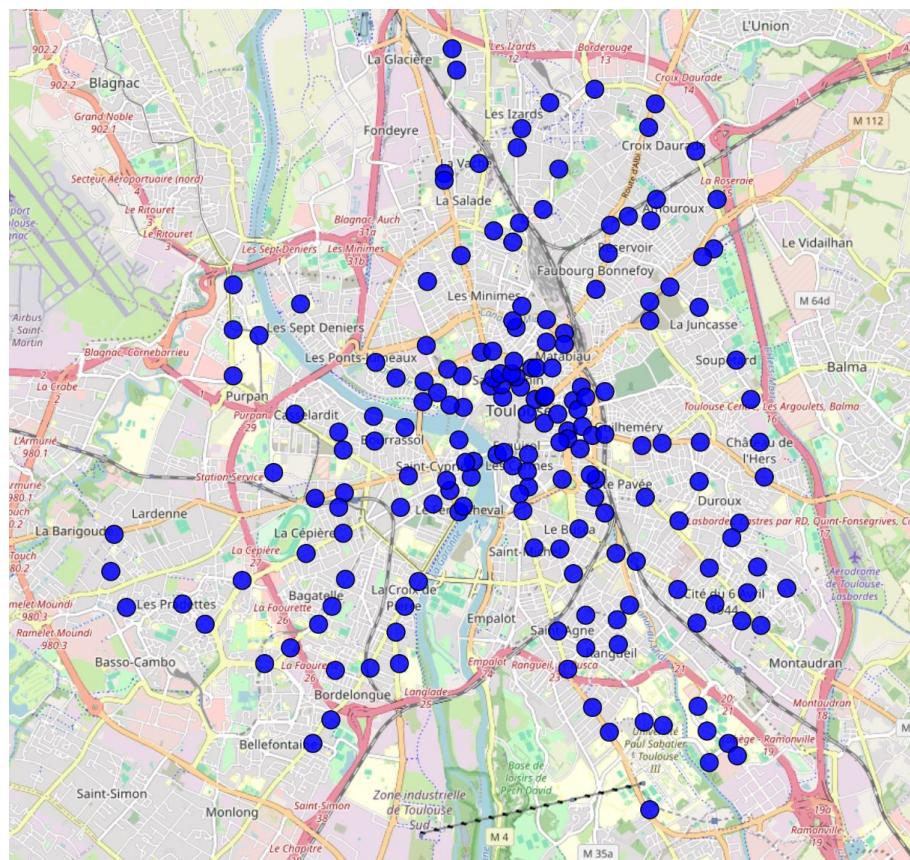
Finalement, une analyse en composantes principales (ACP) sera réalisée afin de rechercher des comportements caractéristiques parmi les stations. L'ACP est une méthode qui permet de diminuer la

dimensionnalité en transformant les variables initiales en un ensemble de variables non corrélées, connues sous le nom de composantes principales. En observant les stations sur les principales composantes, nous serons en mesure de repérer des groupes de stations qui présentent des similitudes. Grâce à cette analyse, nous pourrons également envisager de concevoir des modèles en diminuant le nombre de variables tout en préservant l'essentiel des données.

Toutes ces analyses permettront de mieux comprendre les données et de nous préparer à créer des modèles de prédiction plus solides et précis. Nous pourrons mieux représenter les dynamiques de la disponibilité des vélos en libre-service en connaissant les distributions des données, les relations entre les stations et les structures sous-jacentes.

3.1 La distribution des stations de vélos à Toulouse

Avant de commencer notre analyse statistique, étudions la distribution des stations de vélos de la ville de Toulouse.



Distribution des stations de vélos dans la ville de Toulouse

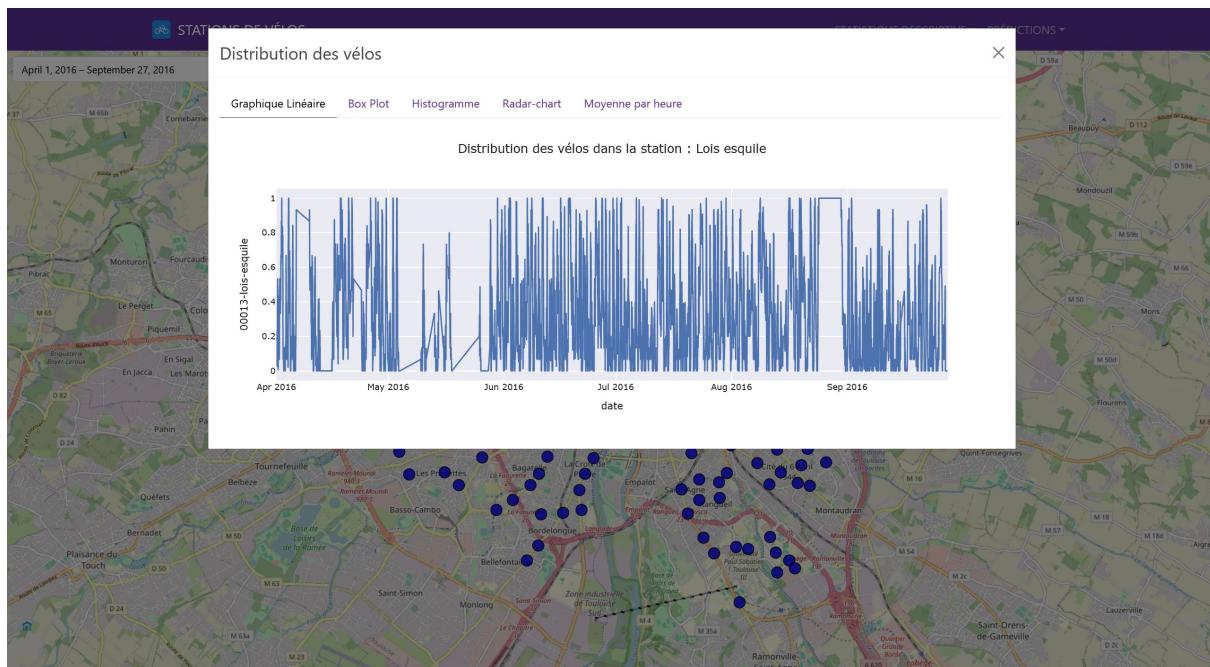
Dans le centre-ville de Toulouse, les stations de vélos sont très nombreuses, ce qui démontre une intégration étroite du service de vélopartage dans la vie quotidienne. On constate qu'il y a une forte densité autour des places centrales, des rues commerçantes et des lieux de vie animés, facilitant ainsi l'accès aux commerces, cafés, restaurants et sites culturels. Les stations semblent être placées de façon stratégique le long des grandes artères de la ville et à proximité des grands carrefours, permettant ainsi un accès facile aux habitants et aux visiteurs qui circulent le long de ces voies. C'est le cas des allées Jean Jaurès, du boulevard de Strasbourg et de la rue de Metz. Les stations se trouvent également dans les quartiers résidentiels, à l'extérieur de l'hypercentre, bien que moins fréquemment. Cela témoigne d'une volonté d'étendre l'utilisation du service de vélopartage au-delà du centre-ville, offrant ainsi une solution de transport pratique pour les déplacements quotidiens ou les activités de loisirs. Plusieurs stations sont situées à proximité d'espaces verts, de sites touristiques, d'établissements scolaires et de complexes sportifs, ce qui laisse entendre que le réseau de vélopartage est destiné à satisfaire les besoins à la fois des habitants et des touristes, mais aussi de la communauté étudiante. La répartition des stations de vélos en libre-service témoigne d'une grande importance accordée à l'accessibilité pour les usagers. En effet, les stations sont situées à des intervalles réguliers dans toute la ville, ce qui facilite grandement la

recherche d'une station à proximité, peu importe l'endroit où l'on se trouve. Cette disposition permet ainsi de favoriser l'utilisation de ce mode de transport écologique et économique.

En somme, la carte des stations de vélos en libre-service à Toulouse illustre une infrastructure de vélopartage bien pensée et adaptée aux besoins de la ville. Les stations sont réparties de manière stratégique, tant dans le centre-ville dense que dans les zones périphériques, offrant ainsi une couverture étendue et pratique pour les déplacements à vélo. Cette disposition permet de répondre aux besoins de tous les usagers, peu importe leur lieu de résidence ou de travail.

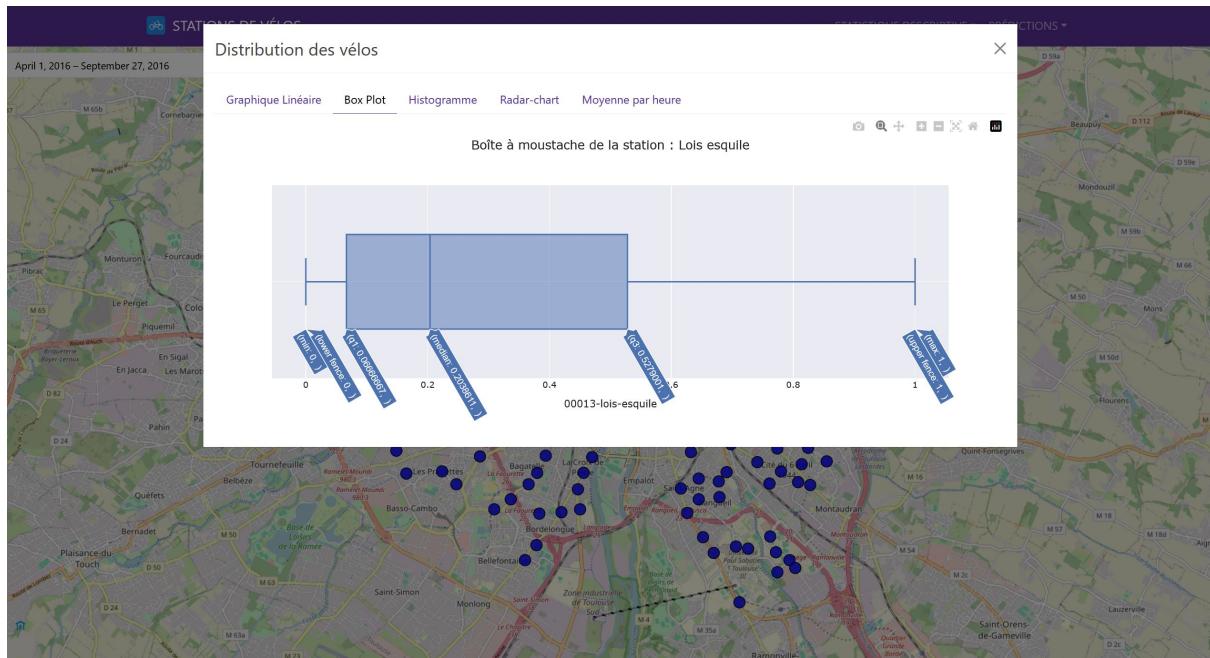
3.2 Statistiques descriptives

La première étape de notre application consiste à effectuer des analyses statistiques sur nos jeux de données de vélos en libre-service. Pour ce faire, nous avons accès à une variété de graphiques pour chaque station, tels que des boîtes à moustaches et des histogrammes. Ces outils visuels nous ont permis de mieux comprendre la répartition de l'utilisation des vélos dans la ville. Nous avons ainsi constaté que les stations situées dans le centre-ville sont en moyenne plus fréquentées et donc plus remplies, tandis que celles en périphérie ont tendance à être moins sollicitées et donc moins remplies. Dans cette première partie du rapport, nous avons décidé de vous présenter la station *00013-lois-esquille*.



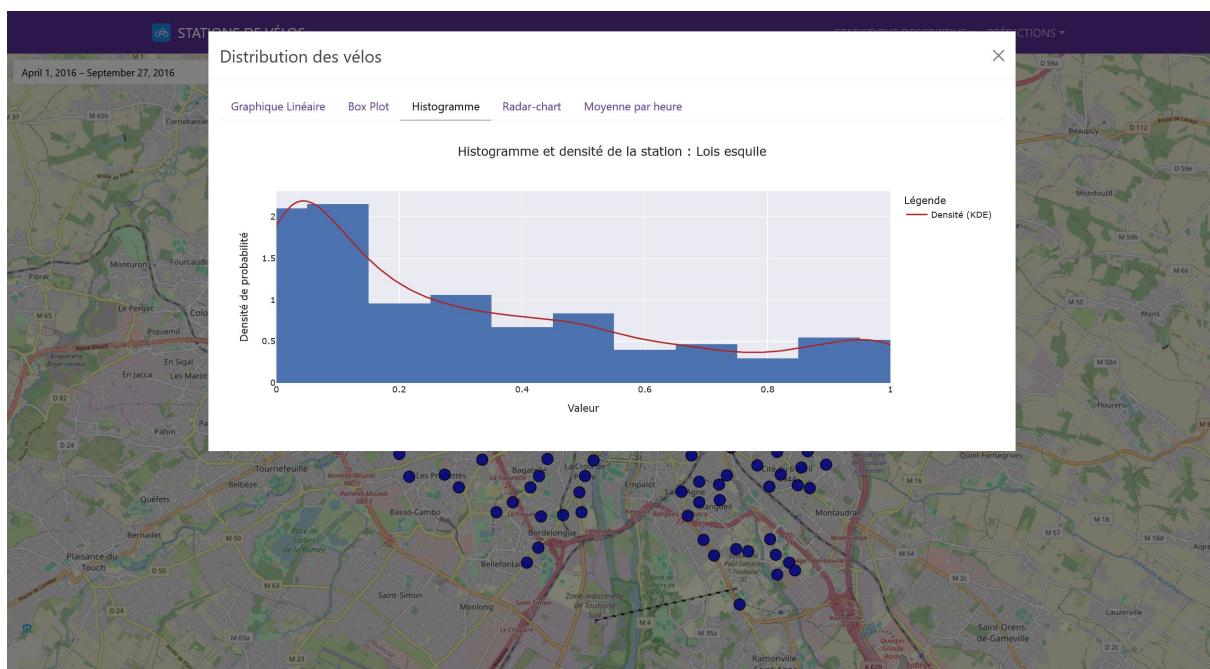
Graphique linéaire de la station 'Lois Esquille' sur les données du 1er avril 2016 au 27 septembre 2016 dans la ville de Toulouse

Le schéma linéaire présenté offre une représentation graphique détaillée de la disponibilité des vélos dans la station "Lois Esquille" sur une période de plusieurs mois. En examinant les variations quotidiennes, il est possible d'identifier des tendances et des fluctuations importantes dans la disponibilité des vélos. Les niveaux élevés et bas réguliers révèlent des schémas de demande, tels que des périodes d'utilisation intense aux heures de pointe et des baisses pendant les périodes de faible demande. La réalisation de cette analyse approfondie permet de mieux comprendre les habitudes d'utilisation des usagers et d'améliorer la gestion et la répartition des vélos dans le réseau. En ajustant la disponibilité des vélos en fonction des besoins réels des usagers, nous pouvons garantir une disponibilité optimale et offrir un service de vélopartage plus efficace et plus pratique.



Boîte à moustaches de la station 'loi Esquille' sur les données du 1er avril 2016 au 27 septembre 2016 dans la ville de Toulouse

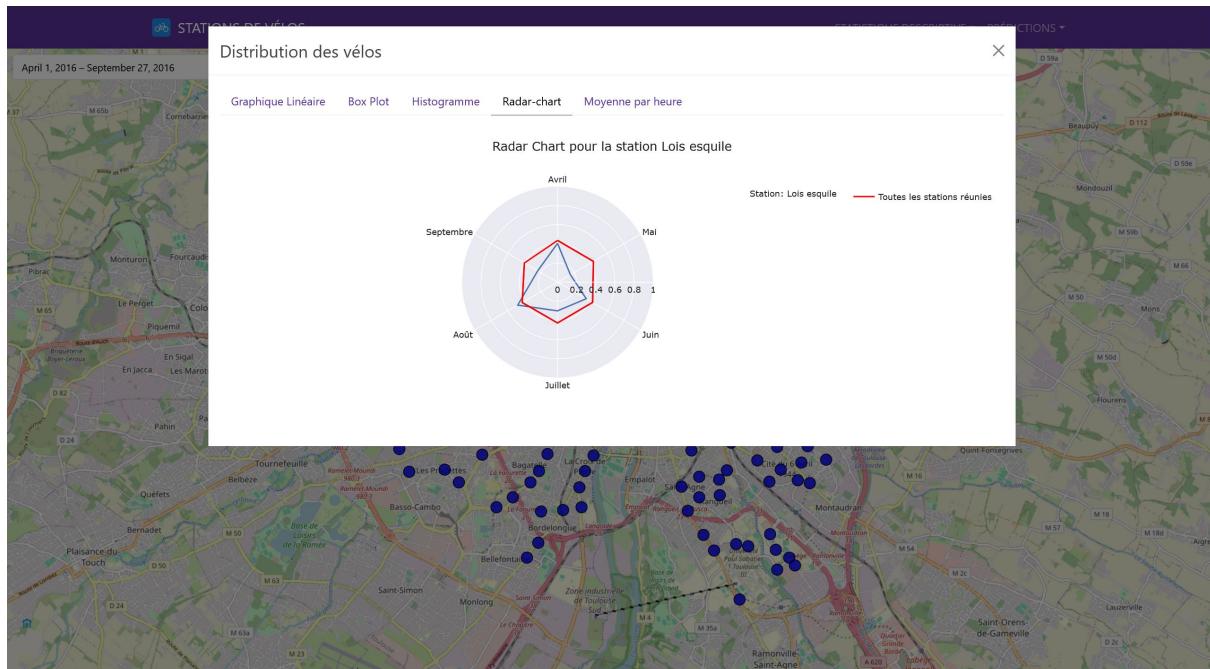
Les boîtes à moustaches fournissent une visualisation simple des quartiles du nombre de vélos disponibles dans une station donnée. En examinant la médiane, nous pouvons déterminer si une station a tendance à être plus remplie ou plus vide en moyenne. Cette analyse peut révéler des tendances intéressantes, comme la prévalence de stations plus fournies dans le centre-ville par rapport à celles en périphérie, qui ont tendance à être moins approvisionnées en vélos. Dans cet exemple, la station se trouve dans le centre-ville, on a donc une médiane proche de 0.2. Cela signifie que la station à tendance à être plus vide que remplie, ce qui est en accord nos précédentes suppositions.



Histogramme de la station 'loi Esquille' sur les données du 1er avril 2016 au 27 septembre 2016 dans la ville de Toulouse

Les histogrammes offrent une représentation visuelle de la distribution des données. En étudiant l'histogramme de la station Loi Esquille, nous pouvons constater que la majorité des observations se

concentrent autour de zéro. Cela suggère que cette station a une tendance à être plus souvent vide que pleine (comme on vient de le voir avec la boîte à moustaches).

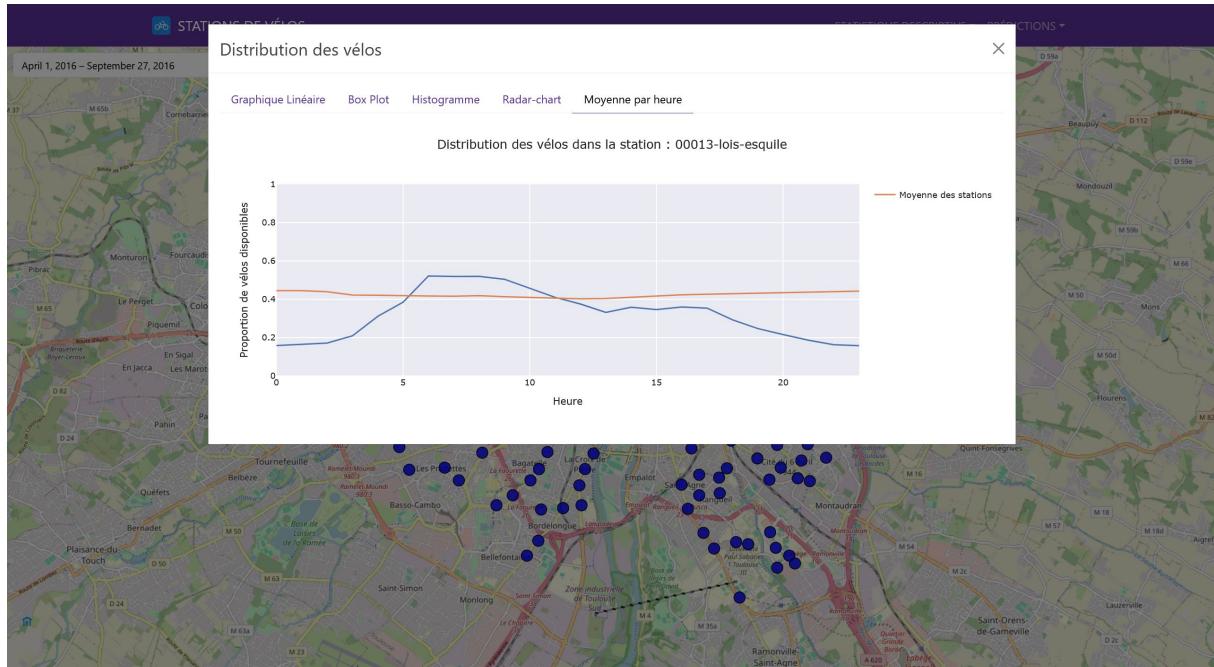


Radar-chart de la station 'Lois Esquille' et toutes les stations sur les données du 1er avril 2016 au 27 septembre 2016 dans la ville de Toulouse

Le graphique radar offre une visualisation comparative de la distribution des vélos pour la station "Lois Esquille" par rapport à la moyenne de toutes les stations sur plusieurs mois. Chaque axe du radar correspond à un mois de 2016, ce qui permet de comparer les variations mensuelles de la disponibilité des vélos pour la station donnée par rapport à l'ensemble des stations de la ville.

En analysant cette représentation, il est possible de repérer des disparités saisonnières dans la disponibilité des vélos. Par exemple, des différences importantes entre les mois estivaux et hivernaux peuvent témoigner de modifications des habitudes d'utilisation en fonction des conditions météorologiques ou des périodes de vacances. En observant les variations par rapport à la moyenne globale (ligne rouge), on peut déterminer si la station "Lois Esquille" présente des tendances similaires ou différentes des autres stations du réseau.

Ces analyses nous permettent de saisir les comportements individuels des stations que nous étudions.



Graphique de la distribution moyenne de vélos de la station 'loi Esquelle' comparée à celle de toutes les stations sur les données du 1er avril 2016 au 27 septembre 2016 dans la ville de Toulouse

En examinant ce graphique, on peut voir les fluctuations de la disponibilité des vélos tout au long de la journée. Par exemple, il y a une forte augmentation de la disponibilité entre 5 heures et 10 heures du matin, ce qui peut indiquer que les vélos sont moins utilisés pendant cette période. À l'inverse, la baisse graduelle après 10 heures peut suggérer que de plus en plus de gens utilisent les vélos pour se rendre à leur travail ou à leurs activités.

Après avoir comparé la ligne de cette station particulière à la moyenne des stations, il est possible de repérer des comportements atypiques ou en accord avec la tendance principale du réseau. En cas de disponibilité constante en dessous ou au-dessus de la moyenne dans cette station, cela peut être dû à des disparités dans la demande locale ou à des problèmes spécifiques de gestion des vélos dans cette station.

En résumé, cette partie de notre application permet d'analyser et de comprendre les comportements spécifiques de chaque station Vélo Toulouse, fournissant ainsi des informations précieuses pour optimiser l'utilisation et la gestion du service de vélos en libre-service de la ville.

3.3 Analyse des corrélations

L'analyse des corrélations est une étape cruciale pour comprendre les relations entre les différentes stations de vélos de la ville de Toulouse. Pour ce faire, nous avons choisi d'utiliser le coefficient de corrélation de Pearson.

3.3.1 Corrélation de Pearson

Le coefficient de corrélation de Pearson, souvent noté r , est une mesure de la force et de la direction de la relation linéaire entre deux variables quantitatives. Il est défini par la formule suivante :

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

où x_i et y_i sont les valeurs des variables, et \bar{x} et \bar{y} sont leurs moyennes respectives.

La valeur de r varie entre -1 et 1. Un coefficient de corrélation de 1 indique une relation linéaire positive parfaite, -1 une relation linéaire négative parfaite, et 0 indique l'absence de relation linéaire [7, 8].

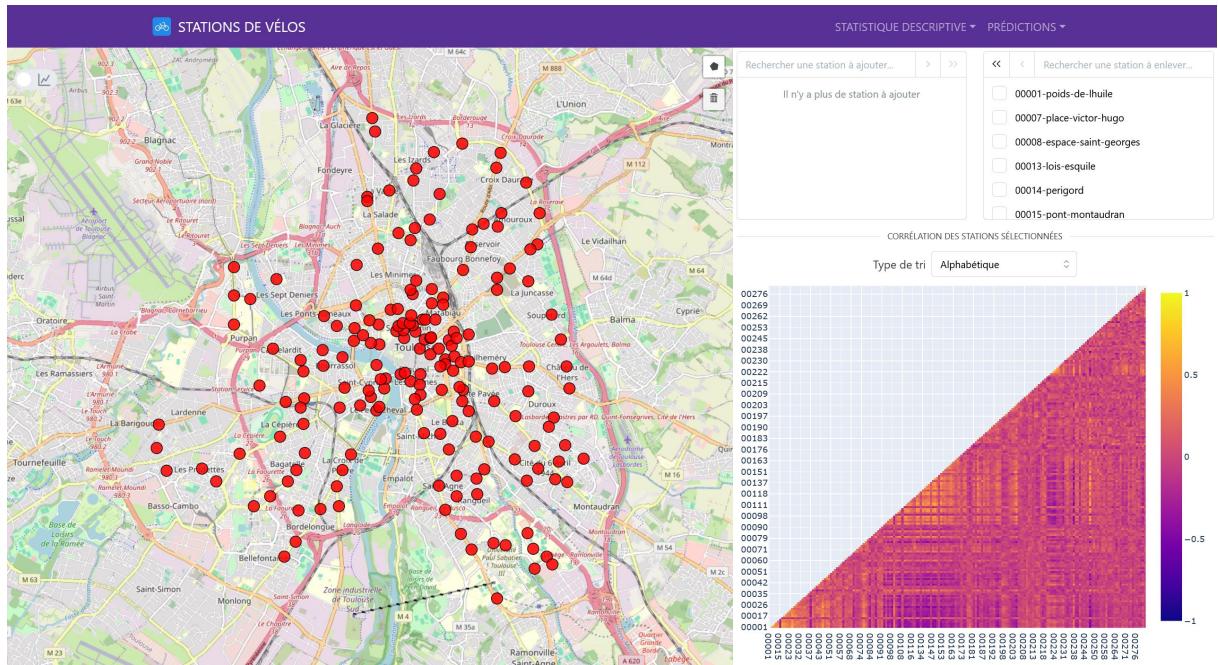
Nous avons décidé d'utiliser le coefficient de corrélation de Pearson dans notre analyse, car il est très efficace pour mesurer les relations linéaires entre les variables continues. Cela correspond parfaitement à notre besoin d'analyser les variations conjointes de la disponibilité des vélos entre les différentes stations [4]. Dans notre cas, les données de disponibilité des vélos sont comprises entre 0 et 1, représentant le pourcentage de vélos disponibles par rapport à la capacité totale de chaque station. Ce type de données

continues convient parfaitement à l'utilisation du coefficient de corrélation de Pearson, qui est conçu pour capturer les relations linéaires entre des variables quantitatives continues. Ainsi, en utilisant cette mesure, nous pouvons obtenir une compréhension précise des corrélations entre les stations, facilitant ainsi l'identification de tendances et de comportements partagés.

3.3.2 Analyse des Corrélations entre les Stations

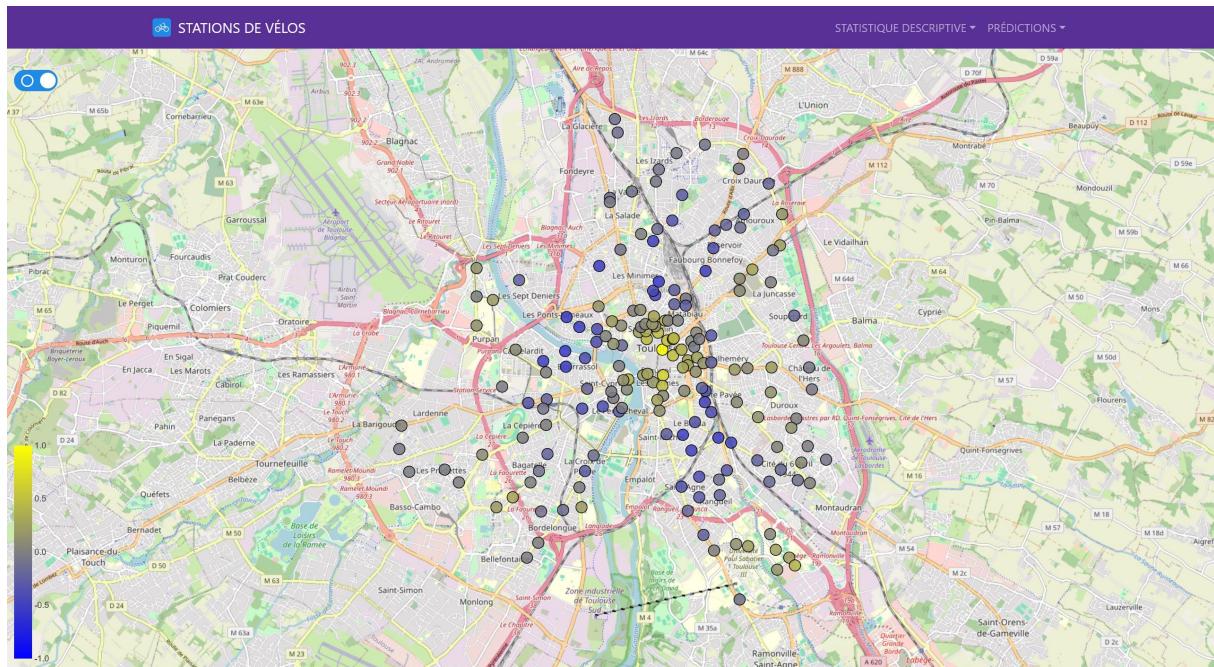
Cette section de notre analyse comprend deux graphiques principaux :

- Une carte interactive accompagnée d'une matrice de corrélation.
- Une carte représentant les corrélations entre une station donnée et les autres stations, avec une échelle de couleur pour visualiser ces corrélations sur la carte.



Carte de la distribution des stations de vélos dans la ville de Toulouse avec sa matrice de corrélation sur les données du 1er avril 2016 au 27 septembre 2016

Sur cette carte interactive, les utilisateurs peuvent sélectionner les stations à analyser directement depuis la carte ou la liste des stations située au-dessus de la matrice. Ce graphique permet de constater qu'en général, les stations proches les unes des autres présentent une corrélation positive.



Carte des corrélations de la station 'Loi Esquille' avec les autres stations dans la ville de Toulouse sur les données du 1er avril 2016 au 27 septembre 2016

Sur cette deuxième carte, nous avons affiché les corrélations entre la station Loi Esquille, située dans le centre-ville, et les autres stations de Vélo Toulouse. Nous avons remarqué que les stations proches de Loi Esquille ont tendance à avoir une corrélation positive avec celle-ci, ce qui signifie que lorsque le nombre de vélos disponibles à Loi Esquille augmente, il en va de même pour les stations avoisinantes et inversement. En revanche, les stations situées en périphérie ont généralement une corrélation négative avec Loi Esquille, ce qui signifie que lorsque le nombre de vélos disponibles à Loi Esquille augmente, il diminue dans les stations périphériques, et vice versa.

En sélectionnant une station en périphérie, nous constatons qu'elle est généralement positivement corrélée avec les autres stations en périphérie, et négativement corrélée avec les stations du centre-ville. Cette observation suggère une relation de corrélation négative entre les stations du centre et celles de la périphérie, et une corrélation positive entre les stations situées dans la même zone géographique.

Ces analyses de corrélation nous permettent de mieux comprendre les relations entre les différentes stations de Vélo Toulouse et d'identifier des tendances et des schémas dans l'utilisation et la disponibilité des vélos en libre-service à travers la ville.

3.4 Analyse en Composantes Principales (ACP)

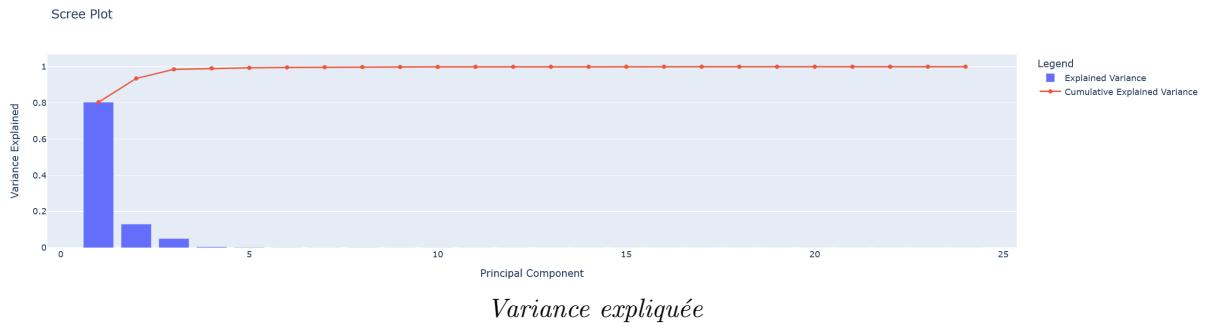
Dans le but d'analyser les dynamiques entre les différentes stations de vélos de Toulouse, nous avons effectué une Analyse en Composantes Principales (ACP). Cette méthode statistique permet de réduire la dimensionnalité des données tout en conservant l'essentiel de l'information, facilitant ainsi l'identification des tendances et des corrélations entre les stations.

3.4.1 Préparation des Données

Pour réaliser cette ACP, nous avons préalablement traité les données en calculant les moyennes horaires pour chaque station. Cela nous a permis d'obtenir un jeu de données composé de 24 colonnes (représentant les 24 heures de la journée) et de k lignes (avec k étant le nombre total de stations). Il est crucial de transformer les données afin de saisir les fluctuations horaires de la disponibilité des vélos.

3.4.2 Analyse des Résultats de l'ACP

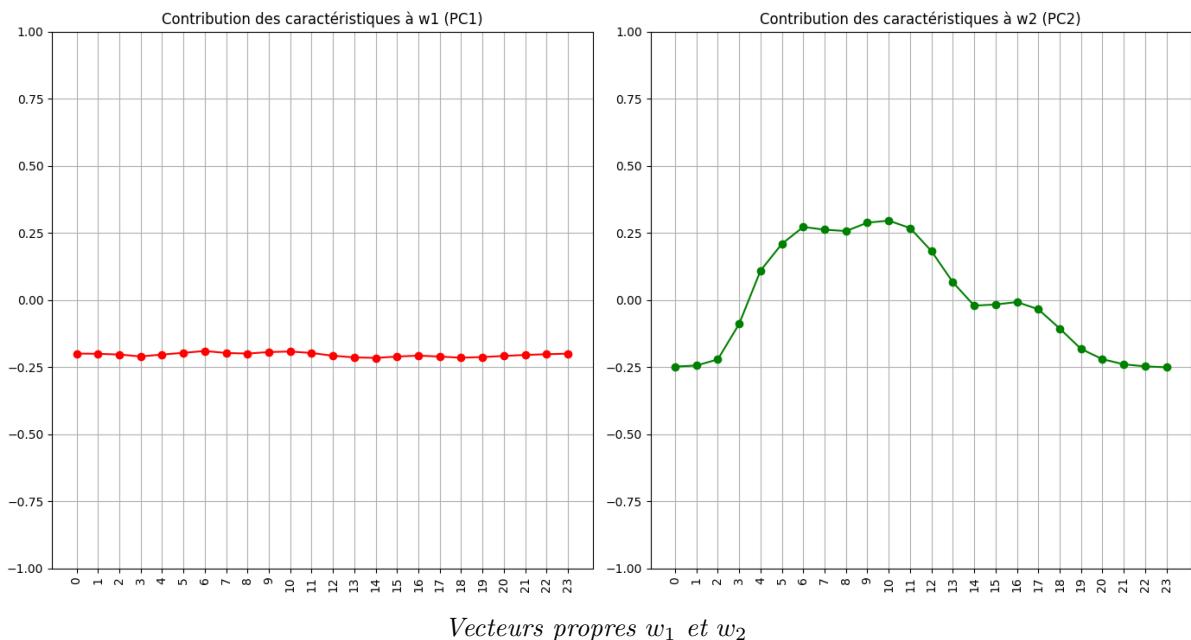
L'ACP nous permet d'identifier et de visualiser les tendances et les corrélations entre les différentes stations de vélos à différents moments de la journée, offrant ainsi une meilleure compréhension des dynamiques d'utilisation du service Vélo Toulouse.



En analysant cet éboulis des valeurs propres, nous pouvons voir que les deux premières composantes de l'ACP expliquent plus de 90% de la variance de nos données. Cela signifie que la majorité des informations présentes dans les données originales peut être représentée par ces deux composantes principales.

3.4.3 Interprétation des Composantes Principales

Pour comprendre les dynamiques capturées par les premières composantes principales, nous examinons les vecteurs propres associés :

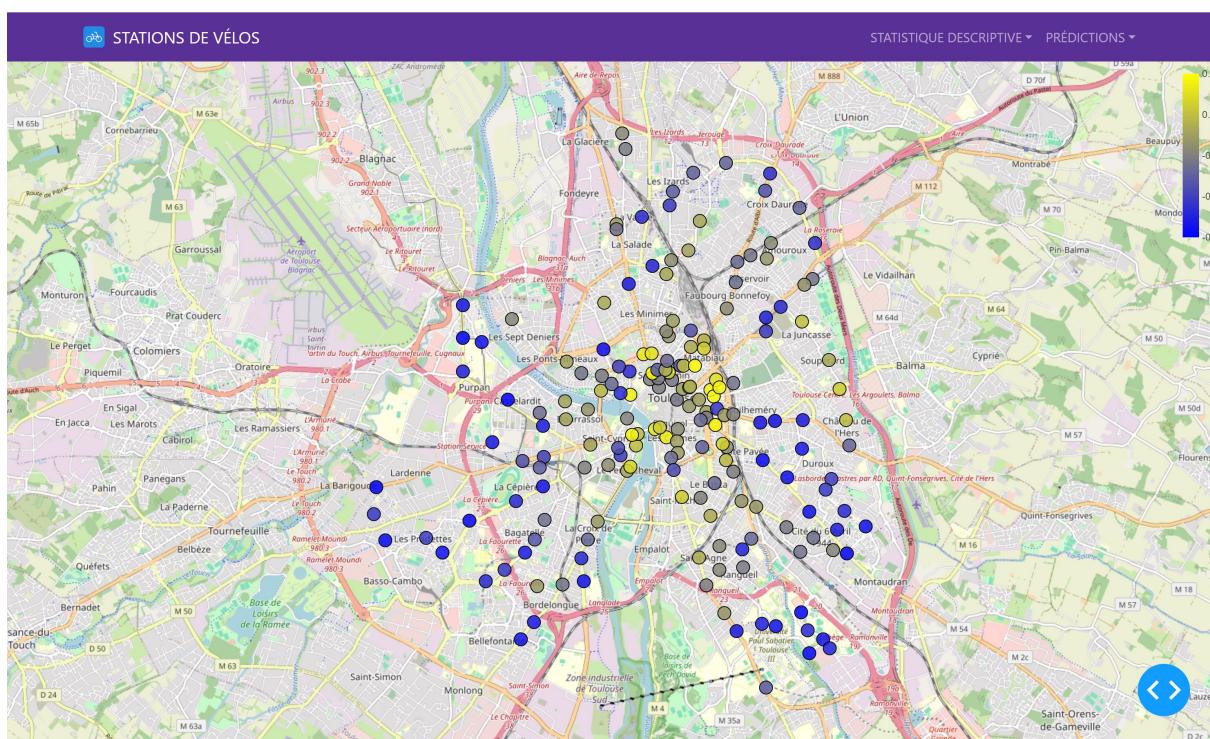
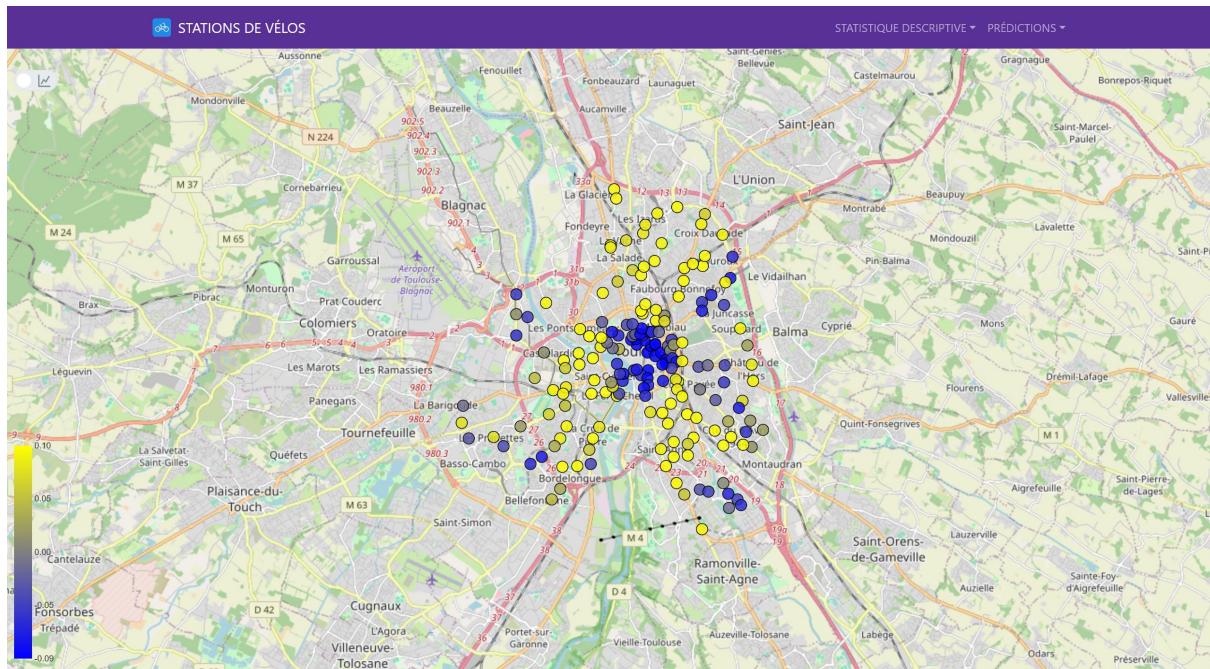


Le premier vecteur propre, étant constant et négatif sur les 24 valeurs, indique qu'une station avec un coefficient positif sur la première composante se vide tout au long de la journée. Inversement, pour une station avec un coefficient négatif, la station se remplit.

Le deuxième vecteur propre, n'étant pas constant, ce dernier montre des dynamiques temporelles plus complexes. Pour une station avec un coefficient positif sur la deuxième composante, il indique que celle-ci se videra entre 17h et 3h, et se remplira entre 3h et 17h. Inversement, pour une station avec un coefficient négatif, la dynamique est opposée.

3.4.4 Visualisation des Coefficients des Composantes

Nous pouvons afficher les coefficients associés à chaque station sur deux cartes, une pour chaque composante principale de l'ACP :



Concernant la première composante de l'ACP, nous observons que les stations situées dans le centre-ville ont généralement des coefficients négatifs, tandis que les stations en périphérie ont des coefficients positifs. Cette observation suggère que les stations du centre suivent une dynamique de remplissage similaire, alors que les stations en périphérie suivent une dynamique inverse. Cette tendance peut être liée à des habitudes de déplacement différentes entre les utilisateurs du centre-ville et ceux de la périphérie.

La deuxième composante de l'ACP a montré une tendance opposée en ce qui concerne les coefficients. Cela suggère que les variations des dynamiques de remplissage des stations peuvent être influencées par divers éléments, tels que la météo, les événements locaux ou les jours de la semaine. Cette remarque met en évidence l'importance de saisir ces différences afin de gérer de manière optimale le service de vélos en libre-service.

3.4.5 Reconstruction des Courbes Moyennes

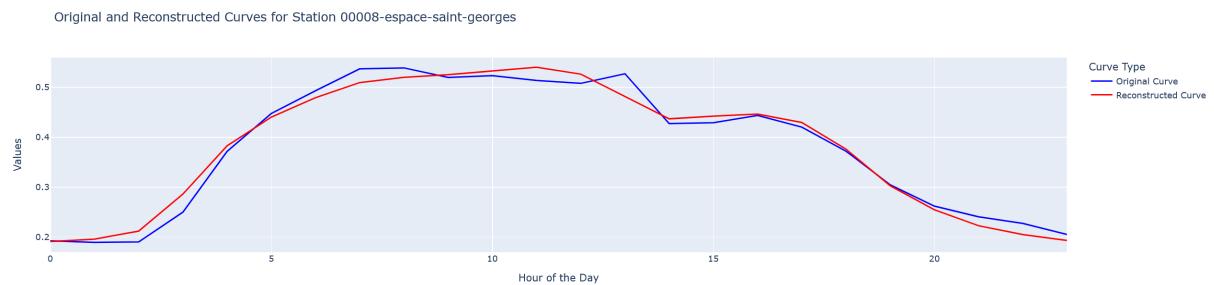
Pour approfondir notre analyse, nous avons employé l'ACP pour reconstruire les courbes de moyenne horaire pour différentes stations. L'objectif est de comprendre comment les variations journalières sont capturées par les composantes principales.

La reconstruction d'une courbe se base sur la formule suivante :

$$\text{Courbe reconstruite} = \bar{X} + \sum_{i=1}^n w_i \cdot c_i$$

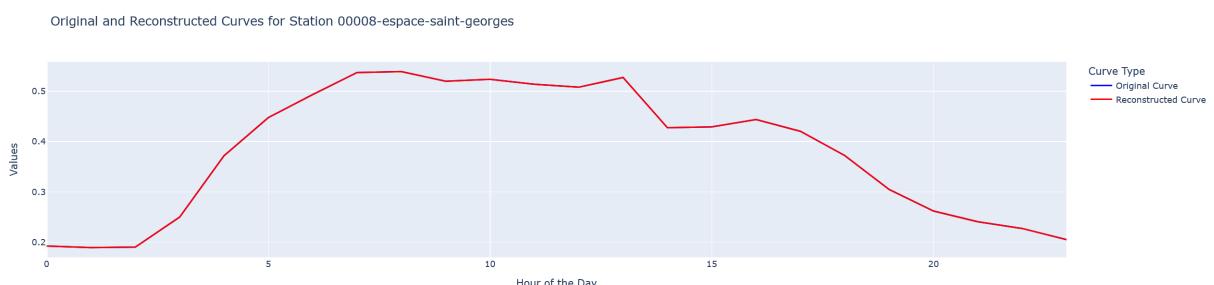
où \bar{X} représente la moyenne générale des stations, w_i sont les poids des composantes principales (vecteurs propres associés à chaque composante principale), et c_i sont les scores des composantes pour une station donnée. Cette formule permet de reconstruire la courbe originale à partir d'une combinaison linéaire des composantes principales, pondérées par leur importance respective dans la variance des données.

Dans l'onglet *ACP* de la section *Statistique Descriptive* de notre application, les utilisateurs peuvent visualiser les vecteurs propres et reconstruire les courbes horaires des stations en sélectionnant un nombre spécifique de composantes principales. Voici un exemple pour la station "00008-espace-saint-georges" :



Reconstruction de la moyenne horaire de la station 00008-espace-saint-georges avec les deux premières composantes de l'ACP.

L'utilisation des deux premières composantes de l'ACP a permis de reconstruire la courbe de manière relativement fidèle, ce qui s'explique par le fait que ces composantes captent plus de 90% de la variance des données. Si nous ajoutons davantage de composantes, la courbe reconstruite devient progressivement plus précise. Par exemple, en sélectionnant les 24 composantes principales — correspondant aux 24 heures de la journée — la courbe reconstruite est identique à la courbe originale :



Reconstruction de la moyenne horaire de la station 00008-espace-saint-georges avec les 24 composantes de l'ACP.

En utilisant toutes les composantes, nous obtenons une reconstruction parfaite de la courbe horaire. Cette fonctionnalité est directement accessible dans notre application, permettant aux utilisateurs d'expérimenter avec le nombre de composantes pour observer les effets sur la reconstruction.

3.4.6 Implications pour les Prédictions Futures

Les résultats de l'ACP mettent en évidence des dynamiques de déplacement de la population entre le centre-ville et la périphérie. Ces dynamiques sont essentielles pour optimiser la gestion et l'utilisation des vélos en libre-service à Toulouse. En utilisant les courbes reconstruites, il est possible de développer des modèles prédictifs plus robustes, tirant parti de la structure sous-jacente des données captée par l'ACP.

En conclusion, l'ACP nous fournit une compréhension approfondie des dynamiques entre les stations de vélos, facilitant ainsi des prédictions plus précises et une gestion optimisée du service Vélo Toulouse.

4 Prédiction des Activités des Stations de Vélo

L'objectif principal de cette section est d'explorer et de comparer divers modèles de prédiction pour estimer l'activité future des stations de vélo à travers la ville. Nous utiliserons une approche empirique, exploitant 70% des données disponibles pour l'entraînement des modèles, ce qui correspond à une période du 1er avril 2016 au 4 août 2016, totalisant un peu plus de 2000 enregistrements. Les 30% restants serviront à tester la performance des modèles.

4.1 Objectifs de Prédiction

Nos analyses se concentreront sur deux horizons de prédiction principaux, chacun adapté à des besoins spécifiques de gestion et d'optimisation des ressources de vélos en libre-service :

- **Prédiction à court terme:** Cet horizon vise à prédire l'activité des stations pour le jour suivant. La prédiction à court terme est particulièrement utile pour répondre aux fluctuations quotidiennes de la demande, qui peuvent être influencées par des facteurs tels que les conditions météorologiques, les événements locaux, ou les variations saisonnières. Des modèles comme les régressions linéaires ou les arbres de décision peuvent être efficaces pour ce type de prédiction en raison de leur capacité à intégrer rapidement de nouvelles données et à ajuster les prédictions en conséquence.
- **Prédiction à moyen terme:** Cette approche a pour objectif de prévoir l'activité hebdomadaire des stations. Les prédictions à moyen terme sont cruciales pour la planification stratégique, incluant la maintenance des vélos et la redistribution optimale des ressources pour satisfaire la demande attendue. Les modèles basés sur des méthodes de séries temporelles peuvent être particulièrement adaptés pour capturer et modéliser les tendances et cycles hebdomadaires.

Ces prévisions nous permettront d'approfondir notre compréhension et de prévoir les fluctuations quotidiennes et hebdomadaires de l'usage des vélos. Il est essentiel d'avoir cette compréhension afin de gérer efficacement les ressources, ce qui permet aux opérateurs de systèmes de vélos en libre-service de maximiser l'utilisation tout en réduisant les coûts opérationnels. Un modèle efficace à court terme peut aider les opérateurs à réagir de manière efficace aux fluctuations soudaines de la demande, tandis qu'une capacité de prédiction à moyen terme élevée peut faciliter la planification des besoins en maintenance et en réapprovisionnement des stations.

En résumé, la sélection du modèle ou de la combinaison de modèles sera fortement influencée par la nature particulière des données disponibles et les objectifs précis de l'opérateur. Il est possible que certains modèles soient plus efficaces pour les prédictions à court terme en raison de leur rapidité et de leur souplesse, tandis que d'autres, peut-être plus complexes et intégrant des analyses de données plus approfondies, seront plus adaptés pour les prédictions à moyen terme.

4.2 Méthode d'entraînement des modèles

Comme nous l'avons mentionné précédemment, pour l'entraînement, nous avons utilisé 70% des données, ce qui correspond à une période du 01-04-2016 au 04-08-2016.

Pour l'entraînement de nos modèles, nous avons utilisé une méthode d'extraction de caractéristiques temporelles afin de créer de nouvelles variables. À partir de la date, nous avons créé les variables suivantes : *Heure (hour)*, *Jour de la semaine (day_of_week)*, *Jour du mois (day_of_month)*, *Est-ce le week-end ? (is_weekend)* et *Est-ce un dimanche ? (is_sunday)*.

4.2.1 Extraction de Caractéristiques Temporelles

L'extraction de caractéristiques temporelles est une méthode qui consiste à transformer des informations de date et d'heure en variables additionnelles que les modèles de machine learning peuvent utiliser pour mieux comprendre les données temporelles. Cette approche est particulièrement utile pour les séries temporelles, car elle permet de capturer des motifs et des tendances récurrents.

4.2.2 Objectifs de la Méthode

L'objectif principal de l'extraction de caractéristiques temporelles est de fournir au modèle des informations contextuelles supplémentaires qui peuvent améliorer la précision des prédictions. En utilisant des informations telles que l'heure de la journée ou le jour de la semaine, le modèle peut mieux comprendre les variations et les tendances saisonnières ou quotidiennes présentes dans les données.

4.2.3 Avantages de la Méthode

Capture des Motifs Saisonniers Cette méthode permet de capturer les motifs saisonniers et les cycles récurrents. Par exemple, les taux d'occupation des vélos peuvent varier significativement entre les jours ouvrables et les week-ends, ou entre les différentes heures de la journée. En incluant ces caractéristiques, le modèle peut mieux prédire ces variations.

Amélioration de la Précision Les informations temporelles supplémentaires enrichissent le jeu de données en fournissant des indices précieux sur les dynamiques temporelles. Cela peut conduire à une amélioration de la précision des prédictions, car le modèle dispose de plus de contextes pour comprendre les variations des données.

Facilité de Mise en Œuvre L'extraction de caractéristiques temporelles est relativement simple à mettre en œuvre. Elle ne nécessite pas de transformations complexes des données, ce qui la rend accessible et rapide à intégrer dans le processus d'entraînement des modèles.

4.2.4 Inconvénients de la Méthode

Complexité Accrue L'ajout de nombreuses caractéristiques temporelles peut augmenter la complexité du modèle, rendant l'entraînement plus long et potentiellement plus sujet au surapprentissage (overfitting), surtout si le nombre de données d'entraînement est limité.

Dépendance aux Données Historiques Les modèles entraînés avec des caractéristiques temporelles peuvent être fortement dépendants des patterns historiques. Si les conditions changent de manière significative (par exemple, un changement dans le comportement des utilisateurs ou des événements imprévus), le modèle peut avoir du mal à s'adapter.

4.2.5 Exemple Pratique

Voici un exemple pratique de la manière dont nous avons implémenté l'extraction de caractéristiques temporelles pour nos modèles :

```
@staticmethod
def create_features_from_date(date_serie: pd.Series) -> pd.DataFrame:
    df_X = pd.DataFrame()
    df_X['hour'] = date_serie.dt.hour.astype('uint8')
    df_X['day_of_week'] = date_serie.dt.dayofweek.astype('uint8')
    df_X['day_of_month'] = date_serie.dt.day.astype('uint8')
    df_X['is_weekend'] = (date_serie.dt.dayofweek >= 5).astype('uint8')
    df_X['is_sunday'] = (date_serie.dt.dayofweek == 6).astype('uint8')
    return df_X
```

Grâce à cette méthode, nous avons pu transformer les informations de date en variables que les modèles peuvent utiliser pour mieux capturer les tendances et les variations des données temporelles.

En résumé, l'extraction de caractéristiques temporelles est une technique puissante et pratique pour améliorer les modèles de prédition de séries temporelles. Elle offre de nombreux avantages en termes de capture des motifs saisonniers et d'amélioration de la précision des prédictions, bien qu'elle puisse également introduire une complexité accrue et une dépendance aux données historiques.

4.3 Métriques d'Évaluation des Modèles

Pour assurer une comparaison rigoureuse des modèles de prédiction, nous avons choisi d'utiliser l'Erreur Moyenne Absolue (Mean Absolute Error : MAE) et l'Erreur Quadratique Moyenne (Between Mean Squared Error : MSE). Ces métriques sont particulièrement adaptées à notre contexte, où les prédictions de pourcentage de vélos présents dans une station sont des valeurs continues comprises entre 0 et 1.

4.3.1 Erreur Moyenne Absolue (MAE)

L'Erreur Moyenne Absolue est définie par la formule :

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

où y_i représente les valeurs réelles et \hat{y}_i les valeurs prédites, pour n observations.

La MAE mesure la différence moyenne absolue entre les valeurs prédites et les valeurs réelles, offrant ainsi une mesure directe de l'erreur moyenne sans direction (positive ou négative). Dans le contexte de notre étude, où nous prévoyons des proportions (donc, des valeurs comprises strictement entre 0 et 1), la MAE nous donne une interprétation claire de l'erreur moyenne en termes de proportion. Par exemple, une MAE de 0.05 indique que, en moyenne, l'erreur de nos prédictions s'écarte de 5% de la valeur réelle, ce qui est très utile pour évaluer la précision de prédictions dans des contextes où les erreurs doivent être minimisées, telles que les systèmes de gestion de ressources.

4.3.2 Erreur Quadratique Moyenne (MSE)

L'Erreur Quadratique Moyenne est exprimée par la formule :

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

où, comme pour la MAE, y_i sont les valeurs réelles et \hat{y}_i les valeurs prédites.

La MSE élève les erreurs au carré avant de les moyenniser, ce qui a pour effet de pénaliser plus sévèrement les grandes erreurs. Ce comportement rend la MSE particulièrement pertinente dans notre étude où une sous-estimation ou une surestimation significative de la disponibilité des vélos pourrait mener à des décisions inefficaces ou insatisfaisantes en termes de gestion des stations de vélo. Une MSE faible indique que le modèle prédit les proportions avec une grande précision et avec peu d'erreurs importantes, ce qui est essentiel pour maintenir une bonne expérience utilisateur dans le contexte du partage de vélos.

En résumé, la MAE et la MSE sont des outils précieux pour évaluer la précision de nos modèles de prédiction dans un contexte où la précision des proportions prédites est cruciale pour la planification et la gestion opérationnelle des stations de vélo.

Voici un tableau résumant les avantages et les inconvénients de ces deux métriques dans notre cas d'application:

Métrique	Avantages	Inconvénients
MAE	<ul style="list-style-type: none"> Fournit une mesure directe et facile à comprendre des erreurs. Moins sensible aux valeurs aberrantes, ce qui est utile lorsque des erreurs extrêmes sont attendues mais non critiques. Bonne indication de l'erreur moyenne dans les prédictions de proportions. 	<ul style="list-style-type: none"> Ne pénalise pas autant les grandes erreurs, ce qui pourrait être un problème si ces erreurs ont des conséquences significatives. Peut ne pas refléter la gravité des erreurs dans certains contextes où les erreurs plus grandes sont plus problématiques.
MSE	<ul style="list-style-type: none"> Pénalise plus sévèrement les erreurs importantes, ce qui aide à identifier et à corriger les prédictions très inexactes. Très utile lorsque les grandes erreurs sont inacceptables, comme dans la gestion stratégique des stations de vélo. Favorise la précision des modèles en mettant l'accent sur la réduction des grandes erreurs. 	<ul style="list-style-type: none"> Peut être excessivement influencée par des valeurs aberrantes ou des erreurs extrêmes. Peut mener à un modèle qui est trop concentré sur les rares grandes erreurs au détriment de nombreuses petites erreurs.

Table 1: Avantages et inconvénients des métriques MAE et MSE pour la prédiction de la présence des vélos en station

4.4 Comparaison des Performances

4.4.1 Comparaisons entre Modèles

Pour évaluer et comparer l'efficacité des différents modèles de prédiction développés dans notre étude, nous mettrons en œuvre une approche systématique basée sur deux métriques statistiques essentielles : l'Erreur Moyenne Absolue (MAE) et l'Erreur Quadratique Moyenne (MSE). Comme vu précédemment, ces métriques nous permettront de quantifier la précision des prédictions fournies par chaque modèle.

Modèle de Base Comme point de référence, nous avons établi un modèle de base qui prédit l'activité des stations en se basant sur la moyenne des données par jour et par heure. Ce modèle simple servira de benchmark pour évaluer la performance des modèles plus sophistiqués. L'objectif principal sera donc de développer des modèles qui surpassent ce modèle de référence en termes de précision de prédiction.

Méthodologie de Comparaison Pour chaque modèle testé, nous calculerons les valeurs de MAE et MSE pour chaque station, ainsi que des valeurs moyennes de ces métriques sur l'ensemble des stations. Cette double approche permettra une analyse détaillée à deux niveaux :

- Analyse Locale** : En effectuant le calcul du MAE et du MSE pour chaque station de manière individuelle, nous pouvons déterminer comment chaque modèle se comporte dans des situations particulières. Cela s'applique notamment aux stations avec des dynamiques d'utilisation différentes, où certains modèles peuvent être plus adaptés que d'autres.
- Analyse Globale** : La moyenne des MAE et MSE sur toutes les stations permet d'obtenir une mesure de performance globale qui résume l'efficacité du modèle sur l'ensemble du réseau. Il est essentiel de prendre en compte cette mesure globale afin d'évaluer l'efficacité du modèle pour prendre des décisions opérationnelles et stratégiques au niveau de la ville.

Périodes de Prédiction Deux périodes de prédiction différentes seront utilisées pour calculer les métriques : quotidienne et hebdomadaire. Grâce à cette segmentation, nous pouvons évaluer la capacité des modèles à gérer des prévisions à court terme par rapport à des prévisions plus long terme. Certains modèles peuvent être performants dans les prévisions à court terme en raison de leur réactivité aux changements récents des données, tandis que d'autres pourraient être plus adaptés à des prévisions à long terme en raison de leur capacité à intégrer et à analyser des tendances sur des périodes plus longues.

Dans ce rapport, nous avons sélectionné des dates précises pour les périodes de prédiction dans le but de comparer de manière cohérente et structurée les performances de nos modèles. Nos tests de prédiction ont été réalisés sur deux périodes différentes : une période à court terme et une période à moyen terme.

Nous procéderons à nos tests sur les données du vendredi 5 août 2016 pour les prédictions à court terme. Nous avons choisi cette journée pour évaluer les performances de nos modèles sur une seule journée, afin d'évaluer leur aptitude à anticiper rapidement et immédiatement les fluctuations de disponibilité des vélos.

Pour les prédictions à moyen terme, nous analyserons une période plus étendue, allant du vendredi 5 août 2016 jusqu'au jeudi 11 août 2016 inclus. En choisissant cette période d'une semaine, nous visons à tester la robustesse et la précision de nos modèles sur une période plus longue, capturant ainsi les variations journalières et les tendances hebdomadaires.

La sélection de ces périodes spécifiques permet d'établir une base de comparaison standardisée pour nos prédictions. En effectuant des tests sur des périodes définies à l'avance, nous pouvons :

- Assurer la reproductibilité de nos tests et résultats.
- Évaluer la performance de nos modèles de manière cohérente sur des horizons temporels différents.
- Identifier les points forts et les limitations de chaque modèle dans des contextes de prédiction variés.

De plus, l'analyse des prévisions à court et à moyen terme nous donne une vision globale de la capacité de nos modèles à s'ajuster à différentes échelles temporelles. Cela se révèle particulièrement bénéfique pour des applications pratiques où les besoins de prédiction peuvent varier, allant de la gestion quotidienne des stations de vélos à la planification stratégique hebdomadaire.

Classement des Modèles Grâce à ces analyses, nous serons en mesure de classer les modèles selon leur performance pour chaque période de prédiction. Ce classement contribuera non seulement à identifier les modèles les plus performants mais aussi à explorer les raisons de leur efficacité, en fournissant des insights précieux sur les caractéristiques des données et des dynamiques urbaines qui influencent le plus les résultats.

En somme, cette méthode de comparaison rigoureuse et multi-niveaux nous aidera à déterminer les modèles les plus prometteurs pour une gestion optimale des stations de vélos en libre-service, tout en mettant en lumière les défis et opportunités associés à la prédiction dans des environnements urbains complexes.

4.4.2 Comparaison Géographique des Performances

En plus de l'analyse numérique utilisant les métriques MAE et MSE, nous inclurons une dimension spatiale dans notre analyse en effectuant une étude géographique des erreurs de prédiction. Grâce à cette méthode, il sera possible de représenter visuellement les performances des divers modèles sur une carte de la ville, ce qui offre une vision visuelle et intuitive des données.

Visualisation des Erreurs sur Carte Pour chaque modèle, nous calculerons le MSE de chaque station et représenterons ces valeurs sur une carte de la ville. Cette cartographie des erreurs de prédiction a plusieurs objectifs:

- **Identification des Zones de Performance** : La visualisation mettra en évidence les stations qui présentent des erreurs de prédiction faibles ou élevées, facilitant ainsi l'identification des zones où les modèles réussissent bien ou échouent.
- **Analyse de la Variabilité Géographique** : En analysant les erreurs à travers différentes zones de la ville, nous pourrons détecter des patterns spécifiques, tels que des prédictions plus précises dans des zones de haute fréquentation comme le centre-ville, où la régularité du flux de vélos pourrait simplifier la modélisation.

Impact sur la Gestion des Stations Il est essentiel d'effectuer cette analyse géographique non seulement afin d'évaluer la qualité des modèles de prédition, mais aussi pour assister les responsables des stations de vélos dans l'optimisation de leurs ressources. En saisissant les endroits et les raisons pour lesquelles certaines prédictions sont moins précises, il est envisageable d'adapter les stratégies de redistribution des vélos ou de mettre en place des campagnes de maintenance spécifiques.

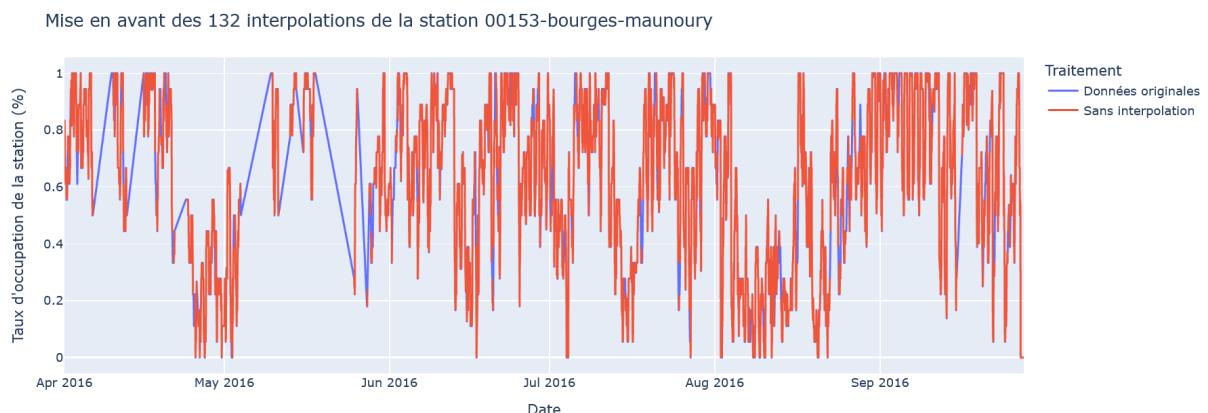
Amélioration de l'Expérience Utilisateur Cette analyse a un impact qui va au-delà de la gestion opérationnelle ; elle joue également un rôle crucial dans l'amélioration de l'expérience des utilisateurs. En repérant les stations avec des prévisions fiables, nous pouvons fournir une meilleure information aux utilisateurs sur les disponibilités prévues, diminuant ainsi le risque de dénicher une station vide ou pleine. En outre, en modifiant les modèles afin d'améliorer les prédictions là où elles sont indispensables, nous accroissons la satisfaction des utilisateurs, qui peuvent avoir plus confiance en l'application pour planifier leurs déplacements.

Perspectives d'Optimisation À long terme, cette analyse des résultats prédictifs nous offre des éléments essentiels pour poursuivre l'amélioration des algorithmes en prenant en considération les spécificités géographiques et comportementales des utilisateurs. Prenons l'exemple des zones à dynamiques très variables, telles que les quartiers d'affaires ou les zones touristiques, où les besoins en vélos peuvent fluctuer considérablement d'un jour à l'autre. D'autres modèles pourraient être élaborés pour les zones plus calmes où les tendances sont plus stables. saisonnières sont moins mises en avant.

4.5 Gestion des interpolations

Nos données, observées sur une période de six mois, sont sujettes à des valeurs manquantes qui varient en fonction des stations. Pour combler ces lacunes, les données nous ont été fournies avec des interpolations linéaires afin de compléter ces valeurs manquantes. Cependant, ces interpolations sont plus ou moins présentes selon les stations et, dans certains cas, les interpolations linéaires couvrent des périodes assez longues, pouvant aller jusqu'à plusieurs jours.

Nous avons donc programmé un algorithme permettant de détecter ces interpolations dans nos données. Dans la section *Prédiction*, vous pouvez cliquer sur l'icône "Dataset" avec un petit "i" entouré pour "info", situé en haut à droite de la page, afin d'observer les interpolations pour nos différentes stations.



Mise en avant des interpolations présentes dans la station 00153-bourges-maunoury

Sur ce graphique, vous pouvez voir les 132 interpolations détectées par notre algorithme. La courbe rouge représente les données sans les interpolations, permettant ainsi de visualiser en bleu les différentes interpolations appliquées.

Les interpolations dans nos données présentent plusieurs inconvénients :

1. Biais dans les Modèles de Prédition Les interpolations linéaires peuvent introduire des biais dans les modèles de prédition. Si les périodes interpolées sont longues, les modèles risquent d'apprendre des tendances artificielles qui ne reflètent pas la réalité.

2. Perturbation des Données Temporelles Les interpolations sur de longues périodes peuvent dissimuler les réelles dynamiques temporelles, ce qui amène à une moins grande fiabilité des prédictions. Par exemple, les interpolations peuvent réduire ou exagérer les variations saisonnières ou les tendances à court terme.

3. Impact sur l'Évaluation des Modèles L'obtention de données précises est essentielle pour évaluer les modèles. Les mesures de performance comme l'Erreur Moyenne Absolue (MAE) et l'Erreur Quadratique Moyenne (MSE) peuvent être altérées par les interpolations. Les interpolations dans les données de test peuvent entraîner une sous-estimation ou une surestimation des performances réelles des modèles.

Pour l'entraînement, nous avons décidé de ne pas conserver les interpolations dans nos données, car la plupart des modèles de prédiction supportent mal les interpolations et les performances des modèles sont impactées. En ce qui concerne les mesures pour le test, nous allons effectuer des mesures de MSE et de MAE à la fois avec et sans interpolations. Cela nous permettra d'obtenir des mesures plus précises pour la comparaison de nos modèles.

L'objectif de notre méthode est de réduire les conséquences néfastes des interpolations linéaires en faisant une distinction claire entre les performances des modèles sur des données interpolées et non interpolées.

4.6 Description des Modèles de Prédiction

Chaque modèle de prédiction sera détaillé dans une sous-section dédiée, où nous discuterons de son fonctionnement, de ses spécificités, et de son application aux données des stations de vélo. Ces descriptions seront accompagnées d'illustrations explicatives pour une meilleure compréhension des mécanismes sous-jacents à chaque méthode de prédiction.

- Moyenne par jours et par heure
- Prédictions via reconstruction de courbe avec l'ACP
- Régression linéaire multiple
- Forêts Aléatoires
- XGBoost
- XGBoost avec ACP

Cette approche structurée nous permettra non seulement de sélectionner le modèle le plus adapté pour nos besoins mais aussi d'approfondir notre compréhension des dynamiques temporelles et spatiales qui influencent l'activité des stations de vélo à travers la ville.

4.7 Modèle sur la moyenne par jours et par heures

Notre premier modèle consiste à calculer la moyenne des disponibilités de vélos par jour et par heure pour chaque station sur les données d'entraînement. Cette moyenne sera ensuite utilisée pour réaliser des prédictions. Comme mentionné précédemment, ce modèle servira de référence, et l'objectif des prochains modèles sera de surpasser ses performances.

4.7.1 Description du modèle

Le modèle de moyenne est basé sur une approche simple mais efficace : il agrège les données historiques pour calculer la disponibilité moyenne des vélos pour chaque combinaison de jour et d'heure. Par exemple, pour une station donnée, nous calculons la moyenne des disponibilités de vélos à 9 heures du matin le lundi, à 10 heures du matin le mardi, et ainsi de suite, pour chaque jour de la semaine et chaque heure de la journée.

4.7.2 Avantages du modèle

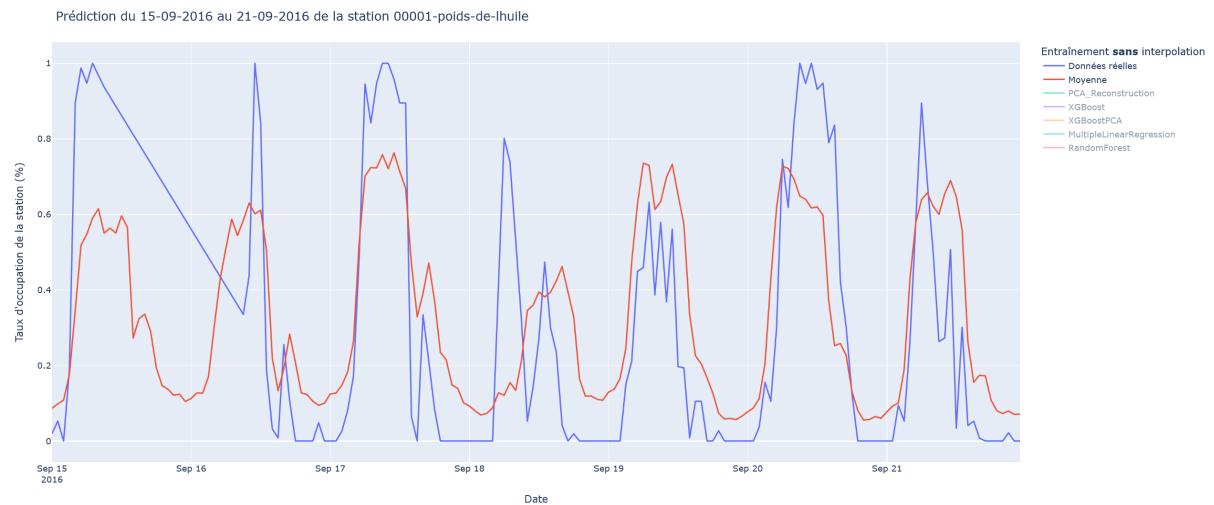
- Facilité d'entraînement :** Ce modèle est très facile à entraîner car il ne nécessite que le calcul de la moyenne des données d'entraînement.
- Robustesse aux données manquantes :** En calculant des moyennes sur des périodes plus longues, ce modèle peut atténuer l'impact des données manquantes ou bruitées.
- Interprétabilité :** Les prédictions générées par ce modèle sont facilement compréhensibles car elles se basent directement sur les moyennes historiques des disponibilités de vélos.

4.7.3 Inconvénients du modèle

- Difficulté à prédire les valeurs extrêmes :** En raison de sa nature, ce modèle a tendance à lisser les valeurs et à prédire des disponibilités proches de la moyenne historique. Cela peut poser problème pour prévoir des situations extrêmes, telles que des stations complètement vides (0) ou pleines (1).
- Limitation des prédictions :** Les prédictions ont tendance à rester dans un intervalle restreint, typiquement entre 0.05 et 0.8, ce qui peut ne pas refléter fidèlement les variations réelles de la disponibilité des vélos, particulièrement lors des périodes de forte demande ou de faible disponibilité.

4.7.4 Exemple de prédition

Pour illustrer le fonctionnement de ce modèle, nous présentons ci-dessous un graphique comparant les prédictions du modèle de moyenne aux valeurs réelles de disponibilité des vélos sur une semaine, pour la station *00001-poids-de-l'huile*.



Prédiction sur une semaine avec le modèle de moyenne, pour la station 00001-poids-de-l'huile

Comme le montre le graphique, les prédictions suivent les tendances générales des données réelles, mais manquent parfois de précision pour capturer les fluctuations soudaines et les valeurs extrêmes. Cela est dû à la nature même du modèle de moyenne, qui lisse les variations pour fournir une estimation stable mais moins réactive aux changements brusques de la demande.

4.7.5 Intérêt et Utilisation

Malgré ses limitations, le modèle de moyenne par jours et par heures présente un intérêt certain pour des applications où la simplicité et la rapidité d'exécution sont cruciales. Par exemple, il peut servir de base pour des systèmes de prévision en temps réel nécessitant une latence minimale. De plus, en tant que modèle de référence, il permet de quantifier l'amélioration apportée par des modèles plus sophistiqués.

En conclusion, le modèle de moyenne par jours et par heures offre une approche simple et intuitive pour prédire la disponibilité des vélos. Bien qu'il présente des limitations, il constitue un point de départ solide pour le développement et la comparaison de modèles prédictifs plus avancés.

4.8 Modèle via reconstruction de courbe avec l'ACP

Notre second modèle utilise l'Analyse en Composantes Principales (ACP) pour reconstruire les courbes de disponibilité des vélos et effectuer des prédictions. En se basant sur les cinq premières composantes principales de l'ACP, ce modèle vise à capturer les variations principales des données tout en réduisant leur dimensionnalité. Ce modèle est une extension de notre analyse ACP précédente, appliquée cette fois-ci aux moyennes par jours et par heures.

4.8.1 Description du modèle

Le modèle de reconstruction de courbe via l'ACP suit une approche structurée pour extraire les composantes principales des données historiques et les utiliser pour reconstruire les courbes de disponibilité des vélos. Voici les étapes clés du modèle :

1. **Calcul des moyennes horaires** : Les données historiques de disponibilité des vélos sont agrégées pour calculer la moyenne des disponibilités par jour et par heure pour chaque station.
2. **Application de l'ACP** : Une ACP est effectuée sur les données moyennes pour extraire les composantes principales, en utilisant les cinq premières composantes principales pour capturer la majeure partie de la variance des données.
3. **Reconstruction des courbes** : Les courbes de disponibilité des vélos sont reconstruites en utilisant une combinaison linéaire des cinq premières composantes principales.
4. **Prédiction** : Les composantes principales obtenues sont utilisées pour effectuer des prédictions sur les disponibilités futures des vélos pour chaque station.

4.8.2 Avantages du modèle

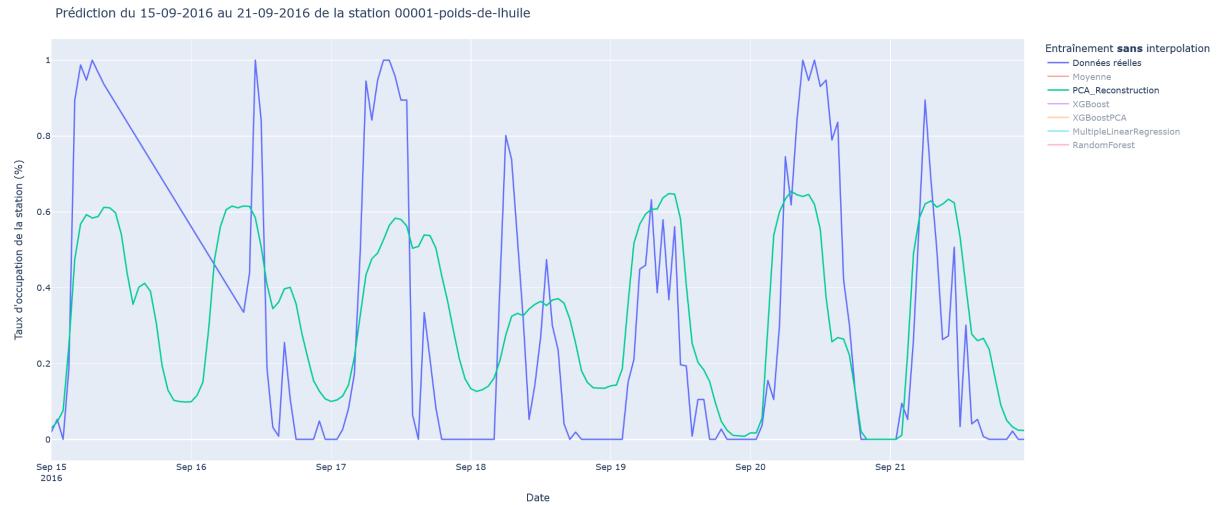
- **Réduction de la dimensionnalité** : En utilisant l'ACP, ce modèle réduit la dimensionnalité des données, capturant ainsi les principales variations avec un nombre limité de composantes. Cela simplifie le modèle et peut améliorer la performance de prédiction.
- **Capturer les tendances principales** : Les composantes principales permettent de capturer les tendances majeures et les structures sous-jacentes des données, ce qui peut rendre les prédictions plus robustes et précises par rapport aux simples moyennes.
- **Adaptation aux variations temporelles** : En se basant sur les composantes principales, le modèle peut mieux s'adapter aux variations temporelles complexes et aux dynamiques de la disponibilité des vélos.

4.8.3 Inconvénients du modèle

- **Complexité accrue** : Par rapport au modèle de moyenne simple, ce modèle nécessite un calcul supplémentaire pour effectuer l'ACP et reconstruire les courbes, ce qui peut augmenter la complexité computationnelle.
- **Sensibilité aux données d'entraînement** : La performance du modèle dépend fortement de la qualité et de la représentativité des données d'entraînement utilisées pour calculer les composantes principales.

4.8.4 Exemple de prédiction

Pour illustrer le fonctionnement de ce modèle, nous présentons ci-dessous un graphique comparant les prédictions du modèle de reconstruction de courbe via l'ACP aux valeurs réelles de disponibilité des vélos sur une semaine, pour la station *00001-poids-de-l'huile*.



Prédiction sur une semaine avec le modèle de reconstruction de courbe via l'ACP, pour la station 00001-poids-de-l'huile

Comme le montre le graphique, les prédictions générées par le modèle ACP sont similaires aux prédictions faites avec le modèle de moyenne. Cependant, les estimations obtenues avec l'ACP sont plus lisses et plus générales, ce qui peut être pertinent dans une optique de prévision à long terme. Malgré cela, ce modèle présente encore la même limitation que le modèle de moyenne : il a du mal à capturer les valeurs extrêmes et les pics de disponibilité des vélos, ce qui peut réduire sa précision lors des périodes de forte demande ou de disponibilité très faible.

4.8.5 Intérêt et Utilisation

Le modèle de reconstruction de courbe via l'ACP présente un intérêt significatif pour les applications nécessitant des prédictions plus lisses et une meilleure capture des tendances générales. En utilisant les composantes principales, ce modèle permet de mieux comprendre les dynamiques sous-jacentes des disponibilités de vélos.

Cependant, il est important de noter que, tout comme le modèle de moyenne, le modèle ACP a des difficultés à prédire les valeurs extrêmes. Cela signifie qu'il peut ne pas être aussi efficace pour prévoir les périodes de forte affluence ou de pénurie, où les disponibilités des vélos peuvent atteindre des niveaux très bas ou très élevés.

En conclusion, bien que le modèle de reconstruction de courbe via l'ACP offre une approche avancée pour prédire la disponibilité des vélos, il partage certaines limitations avec le modèle de moyenne. Néanmoins, en tant que méthode de prévision à long terme, il reste un outil précieux pour analyser et anticiper les tendances générales de la disponibilité des vélos en libre-service.

4.9 Modèle via Régression Linéaire Multiple

Notre troisième modèle utilise la régression linéaire multiple pour prédire la disponibilité des vélos en fonction de plusieurs variables explicatives. Ce modèle tente de capturer les relations linéaires entre les variables d'entrée (comme l'heure de la journée, le jour de la semaine, ...) et la disponibilité des vélos.

4.9.1 Description du modèle

Le modèle de régression linéaire multiple suit une approche structurée pour modéliser la relation entre les variables explicatives et la disponibilité des vélos. Voici les étapes clés du modèle :

- Sélection des variables explicatives** : Identification des variables pertinentes qui influencent la disponibilité des vélos, telles que l'heure de la journée, le jour de la semaine, la température, les précipitations, etc.
- Préparation des données** : Agrégation et transformation des données historiques pour inclure les variables explicatives sélectionnées. Les données sont normalisées si nécessaire.

3. **Ajustement du modèle** : Utilisation d'une régression linéaire multiple pour ajuster un modèle aux données, en estimant les coefficients de régression pour chaque variable explicative.
4. **Prédiction** : Application du modèle ajusté pour prédire la disponibilité des vélos en fonction des valeurs des variables explicatives.

4.9.2 Avantages du modèle

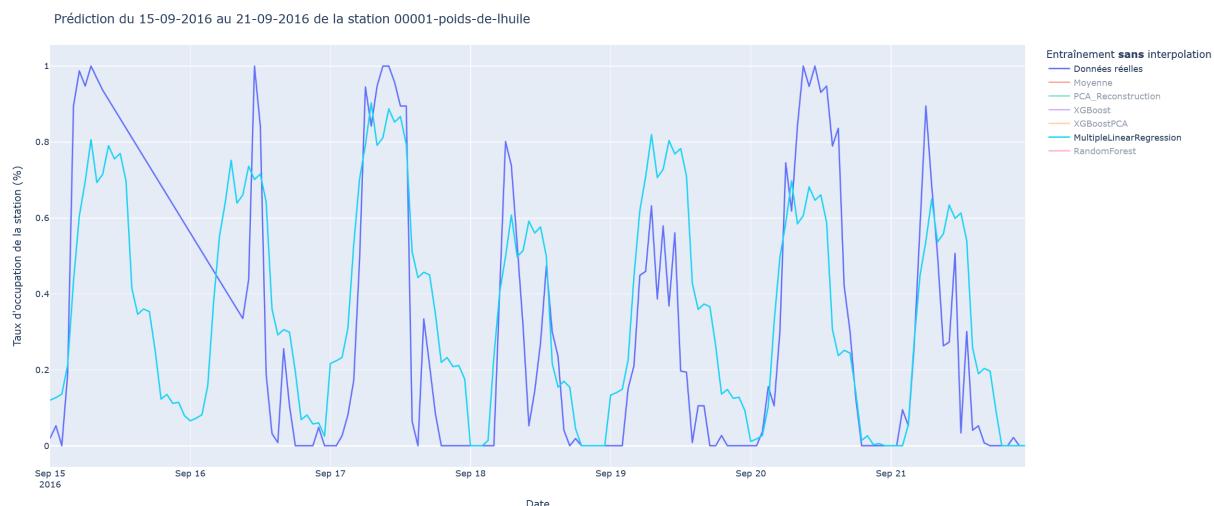
- **Simplicité et interprétabilité** : Le modèle de régression linéaire multiple est simple à mettre en œuvre et à interpréter. Chaque coefficient de régression indique l'impact de la variable correspondante sur la disponibilité des vélos.
- **Prise en compte des variables multiples** : Ce modèle peut intégrer plusieurs variables explicatives, permettant de capturer les relations linéaires complexes entre les différentes influences et la disponibilité des vélos.
- **Facilité d'implémentation** : La régression linéaire multiple est une technique bien établie et largement disponible dans de nombreux logiciels statistiques et bibliothèques de machine learning.

4.9.3 Inconvénients du modèle

- **Limitation aux relations linéaires** : La régression linéaire multiple ne peut capturer que les relations linéaires entre les variables explicatives et la disponibilité des vélos, ce qui peut limiter sa précision si les relations réelles sont non linéaires.
- **Sensibilité aux valeurs aberrantes** : Les valeurs aberrantes peuvent fortement influencer les coefficients de régression et, par conséquent, affecter la précision des prédictions.
- **Multicolinéarité** : Si les variables explicatives sont fortement corrélées entre elles, cela peut entraîner des problèmes de multicolinéarité, rendant les coefficients de régression instables et difficiles à interpréter.

4.9.4 Exemple de prédiction

Pour illustrer le fonctionnement de ce modèle, nous avons choisi de vous présenter un graphique comparant les prédictions du modèle de régression linéaire multiple aux valeurs réelles de disponibilité des vélos sur une semaine, pour la station *00001-poids-de-l'huile*.



Prédiction sur une semaine avec le modèle de régression linéaire multiple, pour la station 00001-poids-de-l'huile

Comme le montre le graphique, les prédictions générées par le modèle de régression linéaire multiple suivent les tendances générales des valeurs réelles, avec une bonne précision globale. Cependant, ce

modèle peut également avoir des difficultés à capturer les variations extrêmes de la disponibilité des vélos (en effet, on remarque que les prédictions ont du mal à dépasser un taux d'occupation de 80%), notamment lors des périodes de forte demande ou de faible disponibilité.

4.9.5 Intérêt et Utilisation

Les applications qui requièrent une interprétation claire des relations entre les variables explicatives et la disponibilité des vélos sont particulièrement intéressées par le modèle de régression linéaire multiple. Grâce à sa capacité à prendre en compte différentes variables, ce modèle permet de prendre en compte différentes influences sur la disponibilité des vélos.

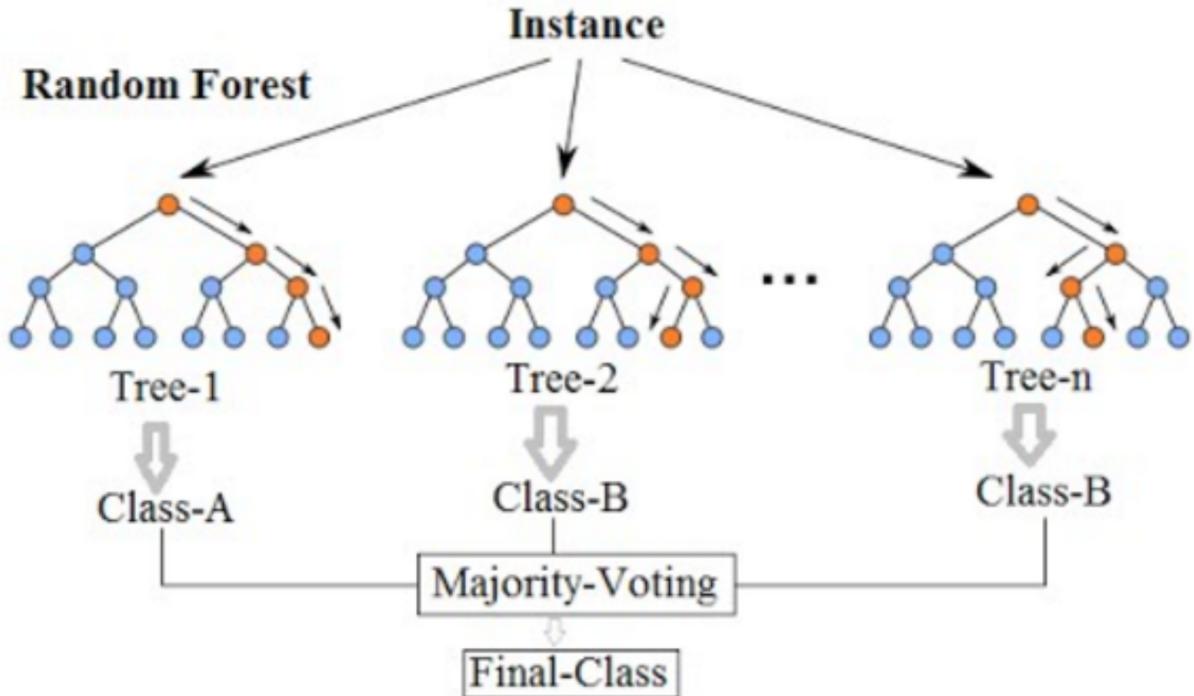
Toutefois, il convient de souligner que ce modèle peut présenter des limites en ce sens qu'il ne peut modéliser que les relations linéaires. Pour des phénomènes plus complexes, il pourrait être nécessaire d'utiliser d'autres modèles plus avancés. Cependant, en tant qu'approche fondamentale, la régression linéaire multiple demeure un outil précieux pour les analyses préliminaires et les prédictions rapides.

En conclusion, malgré sa simplicité et son interprétation facile, le modèle de régression linéaire multiple comporte certaines limitations en ce qui concerne la prise en compte des relations non linéaires et la gestion des valeurs aberrantes. Néanmoins, il reste un modèle pratique pour appréhender les éléments qui influencent la disponibilité des vélos et pour faire des prédictions rapides et facilement interprétables.

4.10 Modèle via Forêts Aléatoires

Notre quatrième modèle utilise les Forêts Aléatoires [2] (Random Forests) pour prédire la disponibilité des vélos. Les Forêts Aléatoires sont des ensembles d'arbres de décision, et elles sont bien connues pour leur capacité à gérer des données complexes et non linéaires. Ce modèle vise à surpasser les performances des modèles précédents en capturant plus efficacement les relations complexes entre les variables.

Random Forest Simplified



Exemple simplifié de l'utilisation d'une forêt aléatoire

4.10.1 Description du modèle

Le modèle de Forêts Aléatoires suit une approche structurée pour exploiter les données historiques et prédire les disponibilités de vélos. Voici les étapes clés du modèle :

1. **Prétraitement des données :** Les données historiques de disponibilité des vélos sont nettoyées et préparées. Cela inclut la gestion des valeurs manquantes (énormément présentes dans nos données), la normalisation des données, et la création de nouvelles variables (features) telles que les conditions météorologiques, les événements spéciaux, etc.
2. **Entraînement du modèle :** Un ensemble d'arbres de décision est entraîné sur les données d'entraînement. Chaque arbre est construit en utilisant un échantillon aléatoire des données et une sous-ensemble aléatoire des variables (échantillon défini précédemment).
3. **Prédiction :** Les prédictions sont faites en agrégant les résultats de tous les arbres de la forêt. Pour chaque station et chaque intervalle de temps, la disponibilité des vélos est prédite en prenant la moyenne des prédictions de tous les arbres.

4.10.2 Avantages du modèle

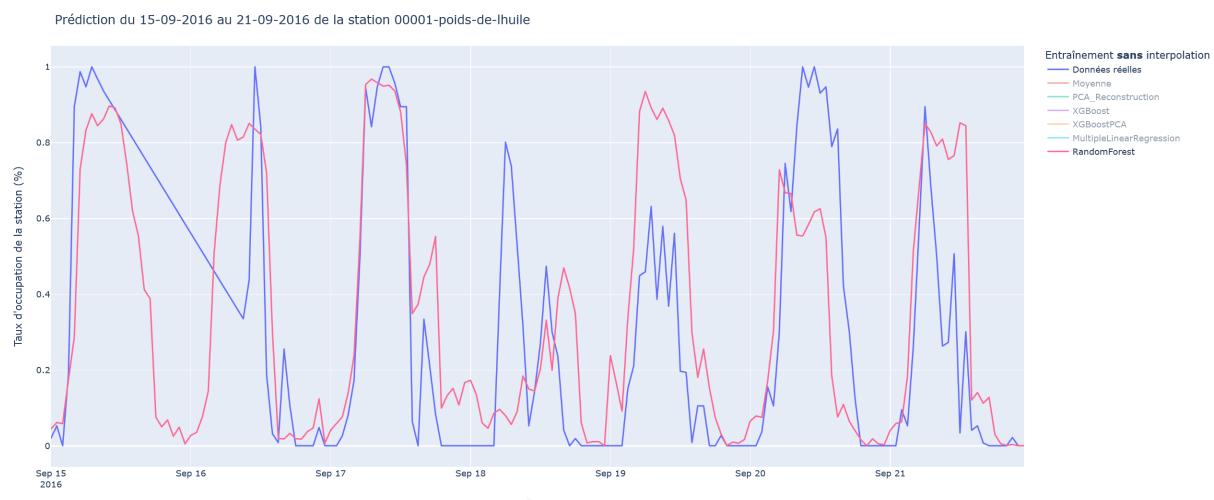
- **Capture des relations non linéaires :** Les Forêts Aléatoires peuvent capturer des relations complexes et non linéaires entre les variables, ce qui permet des prédictions plus précises.
- **Robustesse et réduction du surapprentissage :** En utilisant plusieurs arbres de décision, les Forêts Aléatoires réduisent le risque de surapprentissage (overfitting) et sont robustes aux variations des données.
- **Importance des variables :** Le modèle peut fournir des informations sur l'importance relative des différentes variables utilisées, aidant ainsi à mieux comprendre les facteurs influençant la disponibilité des vélos.

4.10.3 Inconvénients du modèle

- **Complexité computationnelle :** Les Forêts Aléatoires nécessitent des ressources computationnelles importantes pour l'entraînement et la prédiction, surtout lorsque le nombre d'arbres et de variables est élevé.
- **Interprétabilité :** Bien que le modèle puisse indiquer l'importance des variables, les prédictions individuelles peuvent être difficiles à interpréter par rapport à des modèles plus simples comme la moyenne.
- **Temps d'entraînement :** En raison de la complexité des calculs, le temps d'entraînement peut être significativement plus long comparé aux modèles plus simples.

4.10.4 Exemple de prédiction

Pour illustrer le fonctionnement de ce modèle, nous présentons ci-dessous un graphique comparant les prédictions du modèle de Forêts Aléatoires aux valeurs réelles de disponibilité des vélos sur une semaine, pour la station *00001-poids-de-l'huile*.



Prédiction sur une semaine avec le modèle de prédiction de forêt aléatoire, pour la station 00001-poids-de-l'huile

Comme le montre le graphique, les prédictions des Forêts Aléatoires suivent de manière plus précise les fluctuations réelles de la disponibilité des vélos, capturant à la fois les tendances générales et les variations plus subtiles que les modèles de moyenne et d'ACP pourraient manquer.

4.10.5 Intérêt et Utilisation

Le modèle de Forêts Aléatoires est particulièrement utile pour des applications nécessitant des prédictions précises et robustes, même dans des environnements de données complexes. Par exemple, il peut être utilisé pour des systèmes de gestion de flotte de vélos en temps réel, où des prédictions précises sont cruciales pour équilibrer la distribution des vélos à travers différentes stations.

En résumé, même si le modèle de Forêts Aléatoires est plus complexe et nécessite davantage de ressources, il offre une capacité accrue à prendre en compte les dynamiques complexes de la disponibilité des vélos. Ce modèle joue un rôle crucial dans l'amélioration de la précision des prédictions par rapport aux modèles plus basiques, tout en offrant des données précises.

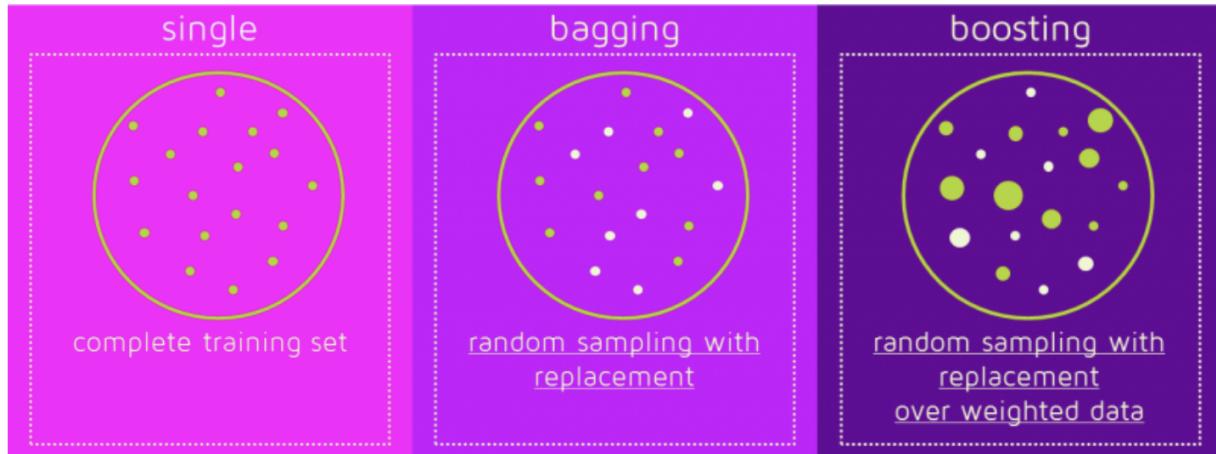
4.11 Modèle XGBoost

XGBoost, ou eXtreme Gradient Boosting [3], est un algorithme de machine learning très performant utilisé principalement pour les tâches de classification et de régression. Il se distingue par sa capacité à gérer efficacement de grandes quantités de données et à fournir des prédictions précises.

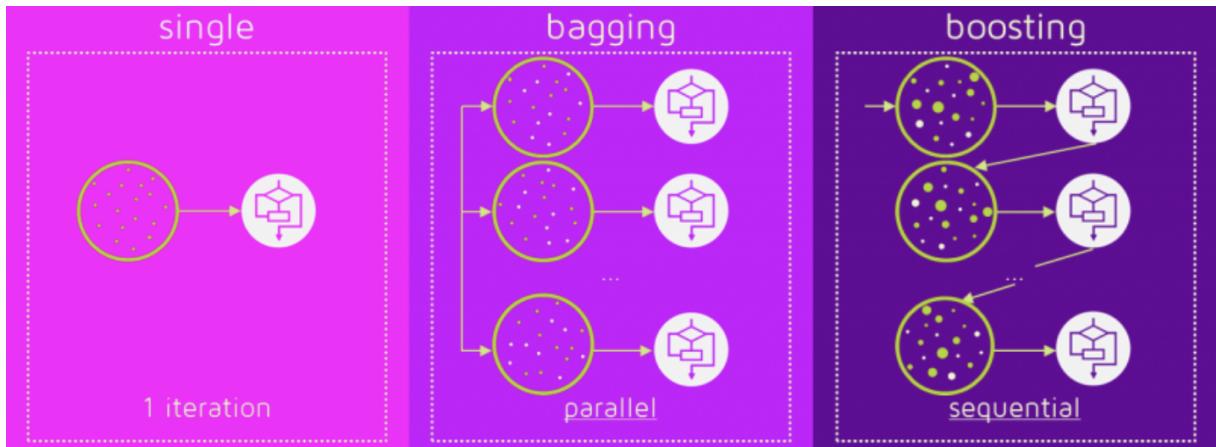
4.11.1 Description du modèle

XGBoost est basé sur le principe du boosting, une méthode d'ensemble où des modèles faibles sont combinés pour créer un modèle robuste. Les étapes clés de ce modèle sont les suivantes :

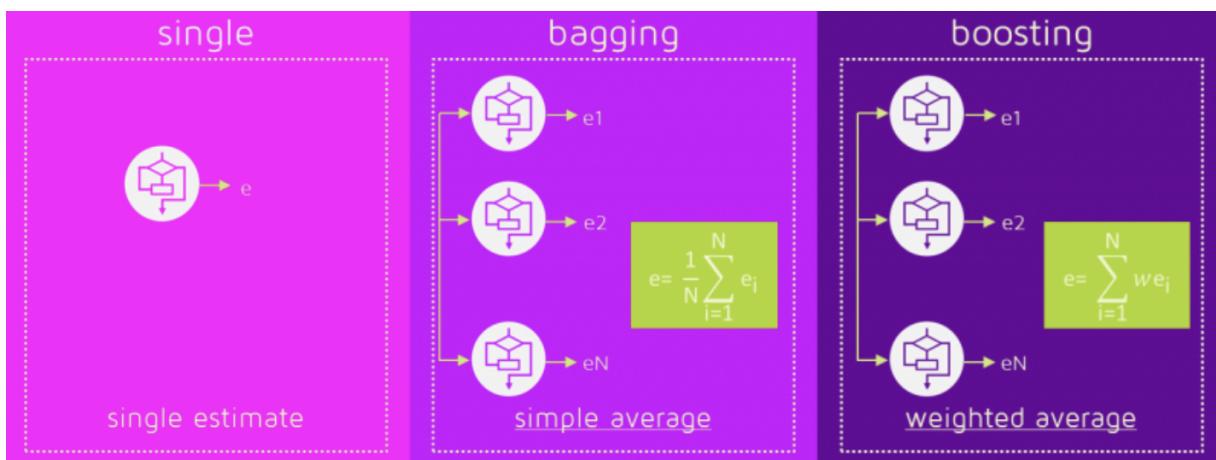
1. La première étape consiste à créer un premier modèle de base à partir d'un algorithme choisi. Il est entraîné sur les données. Au début, on attribue des poids égaux à toutes les observations. À partir des résultats obtenus de ce modèle, si une observation est mal classée, cela augmente son poids.



2. Par la suite, un deuxième modèle est élaboré afin de chercher à rectifier les erreurs du premier modèle. Il est formé en utilisant les données pondérées recueillies lors de la première étape. On poursuit cette procédure et on ajoute des modèles jusqu'à ce que l'ensemble des données de formation soit correctement prédictible ou que le nombre maximal de modèles soit ajouté.



Les prédictions du modèle ajouté récemment seront les prédictions globales pondérées fournies par les anciens modèles d'arbres.



3. **Information** : Différents modèles reposent sur le principe de boosting et utilisent diverses méthodes pour évaluer les poids (AdaBoost, LPBoost, XGBoost, GradientBoost, BrownBoost sont des exemples).

4.11.2 Avantages du modèle

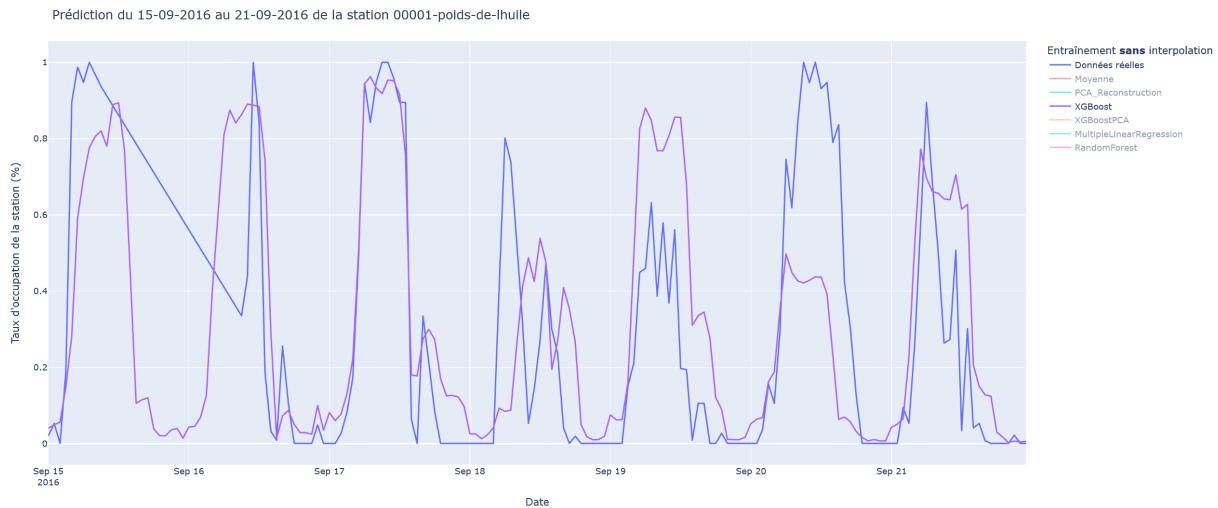
- **Performance élevée** : XGBoost est connu pour sa rapidité et son efficacité, souvent en tête des compétitions de data science.
- **Gestion des valeurs manquantes** : Il gère efficacement les valeurs manquantes (qui sont assez conséquentes dans notre jeu de données), rendant le modèle robuste face à des jeux de données imparfaits.
- **Régularisation** : Des techniques de régularisation intégrées préviennent le surapprentissage, rendant les prédictions plus généralisables.
- **Flexibilité** : XGBoost peut être utilisé pour divers types de problèmes, qu'il s'agisse de classification binaire, de classification multi-classes, ou de régression.

4.11.3 Inconvénients du modèle

- **Complexité accrue** : La mise en œuvre et l'optimisation de XGBoost peuvent être complexes et nécessiter des compétences techniques avancées.
- **Sensibilité aux hyperparamètres** : La performance du modèle dépend fortement de l'optimisation précise des hyperparamètres, ce qui peut nécessiter du temps et des ressources.

4.11.4 Exemple de prédiction

Pour illustrer le fonctionnement de ce modèle, nous présentons ci-dessous un graphique comparant les prédictions du modèle XGBoost aux valeurs réelles de disponibilité des vélos sur une semaine, pour la station *00001-poids-de-l'huile*.



Prédiction sur une semaine avec le modèle XGBoost, pour la station 00001-poids-de-l'huile

Comme le montre le graphique, les prédictions générées par le modèle XGBoost sont précises et suivent bien les tendances des valeurs réelles. Ce modèle est capable de capturer à la fois les tendances générales et les variations fines des données de disponibilité des vélos.

4.11.5 Intérêt et Utilisation

Le modèle XGBoost est particulièrement utile pour les applications nécessitant des prédictions précises et robustes, telles que la gestion de la disponibilité des vélos en libre-service. Grâce à ses capacités de gestion des valeurs manquantes et de prévention du surapprentissage, XGBoost assure des performances élevées et des prédictions fiables.

En conclusion, XGBoost propose une solution puissante et performante pour prédire la disponibilité des vélos, mettant en avant la précision et la solidité tout en gérant les défis liés aux données complexes et larges.

4.12 Modèle XGBoost avec ACP

Le modèle XGBoost est un algorithme d'apprentissage automatique qui utilise l'approche du boosting, réputé pour ses performances élevées et sa rentabilité. En associant XGBoost à l'Analyse en Composantes Principales (ACP), il est possible d'améliorer davantage la précision des prédictions en diminuant la dimensionnalité des données et en prenant en compte les principales variations présentes dans les données.

4.12.1 Description du modèle

L'intégration de XGBoost avec l'ACP suit une approche structurée pour traiter les données et effectuer des prédictions précises :

- 1. Prétraitement des données :** Les données brutes sont prétraitées pour corriger les valeurs manquantes et normaliser les valeurs numériques.
- 2. Application de l'ACP :** L'ACP est appliquée pour extraire les composantes principales des données. Cela permet de réduire la dimensionnalité tout en conservant la majorité de la variance des données (comme on l'a vu précédemment, avec 2 composantes principales, nous avons pu obtenir 90% de la variance des données).

3. **Construction du modèle XGBoost** : Un modèle XGBoost est construit en utilisant les composantes principales comme variables d'entrée. Le modèle est formé pour minimiser les erreurs de prédiction.
4. **Optimisation des hyperparamètres** : Les hyperparamètres du modèle XGBoost sont optimisés pour améliorer les performances. Cela inclut le réglage du taux d'apprentissage, de la profondeur des arbres, et du nombre d'arbres.
5. **Prédiction** : Les composantes principales obtenues sont utilisées pour effectuer des prédictions sur les disponibilités futures des vélos pour chaque station.

4.12.2 Avantages du modèle

- **Réduction de la dimensionnalité** : L'utilisation de l'ACP permet de simplifier les données et de réduire la charge computationnelle sans perdre des informations essentielles.
- **Performance et précision** : XGBoost est réputé pour ses performances et sa capacité à gérer les données complexes, ce qui se traduit par des prédictions précises.
- **Gestion des valeurs manquantes** : XGBoost gère efficacement les valeurs manquantes, ce qui en fait un modèle robuste face à des jeux de données imparfaits.
- **Prévention du surapprentissage** : Les techniques de régularisation intégrées à XGBoost aident à prévenir le surapprentissage, rendant les prédictions plus généralisables.

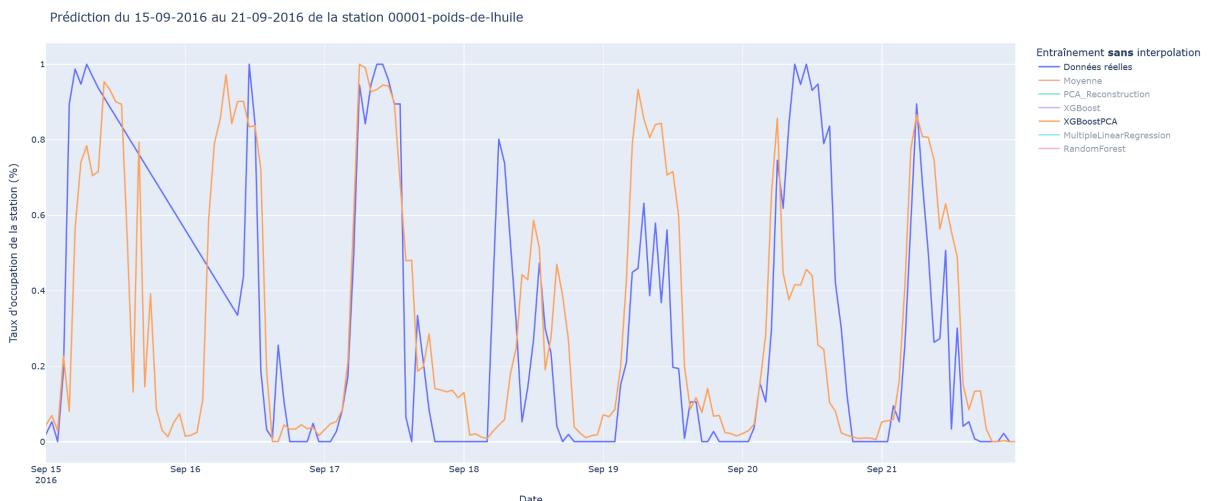
Finalement, ce modèle possède toutes les qualités de l'utilisation de l'ACP et du modèle de XGBoost réunis. Mais il possède aussi les inconvénients de ces derniers comme nous allons le voir tout de suite.

4.12.3 Inconvénients du modèle

- **Complexité accrue** : La combinaison de l'ACP et de XGBoost augmente la complexité du modèle, nécessitant plus de temps de calcul et de ressources.
- **Sensibilité aux hyperparamètres** : La performance du modèle dépend fortement de l'optimisation des hyperparamètres, ce qui peut nécessiter des compétences et du temps.

4.12.4 Exemple de prédiction

Pour illustrer le fonctionnement de ce modèle, nous présentons ci-dessous un graphique comparant les prédictions du modèle XGBoost avec ACP aux valeurs réelles de disponibilité des vélos sur une semaine, pour la station *00001-poids-de-l'huile*.



Prédiction sur une semaine avec le modèle de prédiction de XGBoost avec ACP, pour la station 00001-poids-de-l'huile

Comme le montre le graphique, les prédictions générées par le modèle XGBoost avec ACP sont très précises et suivent bien les tendances des valeurs réelles. Ce modèle est capable de capturer à la fois les tendances générales et les variations fines des données de disponibilité des vélos.

4.12.5 Intérêt et Utilisation

L'utilisation du modèle XGBoost avec ACP est particulièrement bénéfique pour les applications qui requièrent des prédictions précises et solides, comme la gestion de la disponibilité des vélos en libre service. L'ACP permet au modèle de traiter de manière efficace des jeux de données de grande taille en diminuant leur complexité. XGBoost offre des performances élevées et des prédictions fiables grâce à ses compétences en gestion des valeurs manquantes et en prévention du surapprentissage.

En résumé, l'association de XGBoost avec l'ACP permet d'obtenir une solution puissante et performante pour prédire la disponibilité des vélos, en combinant précision et solidité tout en gérant les défis liés aux données complexes et larges. Nous pouvons dès à présent supposer que ce modèle de prédiction est le plus performant. Nous allons vérifier cela dans la partie suivante.

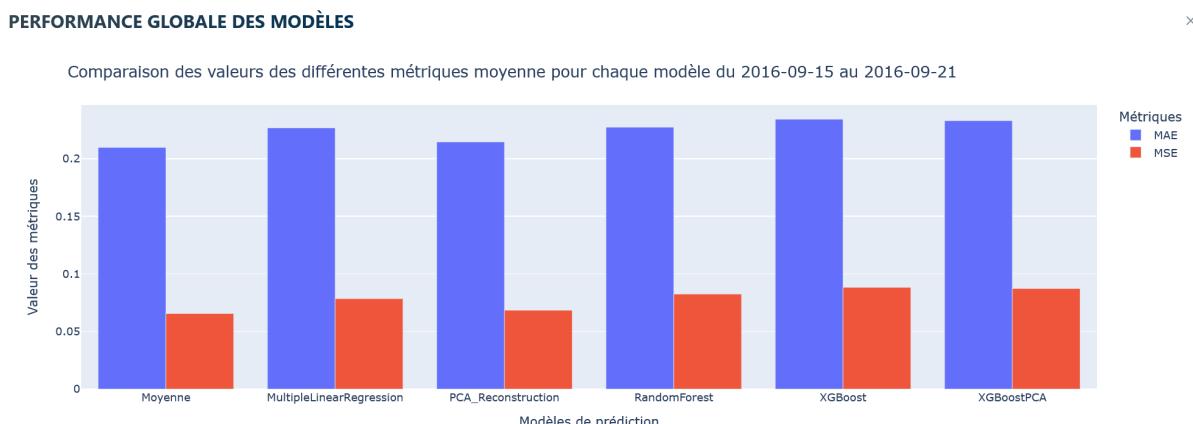
5 Comparaison des modèles

Dans cette étude, nous nous plongeons dans l'analyse comparative des performances des modèles de prédiction. Notre objectif est d'évaluer et de comparer les différents modèles en examinant leurs performances moyennes à l'aide de métriques clés telles que la Mean Absolute Error (MAE) et la Mean Squared Error (MSE).

Nous cherchons à déterminer quel modèle se distingue en termes de précision et de stabilité dans ses prédictions. Cette exploration nous permettra d'identifier les modèles les plus adaptés à notre ensemble de données et de prendre des décisions quant à leur utilisation dans des scénarios pratiques.

5.1 Performances globales des modèles

Nous allons ici commencer par analyser les performances globales des modèles. Nous avons réalisé sur notre site un graphique à barres dans lequel nous comparons la MSE et MAE de chaque modèle pour une prédiction d'une semaine allant du 15 septembre 2016 au 21 septembre 2016 :



Performances globales des modèles

5.1.1 Analyse détaillée des modèles

1. Moyenne

- **MAE:** ~0.20
- **MSE:** ~0.06
- **Interprétation:** Utiliser la moyenne des valeurs pour prédire est un modèle de base. La MAE élevée indique que ce modèle est moins précis, et le MSE montre une dispersion des erreurs relativement modérée.

2. Multiple Linear Regression

- **MAE:** ~0.22
- **MSE:** ~0.07
- **Interprétation:** Ce modèle linéaire n'améliore pas la précision par rapport à la simple moyenne. Les erreurs absolues et quadratiques sont légèrement plus élevées, indiquant une moins bonne performance globale.

3. PCA Reconstruction

- **MAE:** ~0.20
- **MSE:** ~0.05
- **Interprétation:** La reconstruction par PCA (Analyse en Composantes Principales) offre des performances similaires à la moyenne, suggérant qu'elle capte les mêmes niveaux de variance dans les données.

4. Random Forest

- **MAE:** ~0.21
- **MSE:** ~0.07
- **Interprétation:** Ce modèle n'améliore pas la précision par rapport à la simple moyenne. Les erreurs absolues et quadratiques sont légèrement plus élevées, indiquant une moins bonne performance globale.

5. XGBoost

- **MAE:** ~0.22
- **MSE:** ~0.08
- **Interprétation:** Comme la forêt aléatoire, XGBoost présente une MAE relativement élevée mais une MSE élevée.

6. XGBoost PCA

- **MAE:** ~0.22
- **MSE:** ~0.08
- **Interprétation:** Ce modèle combine XGBoost et PCA, les métriques observées n'offrent pas de meilleurs résultats.

5.1.2 Interprétation des résultats

En examinant les mesures globales pour la période choisie, il apparaît que le modèle de reconstruction par l'Analyse en Composantes Principales (ACP) se distingue comme le plus performant, suivi de près par le modèle utilisant la moyenne. Les autres modèles, bien qu'ayant des performances légèrement inférieures, ne sont pas loin derrière. Ces résultats peuvent sembler surprenants, surtout lorsqu'on s'attendrait à ce que des modèles avancés comme XGBoost ou la Forêt Aléatoire surpassent les modèles de base.

Cette observation s'explique en partie par la nature de notre jeu de données. Certaines stations de vélos présentent des tendances temporelles avec une saisonnalité très régulière, où des modèles comme XGBoost et la Forêt Aléatoire excellent en capturant ces cycles. Cependant, d'autres stations montrent des comportements plus erratiques et moins prévisibles. Dans ces cas, XGBoost et Random Forest peinent à saisir ces irrégularités, ce qui réduit leur performance globale.

De plus, notre jeu de données contient un nombre important de ces stations aux comportements irréguliers, ce qui impacte la moyenne des métriques sur l'ensemble des stations. Il est aussi crucial de noter que nous disposons uniquement de six mois de données, dont environ quatre mois pour l'entraînement. Cette période peut être insuffisante pour capturer les variations saisonnières complètes, qui varient d'une année à l'autre. Avoir une année complète de données d'entraînement pourrait améliorer considérablement les performances des modèles en leur fournissant des informations plus exhaustives sur les cycles annuels.

Enfin, les performances des modèles varient considérablement d'une station à l'autre. C'est pourquoi nous allons également procéder à des analyses locales. Cela nous permettra de déterminer, pour chaque station et en fonction de son comportement spécifique, quel modèle est le mieux adapté.

5.1.3 Analyses locales

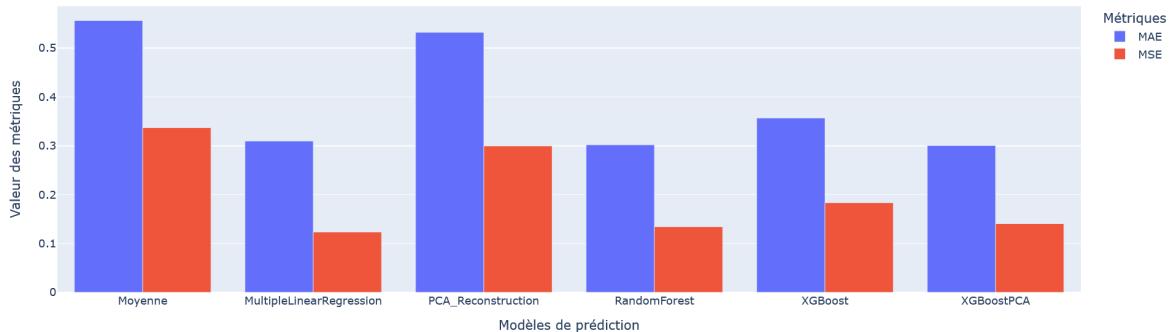
Pour notre analyse locale, nous allons étudier deux stations :

- **Quai de Tounis**, qui présente des tendances régulières,
- **Vauquelin**, qui affiche des tendances irrégulières.

Voici les métriques des modèles associés à ces deux stations :

PERFORMANCE DES MODÈLES SUR LA STATION

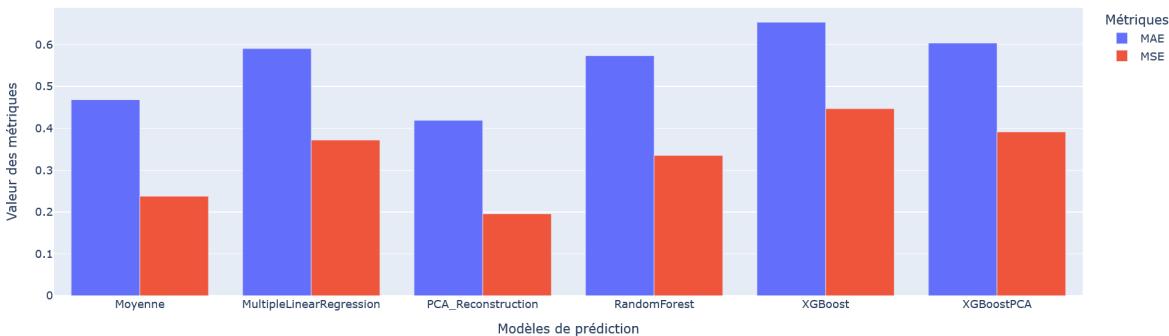
Comparaison des valeurs des différentes métriques pour chaque modèle sur la station : 00027-quai-de-tounis du 2016-09-15 au 2016-09-15



Performances des modèles sur la station Quai de Tounis

PERFORMANCE DES MODÈLES SUR LA STATION

Comparaison des valeurs des différentes métriques pour chaque modèle sur la station : 00221-vauquelin du 2016-09-15 au 2016-09-15

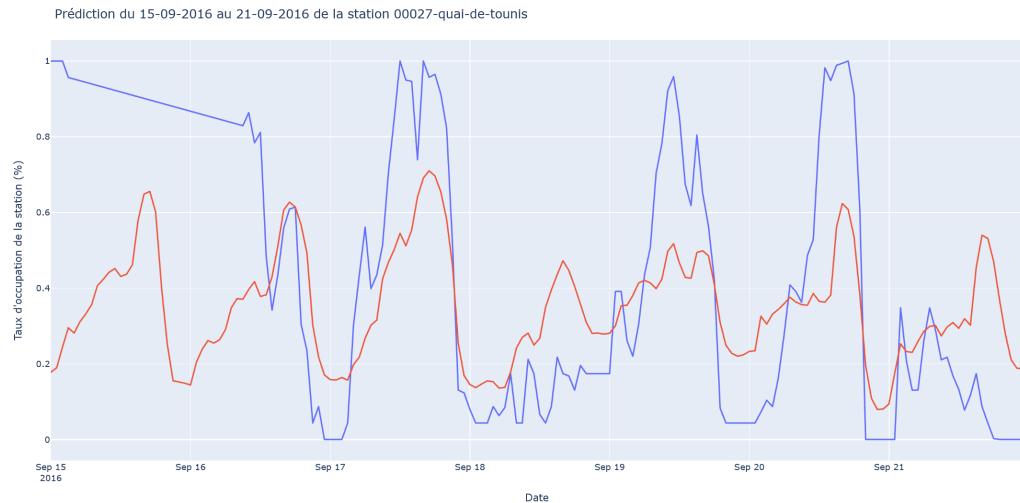


Performances des modèles sur la station Vauquelin

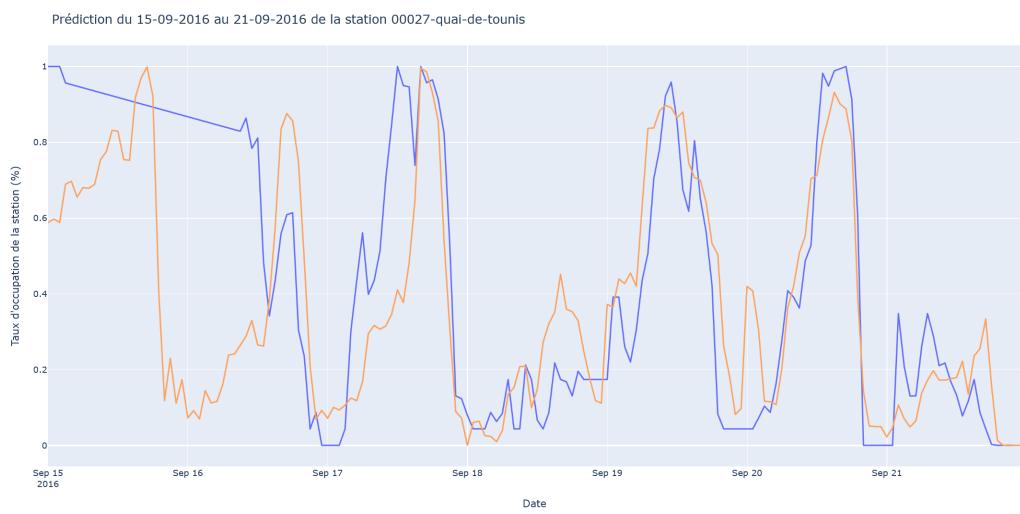
Sur la station Quai de Tounis, les performances des modèles XGBoost, RandomForest et Régression linéaire multiple sont nettement meilleures que celles des modèles de Moyenne et de Reconstruction par l'ACP. Cependant, sur la station Vauquelin, c'est l'inverse qui se produit. Les modèles de Moyenne et de Reconstruction par l'ACP surpassent les autres modèles, bien que leurs valeurs de MSE et de MAE ne soient pas faibles.

Pour comprendre pourquoi cela se produit, observons les courbes de prédiction de deux modèles par station (nous prendrons les modèles XGBoostPCA et Moyenne).

Commençons par analyser la première station :



Prédiction sur la station Quai de Tounis avec le modèle de Moyenne



Prédiction sur la station Quai de Tounis avec le modèle XGBoostPCA

En observant ces graphiques de prédiction pour nos deux modèles, le premier graphique montre la prédiction réalisée avec le modèle de Moyenne. On remarque que ce modèle a du mal à prédire les valeurs extrêmes. En revanche, sur le deuxième graphique, le modèle XGBoostPCA parvient très bien à anticiper le comportement de la station. La courbe des valeurs réelles de la station (hormis l'interpolation) montre un comportement avec des tendances saisonnières assez régulières. C'est sur ce type de station que les modèles comme XGBoost ou RandomForest sont particulièrement performants.

Passons maintenant aux prédictions de nos deux modèles sur la station Vauquelin à la même date :



Prédiction sur la station Vauquelin avec le modèle de Moyenne



Prédiction sur la station Vauquelin avec le modèle XGBoostPCA

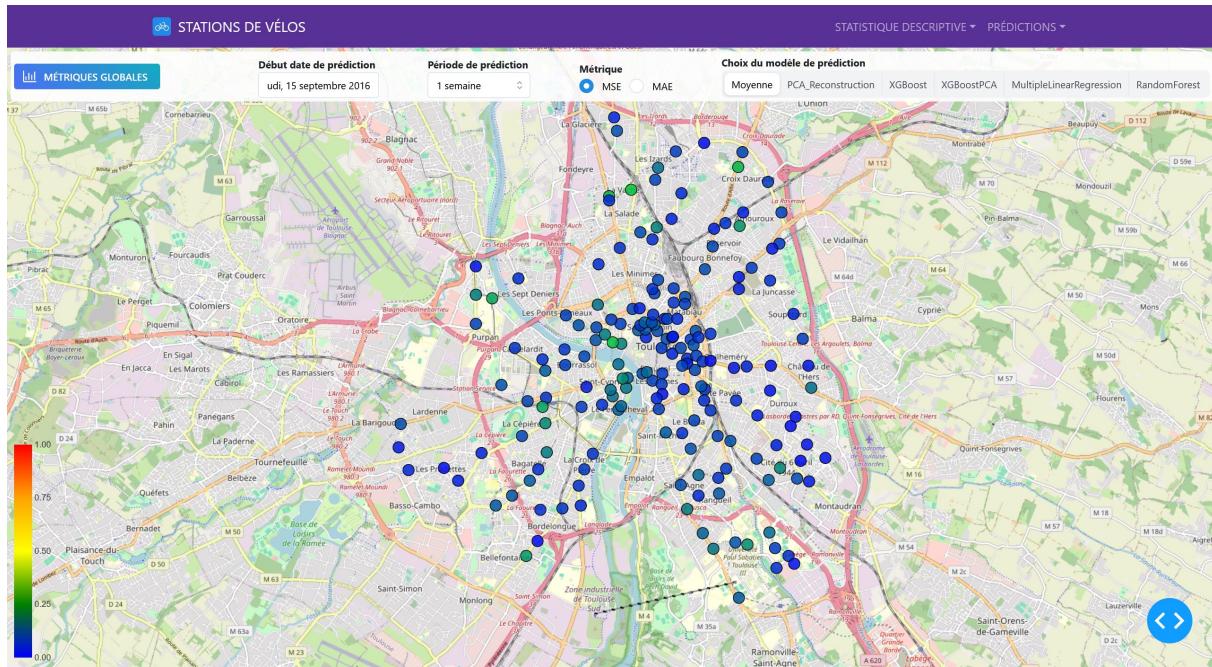
Nous constatons que les deux modèles ont du mal à prédire le comportement de la station, en raison des tendances très irrégulières de cette dernière. Toutefois, le modèle de Moyenne se rapproche légèrement plus de la réalité, car il utilise la moyenne pour prédire. Autrement dit, l'intervalle des valeurs prédictes par ce modèle se situe entre 0.2 et 0.45 environ, ce qui signifie qu'il ne prend pas de risques.

Ces observations sur ces deux stations aux comportements différents montrent que les modèles XGBoost et RandomForest sont plus performants sur des stations avec un comportement régulier.

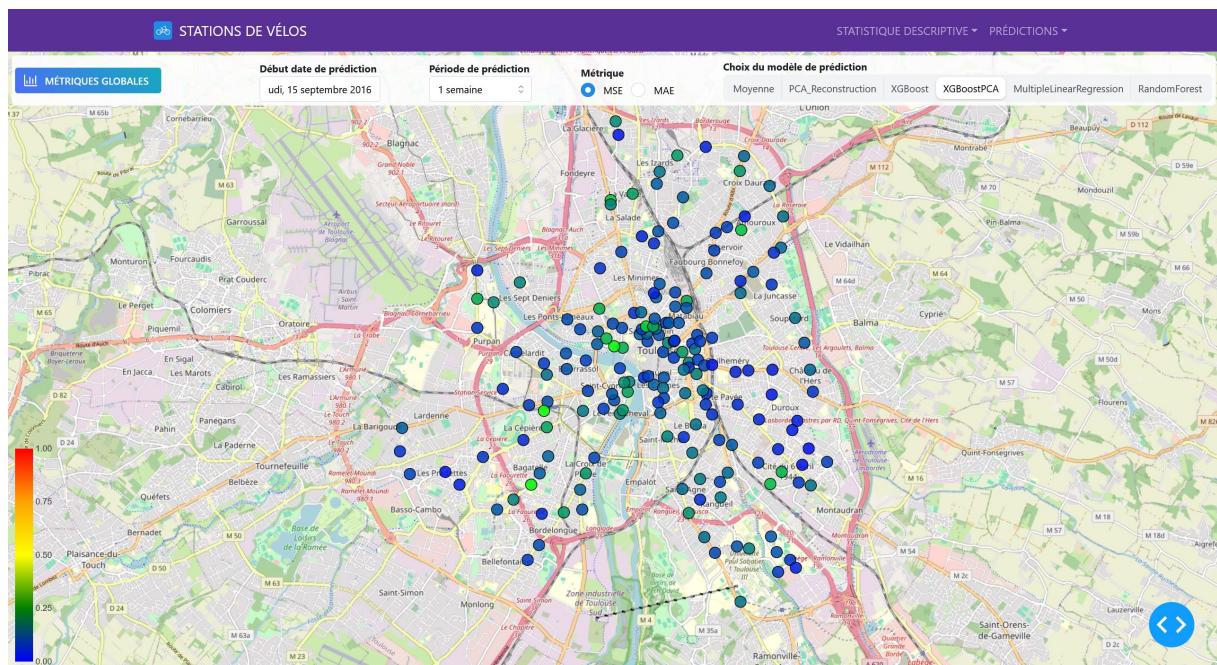
Grâce à ces analyses locales, nous pouvons mieux comprendre les forces et les faiblesses de chaque modèle en fonction du type de station. Cette compréhension est essentielle pour optimiser nos prédictions et améliorer les performances globales de notre système.

Essayons maintenant d'effectuer une analyse Géographique des métriques afin de déterminer si il y a des zones où les modèles arrivent mieux à performer.

5.2 Analyse Géographique des métriques



Carte de la métrique MSE pour le modèle de la moyenne



Carte de la métrique MSE pour le modèle XGBoostPCA

Ces deux graphiques représentent les mesures de MSE pour chaque stations sur une carte. La première carte correspond au modèle de moyenne, tandis que la deuxième au modèle XGBoostPCA.

On peut voir que les mesures de MSE pour le modèle XGBoost varient plus que celles du modèle de moyenne. Les stations où les mesures sont les plus élevées sont les stations avec des tendances irrégulières. Cela montre bien que les performances de XGBoost sont très variables en fonction du comportement des stations.

6 Conclusion du projet et observations

Dans ce projet, nous nous sommes penchés sur l'apprentissage statistique appliqué à un réseau de capteurs, avec une attention particulière à la reconstruction de la dynamique temporelle des stations de vélos en libre-service. Le projet s'est déroulé en deux grandes étapes, chacune apportant ses propres insights et avancées.

6.1 Première partie : Analyse statistique des données

La première partie de notre projet était consacrée à l'analyse statistique des données provenant des stations de vélos. Nous avons minutieusement exploré les données pour en extraire des tendances et des corrélations significatives. Grâce à l'Analyse en Composantes Principales (ACP) et à l'analyse des corrélations, nous avons pu identifier les tendances suivies par différentes stations.

Cette analyse nous a permis de diviser les stations en deux groupes distincts :

- Les stations situées au centre-ville
- Les stations en périphérie

Ces deux groupes présentent des comportements généralement inversés, ce qui est un point crucial pour nos prédictions.

6.2 Deuxième partie : Prédictions et comparaison des modèles

Dans la deuxième partie, nous avons entrepris de réaliser des prédictions en utilisant divers modèles. Parmi les modèles testés, on compte :

- Le modèle basé sur la moyenne par jour de la semaine et par heure
- La régression linéaire multiple
- La Forêt Aléatoire
- XGBoost

Nous avons comparé les performances de ces modèles et observé que la prédiction était globalement plus difficile pour les stations aux tendances irrégulières. En revanche, les modèles comme XGBoost et la Forêt Aléatoire se sont avérés très performants pour les stations présentant des tendances régulières.

6.3 Réflexions et perspectives d'amélioration

Les observations de ce projet nous conduisent à plusieurs réflexions. D'une part, nous avons pu constater l'importance de la régularité des tendances dans la performance des modèles prédictifs. D'autre part, il est clair que des stations présentant des comportements irréguliers posent un défi plus complexe.

Pour améliorer les performances des modèles de prévision de séries temporelles, nous pourrions envisager plusieurs approches :

- Augmenter la durée de la période d'entraînement pour inclure une année complète de données, afin de capturer les variations saisonnières plus précisément.
- Utiliser des modèles plus sophistiqués de séries temporelles, comme les modèles ARIMA ou LSTM (Long Short-Term Memory), qui sont spécifiquement conçus pour gérer des données temporelles.
- Incorporer des variables exogènes dans nos modèles, telles que les conditions météorologiques ou les événements locaux, qui peuvent influencer l'utilisation des vélos.

En conclusion, ce projet nous a permis de mieux comprendre les dynamiques complexes des stations de vélos en libre-service et de tester l'efficacité de différents modèles de prédiction. Nous avons identifié des pistes claires pour améliorer la précision des prédictions à l'avenir, ouvrant ainsi la voie à des applications pratiques et efficaces pour la gestion de ces services.

7 Méthodes envisagées pour aller plus loin

Dans cette section, nous présentons deux méthodes supplémentaires que nous aurions aimé explorer plus en profondeur si nous avions disposé de plus de temps. Ces techniques, bien que non utilisées dans notre étude actuelle, offrent un potentiel considérable pour améliorer la prédiction des séries temporelles.

7.1 Méthode d'entraînement des modèles

7.1.1 La Méthode Rolling Window

La méthode de la fenêtre glissante (rolling window) consiste à utiliser une fenêtre temporelle fixe pour entraîner le modèle, puis à la déplacer progressivement dans le temps pour effectuer des prédictions successives. Concrètement, nous aurions pu utiliser une fenêtre d'un mois de données pour entraîner notre modèle et effectuer une prédiction pour le jour suivant. Ensuite, la fenêtre serait décalée d'un jour, et le processus répété pour chaque nouvelle prédiction.

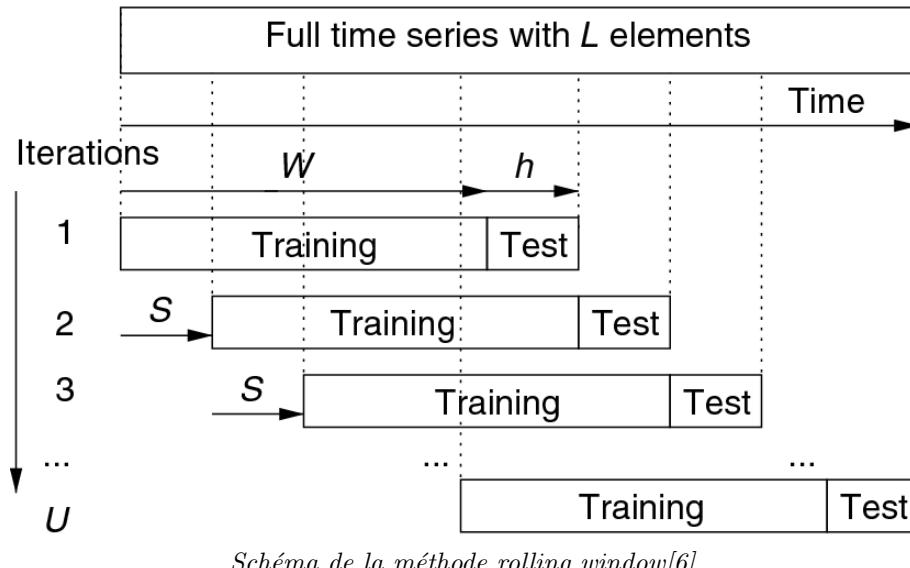


Schéma de la méthode rolling window[6]

Les avantages de cette méthode incluent :

- **Capture des changements temporels** : Permet de mieux comprendre les dynamiques et les tendances changeantes dans les données.
- **Réduction du risque de surapprentissage** : Utilise des sous-ensembles de données, aidant le modèle à mieux généraliser.
- **Évaluation réaliste** : Simule un scénario de production en prédisant un jour à la fois avec des données non vues.
- **Adaptation aux tendances récentes** : Entraînement sur des données récentes, améliorant la précision des prédictions.
- **Gestion des données en flux** : Adaptée aux systèmes en temps réel avec des données continues.
- **Robustesse aux anomalies** : Limite l'impact des anomalies grâce à une fenêtre d'entraînement plus petite et mobile.

7.1.2 La Méthode de Lagging

La méthode de lagging utilise les valeurs passées de la série temporelle comme caractéristiques pour entraîner le modèle. Par exemple, pour prédire la disponibilité des vélos dans les 24 heures à venir, on pourrait utiliser les informations des 24 heures précédentes.

Les étapes incluraient :

- **Calcul des caractéristiques temporelles** : Extraction des informations horaires, journalières et mensuelles.
- **Génération des caractéristiques de lagging** : Création de décalages (lags) pour les périodes précédentes.
- **Entraînement du modèle** : Utilisation des données disponibles avec les caractéristiques temporelles et les lags pour l'entraînement.

Les avantages de cette méthode incluent :

- **Capture des dépendances temporelles** : Saisit les liens temporels et les tendances dans les données.
- **Réduction du risque de surapprentissage** : Évite que le modèle ne devienne trop spécifique aux données d'entraînement.
- **Évaluation réaliste** : Prédiction basées sur des données non analysées par le modèle.
- **Adaptation aux tendances récentes** : Entraînement constant sur les données les plus récentes.

7.2 Conclusion

Ces deux méthodes, la fenêtre glissante et le lagging, représentent des approches robustes pour améliorer la prédiction des séries temporelles. Bien que nous n'ayons pas eu l'opportunité de les explorer pleinement dans cette étude, leur potentiel pour capturer les dynamiques temporelles, réduire le surapprentissage, et fournir des évaluations réalistes et adaptatives est indéniable. Avec plus de temps et de ressources, l'implémentation de ces méthodes pourrait significativement enrichir nos modèles de prédiction et la gestion dynamique des vélos en libre-service.

References

- [1] Mairie de Toulouse. Vélôtoulouse, 2024. Accessed: 2024-05-20, <https://abo-toulouse.cyclocity.fr>.
- [2] Niklas Donges. Random forest: A complete guide for machine learning, 2024. Accessed: 2024-05-20, <https://builtin.com/data-science/random-forest-algorithm>.
- [3] Equipe Blent. Xgboost : Tout savoir sur le boosting, 2022. Accessed: 2024-05-20, <https://blent.ai/blog/a/xgboost-tout-comprendre>.
- [4] Penn State Eberly College of Science. 18.1 - pearson correlation coefficient, 2024. Accessed: 2024-05-20, <https://online.stat.psu.edu/stat509/lesson/18/18.1>.
- [5] Plotly. Dash, 2024. Accessed: 2024-05-20, <https://dash.plotly.com/>.
- [6] ResearchGate. Schematic of the rolling window procedure, 2024. Accessed: 2024-05-20, https://www.researchgate.net/figure/Schematic-of-the-rolling-window-procedure_fig6_334900650.
- [7] Shaun Turney. Pearson correlation coefficient (r) — guide & examples, 2024. Accessed: 2024-05-20, <https://www.scribbr.com/statistics/pearson-correlation-coefficient/>.
- [8] Wikipedia. Pearson correlation coefficient, 2024. Accessed: 2024-05-20, https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.
- [9] Wikipedia. VélôToulouse, 2024. Accessed: 2024-05-20, <https://fr.wikipedia.org/wiki/VlToulouse>.