

Utilisation innovante des données opérationnelles par utilisation de l'intelligence artificielle dans un poste de contrôle ferroviaire visant l'amélioration de l'efficacité opérationnelle

Sondoss Chtioui

▶ **To cite this version:**

Sondoss Chtioui. Utilisation innovante des données opérationnelles par utilisation de l'intelligence artificielle dans un poste de contrôle ferroviaire visant l'amélioration de l'efficacité opérationnelle. Intelligence artificielle [cs.AI]. Université Paris-Saclay, 2025. Français. ⟨NNT : 2025UPAST028⟩. ⟨tel-05351859⟩

HAL Id: tel-05351859

https://theses.hal.science/tel-05351859v1

Submitted on 6 Nov 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

université
PARIS-SACLAY

Utilisation innovante des données opérationnelles par l'intelligence artificielle dans un poste de contrôle ferroviaire visant l'amélioration de l'efficacité opérationnelle

Innovative use of operational data through artificial intelligence in a rail control center to improve operational efficiency

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 575, Electrical, optical, bio : physics and engineering (EOBE).
Spécialité de doctorat : Sciences de l'information et de la communication.
Graduate School : Sciences de l'ingénierie et des systèmes.
Référent : Faculté des sciences d'Orsay.

Thèse préparée dans l'unité de recherche ESTACA'Lab, sous la direction de Toufik AZIB, Professeur, le co-encadrement de Sébastien SAUDRAIS, enseignant Chercheur, le co-encadrement de Sebti MOUELHI, enseignant Chercheur, le co-encadrement de Marc ILLE,Egis rail.

Thèse soutenue à Laval, le 4 avril 2025, par

Sondoss CHTIOUI

Composition du jury
Membres du jury avec voix délibérative

Jean-Yves CHOLEY Professeur, Quartz – SUPMECA de Paris	Président
Marie-Éléonore KESSACI Professeure, CRISTAL – Université de Lille	Rapporteuse et Examinatrice
Mahdi ZARGAYOUNA Directeur de recherche, GRETTIA – Université Gustave Eiffel	Rapporteur et Examinateur
Florence OSSART Professeure, Université Paris-Saclay - CNRS	Examinatrice

Thèse de doctorat

NNT : 2025UPAST028

université PARIS-SACLAY

ÉCOLE DOCTORALE
Physique et ingénierie : Electrons, Photons, Sciences du vivant (EOBE)

Titre : Utilisation innovante des données opérationnelles par l'intelligence artificielle dans un poste de contrôle ferroviaire visant l'amélioration de l'efficacité opérationnelle

Mots clés : Poste de Commandement Centralisé, réseaux de tramways, réseaux RER, données d'exploitation, système d'aide à la prise de décision, prévisions court terme, préparation des données, modèles d'apprentissage

Résumé : Ce projet se concentre sur l'optimisation des opérations ferroviaires, en particulier pour les réseaux de tramways et de RER. L'objectif est de proposer une approche permettant l'intégration d'un système d'aide à la prise de décision dans un poste de contrôle ferroviaire (PCC). Les opérateurs doivent garantir la sécurité, optimiser le débit et maintenir la qualité de service, tout en offrant une expérience satisfaisante pour les passagers. Dans ce contexte, le système de gestion du trafic, sous la direction du PCC, cherche à minimiser les retards et à gérer efficacement les flux de passagers. Cependant, la complexité du réseau et la forte densité du trafic, avec des intervalles courts entre les trains, représentent des défis importants.

Pour répondre à ces enjeux, nous avons développé un système d'aide à la prise de décision basé sur des modèles d'apprentissage automatique et des méthodes mathématiques en temps réel. Ce système vise à fournir des prédictions précises sur les retards des trains, anticipant ces derniers jusqu'à 15 minutes avant leur arrivée. L'objectif principal est de soutenir l'opérateur de PCC dans la gestion proactive des retards et de l'afflux de passagers sur les quais

Le principal défi auquel fait face le PCC est la gestion des retards, en particulier pendant les heures de pointe. Les retards ont tendance à s'accumuler, créant un effet domino qui perturbe l'ensemble du réseau. Malgré les procédures existantes pour atténuer ces retards, la contrainte temporelle pour réguler les lignes reste un obstacle majeur, affectant à la fois les trains précédents et suivants, se qui impacte grandement la qualité du service. Ainsi, notre méthodologie s'articule autour de plusieurs étapes et se divise en deux modes : hors-ligne et en ligne.

1. Mode hors-ligne : Le mode hors-ligne commence par un nettoyage approfondi des données d'exploitation pour garantir leur qualité et leur fiabilité. Ces données sont ensuite préparées en intégrant des informations passées, présentes et futures. Le passé comprend les retards historiques et les informations concernant les trains prédécesseurs et successeurs, analysées à l'aide d'une fenêtre glissante. Le présent représente les caractéristiques opérationnelles actuelles, telles que les états des trains et les conditions du réseau. Enfin, le futur anticipe les évolutions des trains sur les stations suivantes, avec un horizon de prédiction fixé à 15 minutes. L'entraînement est effectué périodiquement, en fonction des besoins spécifiques et de l'arrivée de nouveaux ensembles de données. 2. Mode en ligne : En mode en ligne, des données simulées en temps réel sont utilisées en l'absence d'accès direct au PCC. Ces nouvelles données sont prétraitées pour garantir leur cohérence avec celles du mode hors-ligne. Une interface utilisateur ergonomique et une animation en temps réel permettent de suivre les mouvements des trains, d'anticiper les retards et de fournir des prédictions adaptées aux opérations en cours.

Une étude comparative des modèles d'apprentissage automatique et profond a été menée pour identifier les plus adaptés à la prédiction des retards dans les transports publics. Ces modèles ont été optimisés grâce à des techniques d'ajustement des hyperparamètres, atteignant un taux de précision de 95 % pour des prédictions jusqu'à 15 minutes avant l'arrivée des trains. Une étude de sensibilité a également été réalisée pour analyser l'impact des hyperparamètres sur la convergence des modèles et améliorer leur efficacité sur les données ferroviaires.

Title : Innovative Use of Operational Data Through Artificial Intelligence in a Railway Control Center to Improve Operational Efficiency **Keywords :** Central Command Post, tramway networks, RER networks, operational data, decision support system, Short-term forecasts, data preparation.

Abstract : This project focuses on optimizing railway operations. The objective is to propose an approach that enables the integration of a decision support system in an operations control center (OCC). Operators are tasked with ensuring safety, optimizing throughput, and maintaining service quality, all while providing a satisfactory passenger experience. Within this context, the traffic management system seeks to minimize delays and effectively manage passenger flows. However, the complexity of the network and the high traffic density, with short intervals between trains, present significant challenges.

To address these issues, we developed a decision support system based on machine learning models and real-time mathematical methods. This system is designed to provide accurate predictions of train delays, with forecasts available up to 15 minutes before train arrival. The main goal is to support the OCC in proactively managing delays and passenger congestion on platforms. By offering reliable predictions, the system enables the OCC to take anticipatory actions to minimize the impact of delays, improve passenger flow organization, and ultimately optimize overall operational efficiency.

The main challenge the OCC faces is delay management, particularly during peak hours. Delays tend to accumulate, creating a domino effect that disrupts the entire network. Despite existing procedures to mitigate these delays, the time constraint for regulating the lines remains a major obstacle, affecting both preceding and following trains, which significantly impacts service quality. Thus, our methodology is structured around several steps and is divided into two modes : offline and online.

1. Offline Mode : The offline mode begins with an in-depth cleaning of operational data to ensure quality and reliability. These data are then prepared by incorporating past, present, and future information. The past includes historical delays and details about the predecessor and successor trains, analyzed using a sliding window. The present represents current operational characteristics such as train statuses and network conditions. Finally, the future anticipates the behavior of trains at upcoming stations with a prediction horizon set to 15 minutes. This systematic preparation enables the chosen machine learning model to be trained on the entire historical dataset. Training is conducted periodically, based on specific needs and the availability of new data.

2. Online Mode : In online mode, simulated real-time data is used without direct access to the OCC. The new data undergoes preprocessing to ensure consistency with the offline mode data. An ergonomic user interface and real-time animation help track train movements, anticipate delays and provide predictions suited to current operations.

A comparative study of machine learning and deep learning models was conducted to identify the best models for predicting delays in public transportation. These models were optimized using hyperparameter tuning techniques, achieving an accuracy rate of 95 % for predictions up to 15 minutes in advance of train arrivals. A sensitivity study also analyzed the impact of hyperparameters on model convergence to improve efficiency with railway data.

In summary, this project offers a comprehensive approach to railway traffic optimization. By integrating machine learning techniques and real-time data analysis, the decision support system provides operators with the tools to manage delays effectively and improve the overall passenger experience. This approach not only enhances the operational efficiency of the control center but also contributes to more reliable and smoother railway operations.

Maison du doctorat de Université Paris-Saclay
2e étage, aile ouest, École normale supérieure Paris-Saclay
4 avenue des Sciencs
91190 Gif-sur-Yvette, France

Je dédie cette thèse à mes parents.

Le progrès, ce n'est pas de remplacer l'homme, mais de l'aider à mieux se réaliser.

— Antoine de Saint-Exupéry

Remerciements

Au terme de ce travail de thèse, je souhaite exprimer ma profonde gratitude à toutes les personnes qui m'ont accompagnée, soutenue, inspirée et guidée au cours de ces années riches en apprentissages, en défis, mais aussi en accomplissements personnels et professionnels.

Ce parcours a été bien plus qu'un simple effort académique. Il a représenté pour moi une traversée d'expériences humaines, de remises en question, d'efforts constants et de découvertes. Il a aussi été une école de résilience, de patience, et de maturité intellectuelle. Cette partie du manuscrit est donc dédiée à celles et ceux qui ont contribué, de près ou de loin, à faire de cette aventure une réussite.

Je remercie tout particulièrement Toufik Azib, mon directeur de thèse, pour sa confiance, sa vision, son exigence bienveillante et son engagement continu. Ses retours méthodiques, sa capacité à orienter avec clarté et sa rigueur scientifique ont été décisifs à chaque étape clé de ce travail.

Je remercie également Sébastien Saudrais et Sebti Mouelhi, mes encadrants académiques, pour leur engagement, leur disponibilité et leur accompagnement attentif à chaque étape. Leurs conseils méthodologiques et leurs regards complémentaires ont été essentiels à la structuration de cette thèse. Ils m'ont aidée à relier les enjeux scientifiques aux applications concrètes, en gardant toujours le cap sur la cohérence du travail.

Ce travail a été mené dans le cadre CIFRE avec EGIS Rail, que je remercie pour m'avoir offert un environnement propice à la recherche appliquée. Je pense en particulier à Marc Ille, David Chouvin, Helmut Rasamimanana, Olivier Meriot et toute l'équipe opérationnelle avec laquelle j'ai eu le plaisir de travailler. Leur expertise, leur confiance et leur ouverture ont fortement enrichi ma compréhension des enjeux du monde ferroviaire.

Je remercie également l'ESTACA, en particulier l'ensemble du corps enseignant et des équipes pédagogiques, et tout spécialement Sandrine Pioget, pour leur accompagnement, leur ouverture d'esprit et la richesse de nos échanges. L'ESTACA m'a offert un cadre académique à la fois rigoureux et humain, dans lequel j'ai pu évoluer et faire mes preuves.

Je n'oublie pas les membres du jury et rapporteurs, que je remercie pour leur disponibilité, la qualité de leur lecture et leurs retours précieux.

Enfin, je souhaite exprimer toute ma gratitude à ma famille, qui a toujours cru en moi, même dans les moments de doute. Leur soutien moral, leur patience et leur amour m'ont donné la force de continuer chaque jour. Un merci tout particulier à mes proches qui, sans nécessairement comprendre tous les détails de cette recherche, n'ont jamais cessé de me pousser à donner le meilleur de moi-même.

Table des matières

Remerciements **6**

1 Introduction **11**

1.1 Contexte et motivation 11

1.2 Démarche proposée et travaux réalisés 14

1.3 Organisation du manuscrit 14

2 État de l'Art **17**

2.1 Transport ferroviaire urbain 17

2.1.1 Histoire du ferroviaire urbain 18

2.1.2 Contrôle de trafic centralisé (CTC) 19

2.1.2.1 Évolution de la supervision du trafic ferroviaire 19

2.1.2.2 Technologies de contrôle et de supervision 20

2.1.3 Timetable Management (TTM) dans les systèmes ferroviaires 20

2.1.4 Discussion 21

2.2 Systèmes d'aide à la prise de décision dans un PCC 22

2.2.1 Présentation générale des systèmes d'aide à la prise de décision (SAPD) 22

2.2.2 Intégration du SAPD dans un PCC 23

2.2.3 Une vision améliorée de la régulation du trafic ferroviaire grâce au SAPD 25

2.2.4 Discussion 25

2.3 Intelligence artificiel en vue de l'intégration d'un SAPD dans PCC 27

2.3.1 Concepts et cadre général de l'intelligence artificielle 27

2.3.2 Travaux menés dans l'Intelligence Artificielle appliquée au secteur des transports publics 30

2.3.3 Travaux menés dans les méthodes d'évaluation et de validation des modèles 36

2.3.4 Contributions existantes relatives aux modèles d'apprentissage 43

2.3.4.1 Méthodes statistiques 44

2.3.4.2 Méthodes à base d'arbres de décision et d'optimisation bayésienne 44

2.3.4.3 Méthodes d'apprentissage profond 48

2.3.5 Discussion 51

2.4 Conclusion 53

3 Méthodologie : système d'aide à la prise de décision **55**

3.1 Démarche proposée 58

3.2 Approche relationnelle pour la modélisation des données ferroviaires . . 61
3.2.1 Méthode de travail 63
3.2.2 Création du modèle relationnel 64
3.2.3 Cas d'usage 67
3.2.4 Représentation mathématique des données d'exploitation XOP 69
3.3 Prétraitement des données 70
3.3.1 Nettoyage des données 71
3.3.2 Correction des incohérences temporelles 73
3.3.3 Gestion des valeurs aberrantes 77
3.3.4 Enrichissement des données 80
3.3.5 Mise en forme et encodage des variables pour l'entraînement 80
3.4 Analyse des données 82
3.5 Ingénierie des caractéristiques temporelles 86
3.5.1 Structuration temporelle : impact du passé et projection vers le futur . . 86
3.5.2 Données d'entrée : Comportement historique des rames dans une fenêtre temporelle Y^{past} 87
3.5.3 Données d'entrée : Anticipation des rames sur un horizon temporel futur XOP^{+f} 88
3.6 Mécanisme d'Entraînement et Optimisation des Modèles 90
3.7 Évaluation des modèles 91
3.8 Analyse de Sensibilité 93
3.9 Conclusion 94

4 Analyse et résultats **97**
4.1 Cadre de travail 98
4.1.1 Environnement de travail 98
4.1.2 Sources de données (Tram, RER) 100
4.1.3 Présentation des données collectées : organisation et format 101
4.2 Mise en forme et préparation des données 103
4.3 Traitement analytique 104
4.3.1 Relation entre les retards et les variables d'exploitation (XOP) 107
4.3.2 Impact des variables passées sur l'analyse des retards (Y^{past}) 110
4.4 La prévision à court terme du retard : une première analyse exploratoire (tramway) 116
4.4.1 Partitionnement des données pour l'entraînement et la validation : Tramway 116
4.4.2 Méthodes d'entraînement 117
4.4.3 Ajustement des hyperparamètres 119
4.4.4 Performance et convergence des modèles sur l'ensemble d'entraînement 119

4.4.5 Evaluation et comparaison des modèles sur l'ensemble du test 122
4.4.6 Analyse comparative et discussion 123
4.5 Analyse de sensibilité 125
4.5.1 La prévision à court terme du retard : Une 2eme analyse exploratoire (RER) 125
4.5.2 Sensibilité des performances des modèles à la standardisation des infrastructures Tramway-RER 129
4.6 Conclusion 130

5 Implémentation du prototype (Système d'Aide à la Prise de Décision) 135
5.1 Conception de l'interface homme machine 136
5.1.1 Maquette 137
5.1.2 Fonctionnalités clés avec diagramme des cas d'utilisation 138
5.2 Architecture générale du prototype complet (IHM) 139
5.3 Phase de Développement du Prototype 141
5.3.1 Diagramme de séquence : Création de la ligne à simuler 142
5.3.2 Diagramme de séquence : Caractéristiques de la simulation 144
5.3.3 Diagramme de séquence : Caractéristiques + Prédiction des retards . . 145
5.3.4 Diagramme de séquence : Animation des rames avec un code couleur . 146
5.4 Mise en œuvre sur un cas test : Scénario simplifié 148
5.4.1 Interface d'utilisation 149
5.4.2 Analyse de la simulation ferroviaire issue des recommandations de ChatGPT 152
5.4.3 Analyse des contraintes et évaluation de l'IHM (Questionnaire destiné aux anciens opérateurs et experts en systèmes ferroviaires d'Egis Rail . 154
5.5 Conclusion 157

6 Conclusion et Perspectives 159
6.1 Contributions 159
6.2 Réflexion sur les défis rencontrés et les solutions apportées 161
6.3 Perspectives 162

Bibliographie 173

Liste des figures 177

Liste des tableux 179

Liste des mots-clés 181

1 Introduction

Ce chapitre présente l'introduction générale de la méthodologie développée au cours de cette thèse. Il expose les principes fondamentaux et les approches employées pour répondre aux problématiques identifiées, tout en soulignant les choix méthodologiques effectués pour garantir la pertinence et l'efficacité des solutions proposées.

1.1 Contexte et motivation

Le secteur ferroviaire connaît un développement marqué par des évolutions profondes à travers le monde. Cette dynamique s'accompagne de nombreux investissements visant à créer de nouvelles infrastructures et à améliorer la capacité des structures existantes. En France, par exemple, les investissements nets de la SNCF ont dépassé 5,2 milliards d'euros en 2019, avant la crise sanitaire, tandis qu'ils s'élevaient à environ 3,5 milliards d'euros en 2022, reflétant une reprise progressive après la pandémie de COVID-19 [1]. L'objectif est de promouvoir la qualité de service et d'accroître la satisfaction des clients. Pour relever ces nouveaux défis, notamment en matière d'optimisation et d'aide à la décision, de nouveaux outils informatiques s'avèrent indispensables.

Par ailleurs, les perturbations dans les réseaux de métro et tramway en France représentent une problématique croissante, particulièrement dans les grandes agglomérations. Entre 2019 et 2023, d'importantes disparités régionales dans le taux d'annulation ont été observées, la Bretagne affichant le meilleur taux (1,3 %) tandis que l'Occitanie présente le plus élevé (3,5 %), reflétant des écarts similaires à ceux de la ponctualité [2]. À titre d'exemple, en Île-de-France, la ponctualité des RER (arrivée avec un retard inférieur à 5 minutes) est passée de 94,5 % en 2015 à 85,1 % en 2023, reflétant une dégradation progressive des performances [3]. Ces retards sont principalement causés par la saturation des lignes, les dysfonctionnements techniques, et l'interconnexion complexe avec d'autres modes de transport.

De plus, les pannes récurrentes et les conditions météorologiques extrêmes exacerbent la situation, affectant particulièrement les régions densément peuplées. La modernisation insuffisante des infrastructures accentue ces difficultés, entraînant des perturbations régulières et une diminution de la satisfaction des usagers. Ces observations soulignent la nécessité d'une approche proactive pour améliorer la régularité des services et garantir une meilleure expérience aux voyageurs.

Le service au sein d'un centre de contrôle ferroviaire est exigeant, car il doit assurer la fluidité du trafic, la ponctualité des rames et le confort des voyageurs. Chaque jour, le réseau ferroviaire est surveillé en temps réel grâce à des systèmes informatiques

sophistiqués fournissant des informations détaillées notamment sur la position des rames, l'état des signaux, et des voies. Lorsqu'un incident survient, une évaluation rapide de la situation est indispensable pour identifier les causes possibles (comme des défaillances techniques, des conditions météorologiques défavorables, ou des problèmes de signalisation) et optimiser la régulation du trafic afin de limiter les sur-incidents liés à une sur-fréquentation de certaines rames. Cela peut inclure la modification des itinéraires, la coordination avec d'autres centres de contrôle, ou la gestion des priorités de circulation des rames. Chaque décision tient compte de la sécurité des voyageurs, de la ponctualité des services, ainsi que du respect des réglementations en vigueur. Le retard, en tant qu'indicateur clé, permet à l'exploitant d'évaluer la capacité des trains à assurer la prochaine course au départ du prochain terminus. Lorsqu'un train accumule du retard, cela peut entraîner un effet boule de neige : l'augmentation du nombre de voyageurs en attente dans chaque station allonge le temps de fermeture des portes, ce qui accentue encore davantage le retard.

Une ligne de Métro/Tram est exploitée en carrousel : un train (rame) part d'un terminus, dessert chaque station de la ligne sur la voie sur laquelle il circule jusqu'au terminus opposé, puis revient à son terminus de départ en s'arrêtant à toutes les stations sur l'autre voie. Les trains se succèdent tout au long de la journée depuis l'ouverture de la ligne au public jusqu'à sa fermeture au public. Le nombre de trains varient en fonction de l'heure pour adapter l'offre de transport aux heures creuses et heures de pointe. Chaque train (rame) part de chaque terminus à une heure déterminée. Si un train subit un incident, la durée de son parcours jusqu'au terminus suivant est rallongée. Il ne pourra donc pas partir du terminus suivant à son heure prédéterminée. La différence de temps entre son heure de départ prévue et son heure de départ réelle est appelée "retard". C'est une notion purement exploitation qui sert avec d'autre éléments à mesurer la qualité de service. La ligne étant exploitée en carrousel et en intervalle, un client voit une fréquence pas une heure donc pas de retard.

Les incidents génèrent des retards, nécessitant une régulation en temps réel de la part de l'exploitant. Pour une gestion efficace, l'opérateur doit également anticiper les incidents potentielles en s'appuyant sur son expérience ainsi que sur l'analyse des données historiques. Cela implique de prévoir l'impact d'un retard sur les rames suivantes et d'adapter les plannings en conséquence. Cette tâche est rendue encore plus complexe par la nécessité de maintenir une communication constante avec les équipes sur le terrain, les autres postes de contrôle, et parfois même les passagers. Le travail quotidien d'un opérateur de poste de contrôle ferroviaire exige donc une grande réactivité, une capacité d'analyse rapide, et une prise de décision appropriée malgré la pression pour assurer la meilleure qualité de service.

Une bonne régulation du trafic ferroviaire revêt une importance cruciale pour garantir un service efficace, fiable et adaptée aux besoins des voyageurs et conforme aux exigences de l'autorité des transport. En optimisant la gestion des rames en circulation selon la demande, il est possible de mieux répondre aux besoins quotidiens des clients tout en évitant la surcharge du réseau. Cette approche contribue à améliorer la ponctualité, à offrir une meilleure qualité de service, et à maximiser l'utilisation des infrastructures existantes. Par ailleurs, une régulation optimisée permet de réduire l'empreinte environnementale du secteur ferroviaire en minimisant les trajets inutiles, en diminuant les consommations d'énergie, et en limitant les émissions de gaz à effet de serre, contribuant ainsi à une mobilité plus durable et respectueuse de l'environnement.

Chaque jour, la circulation des rames génère un vaste ensemble de données d'exploitation, regroupant des informations cruciales telles que les heures de départ et d'arrivée prévues et réelles, les emplacements des rames, ainsi que les retards enregistrés au cours de la journée. À cela s'ajoutent les données issues des systèmes de surveillance et de billettique, fournissant des informations précieuses sur les flux de voyageurs sur les quais et à bord des rames [4]. Cependant, malgré leur richesse, ces données historiques restent largement sous-utilisées une fois leur fonction initiale — la régulation immédiate du trafic — accomplie. Elles sont généralement stockées sans véritable valorisation ni exploitation ultérieure, limitant ainsi leur potentiel à améliorer les stratégies de régulation, de maintenance préventive, ou encore d'optimisation des flux voyageurs [5].

Ce manque d'exploitation pose un double défi : d'une part, il empêche une meilleure compréhension des dynamiques du réseau ferroviaire sur le long terme ; d'autre part, il prive les gestionnaires de postes de contrôle d'outils analytiques capables d'anticiper et de résoudre les perturbations de manière proactive. Cette problématique met en lumière la nécessité de développer des systèmes capables d'intégrer et de valoriser ces données historiques afin d'optimiser les performances du réseau ferroviaire et de répondre aux besoins croissants des usagers [6].

Notre travail se base sur l'utilisation de données d'exploitation non utilisées, qui au travers de l'IA vont proposer à l'Opérateur du Centre de Contrôle une aide à la décision. Ces données seront affichées sur un moniteur supplémentaire dédié à cet usage. Ce système vise non seulement à fournir une vue analytique sur les performances passées, mais également à anticiper les problèmes potentiels, tels que les retards ou les pannes, grâce à des modèles d'apprentissage automatique capables de produire des prévisions fiables sur le futur proche de la circulation ferroviaire [6].

Cet écran supplémentaire repose sur le principe des systèmes d'aide à la prise de décision (SAPD), intégrant des outils avancés pour réguler le trafic, gérer les retards, et optimiser les flux voyageurs. En exploitant les données historiques et en analysant les besoins réels des voyageurs présents sur les quais, il permet d'ajuster dynamiquement le nombre de rames en circulation. De plus, le système propose une anticipation proactive des perturbations, qu'il s'agisse de retards ou d'avances. Ces informations, transmises en temps réel, offrent aux opérateurs les moyens de réguler efficacement le trafic en appliquant des procédures adaptées, minimisant ainsi l'impact des perturbations sur le réseau.

Ce projet, fruit d'une collaboration entre l'ESTACA et Egis Rail, vise à intégrer et montrer une des applications possibles de l'intelligence artificielle dans les postes de commande centralisés (PCC) afin d'améliorer la gestion dans les PCC. Un écran supplémentaire, à la fois ergonomique et intuitif, a été développé pour accompagner les opérateurs en leur fournissant des estimations de retard précises, destinées à faciliter la gestion des perturbations. Il est proposé que cette interface soit affichée sur un écran dédié, distinct des autres outils de supervision. Ces prévisions, générées par un modèle d'intelligence artificielle exploitant les données historiques des rames ainsi que les caractéristiques des infrastructures, devraient favoriser des prises de décisions plus rapides et fiables. Inspiré des outils existants, cet écran est conçu pour être facilement accepté par les opérateurs et constituer un soutien efficace dans la gestion des environnements ferroviaires complexes.

1.2 Démarche proposée et travaux réalisés

Nous présentons maintenant les travaux réalisés et la démarche proposée pour alimenter le système d'aide à la prise de décision.

La première étape consiste à comprendre la nature des données avec lesquelles nous travaillons, en utilisant un modèle relationnel qui visualise les liaisons existantes entre ces données. Cela nous permet de préparer, nettoyer et enrichir les données avant d'identifier les facteurs essentiels perturbant une ligne ferroviaire dans le contexte des transports publics.

Une fois les données prêtes à être utilisées, le défi consiste à explorer l'espace des variables d'ajustement pour déterminer celles qui influencent le retard sur la ligne ferroviaire, tout en tenant compte des contraintes identifiées lors de l'analyse de l'état de l'art.

Nous avons proposé une méthodologie originale en trois étapes. La première étape consiste à effectuer une analyse afin de déterminer le type de résultats que nous souhaitons obtenir, qu'il s'agisse de prédictions à court ou à long terme, en fonction du jeu de données utilisé. Après avoir choisi le type de prédiction, nous avons appliqué le principe selon lequel le passé impacte le présent et le futur, en réalisant une ingénierie des variables pour l'ensemble des données. Une étude comparative a été menée pour identifier le modèle le plus performant sur les données ferroviaires, en tenant compte de différents critères de performance.

La deuxième étape, après avoir sélectionné le modèle optimal qui maximise les performances sur les données ferroviaires, consiste à réaliser une étude de sensibilité de notre méthodologie. Cela implique de tester la méthode sur une infrastructure différente pour vérifier si le principe selon lequel le passé influence le présent et le futur s'applique de manière cohérente à divers types d'infrastructures ferroviaires de transports publics.

La troisième étape consiste à intégrer le modèle choisi lors de la première étape dans un prototype appelé "écran supplémentaire". Ce prototype sert de simulation en temps réel utilisant les données de la ligne étudiée, permettant de suivre la circulation des rames selon les horaires réels et prévus. Les prédictions sont affichées sur l'écran, indiquant les retards à venir, offrant ainsi un outil de visualisation et d'aide à la prise de décision en temps réel pour les opérateurs.

1.3 Organisation du manuscrit

Le manuscrit s'organise en plusieurs chapitres, chacun traitant un aspect essentiel du projet, comme suit :

1. État de l'Art : Ce chapitre propose une introduction au contexte général des transports ferroviaires, abordant l'histoire, la mission et la régulation du trafic ferroviaire dans les PCC. Il présente également les systèmes d'aide à la prise de décision existants, une revue de la littérature sur les prédictions de retards et les technologies en place, ainsi que les modèles d'apprentissage automatique utilisés dans ce domaine. Cette exploration nous a permis de clarifier notre positionnement scientifique ainsi que l'élaboration de l'architecture de notre outil (approche d'aide à la décision) ciblé.

2. Méthodologie (Système d'Aide à la Prise de Décision) : Ce chapitre détaille la démarche méthodologique adoptée, à commencer par la collecte, le nettoyage et

l'enrichissement des données (tramways et RER). Il décrit ensuite la modélisation relationnelle, l'analyse des données, et les approches de conception des caractéristiques pour la modélisation prédictive. Il présente les différentes méthodes d'apprentissage utilisées (statistiques, arbres de décision, apprentissage profond) et les critères de performance adoptés.

3. Analyse et Résultats : Ce chapitre expose l'environnement de travail, la présentation des données collectées, et les analyses menées sur les variables influençant les retards. Il compare les performances des modèles, ajuste les hyperparamètres et évalue les résultats à travers une étude de sensibilité sur différentes infrastructures, telles que les tramways et le RER.

4. Implémentation du Prototype (SAPD) : Ce chapitre décrit la conception de l'interface homme-machine (IHM), la maquette, et l'architecture générale du prototype. Il aborde également le développement de différents scénarios d'utilisation, y compris les diagrammes de séquence, les caractéristiques du système, et les animations en temps réel. Une évaluation est menée pour valider les résultats obtenus.

5. Conclusion et Perspectives : Ce dernier chapitre résume les contributions du projet, réfléchit sur les défis rencontrés et propose des solutions, ainsi que des perspectives d'amélioration et de futurs travaux pour étendre la portée de l'outil proposé.es, et les résultats des tests menés pour valider son efficacité en conditions réelles.

2

État de l'Art

Sommaire

1.1 Contexte et motivation 11
1.2 Démarche proposée et travaux réalisés 14
1.3 Organisation du manuscrit 14

Ce chapitre présente une revue de l'état de l'art dans le domaine du transport ferroviaire, en mettant l'accent sur les postes de contrôle centralisés (PCC) et l'intégration potentielle d'un Système d'Aide à la Prise de Décision (SAPD). L'objectif est de comprendre les défis et les opportunités liés à cette intégration, tout en explorant les approches existantes pour prédire et gérer les retards.

La première partie aborde une vue d'ensemble du transport urbain, incluant l'évolution historique et les concepts modernes comme la gestion centralisée du trafic et les systèmes de management des horaires (Timetable Management).

La deuxième partie se concentre sur les systèmes d'aide à la prise de décision dans les PCC. Elle explore les concepts fondamentaux de ces systèmes, les cas d'utilisation pour une prise de décision en temps réel, et les approches déjà testées dans des environnements similaires. Ces analyses fournissent des pistes pour concevoir un SAPD adapté aux besoins spécifiques des réseaux ferroviaires.

En parallèle, une analyse approfondie des modèles d'apprentissage machine est réalisée, mettant en avant les techniques les plus prometteuses pour prédire les retards et améliorer les processus de prise de décision. Cette étude est complétée par les informations fournies par Egis-rail, incluant des fichiers internes détaillant les processus opérationnels, les contraintes de régulation du trafic, et les facteurs influençant les retards. Ces données internes sont essentielles pour enrichir la compréhension des défis opérationnels et guider le développement de solutions adaptées.

2.1 Transport ferroviaire urbain

Comme tout système industriel, le système ferroviaire comporte trois composantes principales : les opérateurs, les outils et les procédures. L'organisation définit les interactions entre ces éléments. Les outils et procédures permettent d'exploiter efficacement le réseau ferroviaire, tandis que les opérateurs, formés aux procédures, utilisent ces outils pour assurer un fonctionnement sûr et optimal.

La sécurité technique joue un rôle clé dans ce cadre. Elle garantit la préservation des caractéristiques techniques et fonctionnelles nécessaires à la circulation des trains

sur l'infrastructure ferroviaire. Cette sécurité repose principalement sur le travail des opérateurs d'entretien et de maintenance, qui assurent la fiabilité du matériel roulant ainsi que des installations fixes.

Ces interactions, essentielles au bon fonctionnement du réseau ferroviaire, ont été profondément transformées par l'évolution technologique. Les innovations, notamment dans les domaines de la supervision automatique et du contrôle du trafic, ont permis d'automatiser de nombreuses tâches et de centraliser la gestion des opérations dans des postes dédiés. Cette section explore en détail ces transformations et met en lumière les principaux systèmes technologiques ayant redéfini l'organisation et le fonctionnement du système ferroviaire moderne.

2.1.1 Histoire du ferroviaire urbain

Le développement du transport ferroviaire urbain a joué un rôle crucial dans l'évolution des villes modernes, offrant des solutions efficaces pour la mobilité des populations croissantes. Les premières initiatives remontent au début du XIX^{e} siècle, avec l'introduction de lignes de chemin de fer pour le transport de marchandises, notamment dans les régions minières. En France, la ligne Saint-Étienne-Andrézieux, inaugurée le 1er janvier 1828, est un exemple précoce de ce type d'infrastructure [7]. La véritable expansion des chemins de fer urbains s'est produite avec l'électrification et l'urbanisation rapide des villes. À Paris, le chemin de fer de Petite Ceinture, conçu à l'origine pour des raisons militaires, a été mis en service progressivement à partir de 1852, formant une boucle autour de la ville et facilitant le transport de passagers et de marchandises [8]. Cette ligne a joué un rôle essentiel dans la connexion des différentes gares parisiennes et le désenclavement des quartiers périphériques.

L'introduction des tramways a également marqué une étape significative dans l'histoire du transport ferroviaire urbain. À Paris, les premiers tramways à chevaux ont été mis en service en 1855, suivis par des tramways électriques à partir de 1898 [9]. Cette transition vers l'électrification a permis une augmentation de la capacité et de la vitesse des services, répondant ainsi aux besoins croissants de mobilité urbaine. Le développement du métro parisien, inauguré en 1900, a constitué une avancée majeure dans le transport ferroviaire urbain. Conçu pour désengorger les rues encombrées et offrir un moyen de transport rapide à travers la ville, le métro est devenu un modèle pour de nombreuses autres métropoles [10].

D'autres villes françaises ont suivi cette tendance. Par exemple, le tramway de Strasbourg, initialement mis en service en 1878 avec des véhicules hippomobiles, a évolué vers un réseau électrifié étendu, bien que son exploitation ait cessé en 1960 avant d'être relancé en 1994 [11]. De même, Le Havre a inauguré son premier tramway électrique en 1894, devenant ainsi la première grande ville française à disposer d'un réseau entièrement électrifié [12].

En somme, l'histoire du transport ferroviaire urbain en France reflète une adaptation continue aux besoins de mobilité des populations urbaines, marquée par des innovations technologiques et une expansion progressive des réseaux pour améliorer l'accessibilité et la qualité de vie en milieu urbain.

2.1.2 Contrôle de trafic centralisé (CTC)

Le contrôle de trafic centralisé (CTC), également connu sous le nom de contrôle centralisé du trafic (CCT) chez Alstom [13], [14], est un système de gestion et de régulation du trafic ferroviaire à partir d'un poste centralisé appelé « poste de commandement ». Ce dispositif, initialement développé en Amérique du Nord, vise à optimiser la gestion du trafic ferroviaire et à résoudre les problèmes opérationnels qui en découlent [15].

Le CTC contrôle à distance les signaux et les aiguillages d'un tronçon grâce à des systèmes automatiques informatisés. Le trafic ferroviaire est visualisé sur des écrans où les opérateurs du centre de contrôle peuvent observer la circulation des rames [16]. À partir de ces écrans, ils disposent de commandes pour établir des itinéraires dédiés à chaque train.

2.1.2.1 Évolution de la supervision du trafic ferroviaire

Depuis quelques années, la supervision du trafic ferroviaire a été considérablement transformée. Autrefois bâtie sur l'omniprésence de l'opérateur humain dans les activités ferroviaires, depuis le terrain jusqu'au poste de régulation, l'arrivée des systèmes informatisés de surveillance et de commande a réduit considérablement le nombre d'opérateurs ferroviaires, notamment dans la boucle de supervision [17].

Les systèmes modernes de supervision du trafic ferroviaire, appelés *Automatic Train Supervision* (ATS), tendent à centraliser toute la commande ferroviaire dans un seul poste, appelé « Poste Centralisé de Commandement » (PCC). Ces systèmes assurent plusieurs fonctions clés :

- Contrôle automatique de l'évolution du trafic.
- Commande des itinéraires de circulation.
- Communication avec les circulations.
- Gestion de l'énergie électrique de traction.
- Gestion des ressources (matériel roulant et personnel).
- Contrôle des systèmes auxiliaires (ventilations, escalators, etc.).
- Informations aux passagers.
- Sûreté publique (sécurité, malveillance).

Autrefois garant de la sécurité, certaines fonctions sont confiées à des systèmes techniques de sécurité toujours plus autonomes. Au-delà des systèmes d'enclenchement, qui garantissent la sécurité des circulations sur les voies depuis le début du siècle dernier, des systèmes de protection appelés *Automatic Train Protection* (ATP) ont vu le jour. Ces systèmes offrent un niveau de sécurité tel que l'ATP (Automatic Train Protection) est considéré comme l'acteur principal de la sécurité, tandis que l'ATS (Automatic Train Supervision) n'est plus perçu comme un élément central dans ce domaine [18]. Les opérateurs de l'ATS sont désormais principalement confinés à des tâches de surveillance. Il est important de distinguer les fonctions de l'ATS et de l'ATP : l'ATS gère et contrôle l'exploitation des trains, notamment la commande des itinéraires, le suivi des rames et la régulation du trafic. En revanche, l'ATP est dédié au contrôle de la vitesse des trains, en surveillant les dépassements de vitesse autorisée et en assurant l'arrêt automatique des trains en cas de besoin. Au sommet de l'architecture du *Automatic Train Control* (ATC), le système ATS est intégré dans le PCC (Poste de Commande Centralisé) et offre de nombreuses fonctionnalités, rendues possibles par la généralisation des nouvelles technologies de l'information. L'ATC englobe également l'ATO (Automatic Train Operation), qui gère des fonctions telles que l'arrêt des rames à

quai et la gestion des portes des rames [19].

2.1.2.2 Technologies de contrôle et de supervision

L'automatisation du mouvement des trains a été motivée dès le départ par des considérations de capacité de transport et de qualité de service. Cette automatisation permet de réduire les intervalles entre chaque train, diminuant ainsi le temps d'attente des voyageurs. Le retour d'expérience sur les lignes équipées montre que le trafic s'en trouve fluidifié, avec des intervalles entre trains plus respectés [20].

Les fonctions sécuritaires intégrées aux systèmes de contrôle automatique des trains permettent d'intervenir en cas d'erreur humaine pouvant compromettre la sécurité [21]. Dans les systèmes d'automatisation intégrale, tout risque d'erreur de conduite est écarté. Là encore, le retour d'expérience confirme que le niveau global de sécurité est amélioré. À ce jour, les accidents graves impliquant des métros automatiques sont principalement liés à des erreurs humaines lors de travaux à proximité du système [22].

Avec le développement des technologies ferroviaires, les systèmes modernes tels que le *Communication-Based Train Control* (CBTC) ont vu le jour et sont de plus en plus utilisés. Contrairement à la signalisation ferroviaire conventionnelle, le CBTC utilise un système de positionnement à haute résolution, indépendant des circuits ferroviaires ou des compteurs d'essieux. Ce système permet un échange continu et bidirectionnel de données à grande capacité, tout en assurant des fonctions vitales [23].

Le CBTC intègre des modules ATP, ATO, et ATS. La localisation des trains, assurée par des systèmes embarqués, est transmise aux équipements au sol via des canaux de communication radio. Ces fonctionnalités permettent de réduire les intervalles entre les trains tout en garantissant un niveau élevé de sécurité [24]. En cas d'échec de transmission, par exemple en raison d'erreurs ou de retards, le système déclenche un freinage d'urgence pour éviter tout risque inacceptable [25].

Par ailleurs, la signalisation ferroviaire européenne souffre d'une fragmentation due à des différences techniques entre réseaux, telles que l'écartement des rails ou les types de courant électrique. Le système *European Rail Traffic Management System* (ERTMS) vise à standardiser ces multiples systèmes pour améliorer l'interopérabilité. Ce système contribue à accroître la compétitivité, à réduire les coûts et à améliorer la qualité du transport ferroviaire [26]. ERTMS est ainsi devenu un élément clé pour l'interopérabilité du réseau ferroviaire européen, à la fois pour les trains à grande vitesse et conventionnels.

Toutes ces phases propres à chaque système (CBTC, ERTMS, signalisation, etc.) permettent de collecter et stocker un ensemble de données (vitesse, planification, déplacements, images, etc.) relatives à la circulation des rames sur les voies. Ces informations facilitent la prise de décision des agents dans les cabines de contrôle, notamment en ce qui concerne la gestion du trafic et la prévention des collisions entre rames.

2.1.3 Timetable Management (TTM) dans les systèmes ferroviaires

La gestion des horaires, ou Timetable Management (TTM), constitue une composante essentielle du CTC. Elle repose sur un ensemble de données prescrites pour une période spécifique de trafic, nécessaires au fonctionnement optimal du système de contrôle. Ces données proviennent d'un calculateur externe de gestion des horaires,

lequel alimente la surveillance des trains, la comparaison des horaires, ainsi que la régulation et la supervision automatiques des trains.

L'objectif principal de la gestion des horaires est de fournir des données planifiées aux différents sous-systèmes afin d'assurer un service de trains conforme à ces données. Parmi ses fonctions clés, on peut citer :

— La comparaison des horaires.
— La régulation automatique des rames.
— La supervision automatique des rames.
— Le système d'information des passagers.

Le TTM, est une composante essentielle des systèmes de contrôle ferroviaire. Le timetable prévu (TTP) regroupe des informations détaillées sur les circulations de trains programmées pendant une période donnée, servant de base à la planification et à l'organisation des mouvements des rames. En temps réel, le module de ATS utilise un timetable réel (TTR), mis à jour en permanence pour refléter l'état actuel des circulations. Ce processus repose sur un modèle lien-nœud représentant chaque train par une collection de triplets :

— L'identifiant du nœud (station ou point de passage).
— L'heure d'arrivée prévue et réelle.
— L'heure de départ prévue et réelle.

La mise à jour continue du TTR est cruciale pour optimiser la régulation du trafic et minimiser les perturbations. Cette distinction entre TTP et TTR facilite une régulation proactive en détectant rapidement les écarts ou retards [27].

Des recherches récentes proposent des approches innovantes pour améliorer la gestion des horaires ferroviaires. Gupta et al. (2023) ont développé un modèle de programmation linéaire visant à concevoir des horaires énergétiquement optimaux, appliqué à la ligne 8 du métro de Shanghai, permettant des économies d'énergie de 20,93 % à 28,68 % [28]. De plus, Sander et al. (2023) ont proposé une méthode robuste pour concevoir des infrastructures ferroviaires répondant à la demande d'horaires stratégiques, tout en minimisant les coûts d'infrastructure [29]. Ces travaux soulignent l'importance des modèles d'optimisation pour améliorer l'efficacité et la durabilité des réseaux ferroviaires modernes.

2.1.4 Discussion

Les données circulant dans un PCC sont vastes et hétérogènes. Collectées en temps réel, elles permettent de réguler le trafic et de garantir la sécurité. Cependant, une fois leur rôle opérationnel terminé, ces données sont majoritairement stockées à des fins statistiques annuelles, sans véritable exploitation continue [27]. Parmi ces données, on retrouve celles liées au TTM, comprenant des informations essentielles telles que l'emplacement des rames, les heures d'arrivée et de départ, les vitesses, les flux de passagers, les données audio et vidéo, ainsi que les données énergétiques. Bien que ces informations soient précieuses, elles restent largement sous-utilisées après leur collecte initiale.

Pour répondre à ce défi, nous proposons d'exploiter ces données historiques, ou celles accessibles à court terme, en les intégrant dans un modèle d'apprentissage automatique. Ce modèle pourrait fournir un écran supplémentaire aux opérateurs, exploitant ces données pour offrir des informations prédictives sur l'état futur du PCC. Cependant, plusieurs questions se posent : quelles données spécifiques doivent être

sélectionnées ? Comment les exploiter efficacement ? Et quelles sorties ou prédictions cet écran pourrait-il offrir pour assister les opérateurs dans leur prise de décision ?

L'objectif principal est d'intégrer les données historiques d'un PCC dans ce nouvel écran pour aider les opérateurs à prendre des décisions éclairées, permettant ainsi une régulation proactive et optimisée du trafic. Une telle approche vise à améliorer la gestion quotidienne des opérateurs en leur offrant des outils pour anticiper les événements à court terme et minimiser les perturbations.

Dans la section suivante, nous présentons le concept d'un système d'aide à la prise de décision pour un PCC et son intégration dans le système de régulation, exploitant les données historiques pour soutenir les opérateurs dans leurs tâches quotidiennes.

2.2 Systèmes d'aide à la prise de décision dans un PCC

Un SAPD est conçu pour pallier les limites humaines dans la prise de décision, en tenant compte de contraintes psychologiques, contextuelles et temporelles. Ce concept a été théorisé à partir des années 1950-1960, notamment au Carnegie Institute of Technology, avec un intérêt initial pour le comportement organisationnel et le fonctionnement du cerveau humain. L'objectif était d'utiliser le support informatique pour assister les décideurs, qu'ils soient seuls ou en groupe[30].

Dans les années 1970, le SAPD a évolué en un domaine de recherche distinct, permettant aux utilisateurs d'accéder à des bases de données et à des modèles pour résoudre des problèmes non structurés. À partir des DSS (*Decision Support Systems*) classiques, des variantes comme les GDSS (*Group Decision Support Systems*) et ODSS (*Organizational Decision Support Systems*) ont vu le jour. L'expansion s'est poursuivie dans les années 1990 avec des innovations comme l'entreposage des données et le traitement analytique en ligne (*Online Analytical Processing*, OLAP) [30].

2.2.1 Présentation générale des systèmes d'aide à la prise de décision (SAPD)

Un SAPD recueille, analyse et synthétise les données pour produire des rapports d'information complets. Contrairement à une application d'exploitation ordinaire, qui se limite à collecter des données, un SAPD est conçu pour fournir des informations structurées et exploitables permettant de prendre des décisions stratégiques [31]. Ce type de système peut être entièrement informatisé, alimenté par des humains, ou combiner les deux. Les systèmes les plus avancés analysent les données et proposent des décisions directement à l'utilisateur. À tout le moins, ils aident les utilisateurs à prendre des décisions plus éclairées et plus rapidement [32].

Un SAPD se compose principalement de trois éléments fondamentaux. Le premier est une base de données ou une base de connaissances, qui stocke des informations provenant de diverses sources internes et externes. Cette base sert de fondation pour l'analyse et la prise de décision [30]. Le deuxième élément est un système logiciel ou modèle, composé de méthodes mathématiques et analytiques. Ces méthodes permettent de traiter les données et de simuler différents scénarios, facilitant ainsi l'évaluation de situations en fonction de paramètres spécifiques [33]. Enfin, une interface utilisateur joue un rôle clé en permettant aux utilisateurs d'interagir avec le système. Elle présente les informations sous des formes variées, telles que des graphiques, des

tableaux ou du texte, rendant le SAPD intuitif et accessible pour une prise de décision efficace. Ces trois composantes, interconnectées, assurent une assistance optimale dans des environnements complexes où les décisions doivent être prises rapidement et précisément.

Dans le contexte d'une organisation, un SAPD est fréquemment utilisé par la direction des opérations et les départements de planification pour transformer les données collectées en renseignements exploitables. Ces systèmes, souvent destinés aux cadres moyens et supérieurs, permettent d'optimiser les stratégies opérationnelles avec une efficacité accrue [34].

Dans le cadre d'un centre ferroviaire, une discussion avec des experts du domaine chez egis-rail a montré qu'un SAPD peut être exploité pour assister les agents en salle de contrôle à prendre rapidement des décisions pertinentes et précises. Ce système serait particulièrement utile pour :

- Le contrôle du trafic : supervision des rames et régulation en temps réel.
- Le contrôle des gares : gestion des flux de passagers et des installations.
- La maintenance : surveillance proactive et planification des réparations.
- Les caractéristiques d'exploitation : ajustement des vitesses, suivi de la signalisation, etc.
- L'information des voyageurs : communication rapide et précise en cas de perturbations.
- La sécurité publique : gestion des incidents et prévention des risques.
- Le contrôle des événements exceptionnels : réponse efficace aux situations imprévues (pannes, conditions météorologiques, etc.).

Grâce à l'intégration d'un SAPD dans un environnement ferroviaire, il est possible d'améliorer significativement la réactivité, la sécurité et l'efficacité globale des opérations. De plus, l'analyse des données historiques et en temps réel permettrait d'anticiper les perturbations et de minimiser leur impact, offrant ainsi une expérience optimisée pour les opérateurs et les usagers. L'intégration de ce système nécessite une infrastructure informatique robuste et une collaboration étroite entre les départements d'opérations, de maintenance et de sécurité. Des travaux futurs pourraient se concentrer sur l'amélioration des algorithmes d'analyse et sur l'intégration de l'intelligence artificielle pour automatiser davantage les prises de décision [6].

2.2.2 Intégration du SAPD dans un PCC

L'installation d'un SAPD dans un PCC ne pose pas de problème majeur. En effet, les PCC utilisent déjà des systèmes informatiques avancés pour collecter les données relatives à la circulation des rames et les transmettre au centre de contrôle, notamment via des technologies telles que le CBTC [35] et le ERTMS [36]. Ces infrastructures informatiques offrent une base solide pour intégrer un nouveau système comme le SAPD, tout en assurant sa compatibilité avec les systèmes existants. Par conséquent, la mise en œuvre d'un SAPD et son interconnexion avec l'architecture informatique actuelle des PCC peuvent être réalisées de manière fluide et efficace.

L'intégration d'un SAPD dans un PCC repose sur l'interconnexion harmonieuse avec les systèmes existants, comme illustré dans la figure 2.1. Cette architecture générale met en évidence les principales interactions entre les différents composants du système ferroviaire, notamment la base de données, les systèmes de communication, et la cabine de contrôle. L'objectif est de fournir une assistance décisionnelle efficace aux

opérateurs en salle de commande tout en s'appuyant sur les infrastructures en place.

L'un des éléments clés de cette intégration est l'ajout d'un écran dédié au SAPD au sein du PCC. Cet écran présente des informations cruciales comme les prédictions, les alertes, ou encore les analyses stratégiques basées sur les données collectées. Ces informations, générées par le SAPD, sont destinées à améliorer la réactivité et la pertinence des décisions prises par les opérateurs.

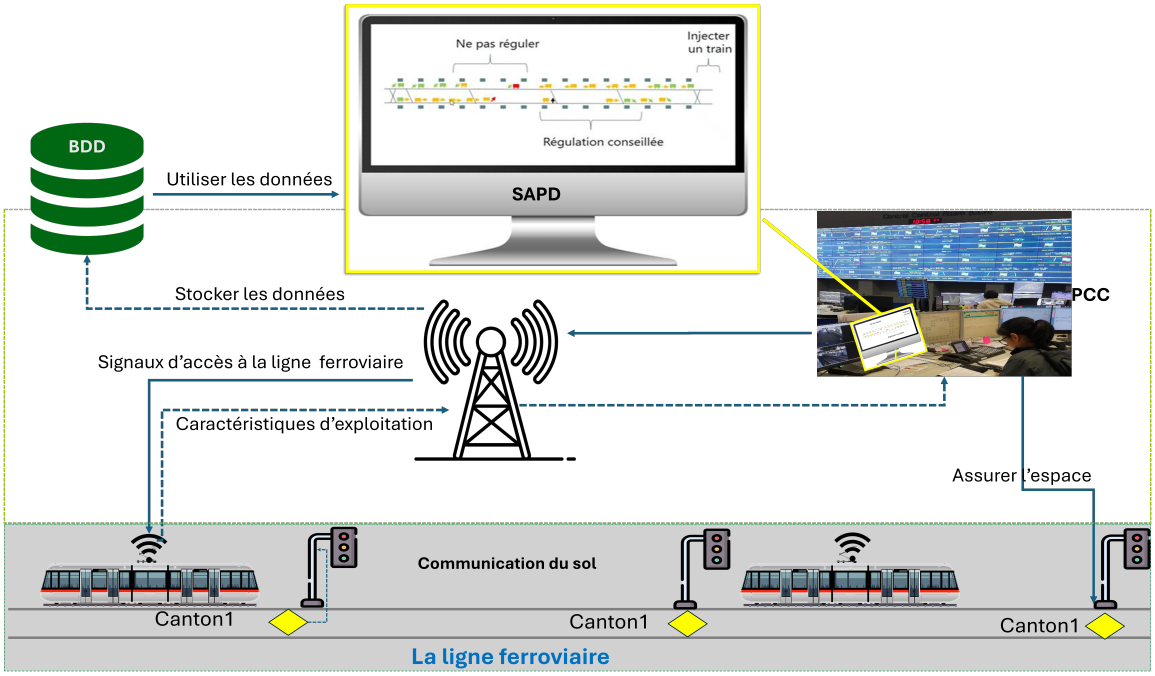


Figure 2.1 – Schéma général de l'intégration d'un SAPD dans un PCC.

Pour assurer la compatibilité avec les systèmes en place, le SAPD est directement connecté à la base de données existante du PCC. Cette base centralise les informations issues des systèmes de communication de bord et de sol, ainsi que du système de communication radio. Ces données incluent des signaux d'accès à la ligne de train, des caractéristiques d'exploitation, ou encore des indicateurs liés à la sécurité. Cette interconnexion garantit que le SAPD repose sur des données fiables et actualisées, tout en évitant les redondances [30, 33].

Un autre aspect essentiel de l'intégration est l'interface utilisateur. L'écran dédié au SAPD doit être conçu de manière intuitive et intégré harmonieusement dans l'interface globale du PCC. Cela permet aux opérateurs de naviguer facilement entre les différents outils disponibles, tout en maintenant un haut niveau d'efficacité opérationnelle. Grâce à cette approche, le SAPD devient un complément stratégique aux systèmes existants, offrant des analyses approfondies et des recommandations basées sur des données complexes.

En conclusion, l'intégration d'un SAPD dans un PCC repose sur trois piliers : la connexion à une base de données centralisée, l'ajout d'un écran dédié pour les opérateurs, et une interface utilisateur intuitive. Ces éléments interagissent pour transformer le PCC en un environnement de prise de décision amélioré, capable de relever les défis des opérations ferroviaires modernes.

2.2.3 Une vision améliorée de la régulation du trafic ferroviaire grâce au SAPD

Une vision globale a été décrite par les experts exploitation ferroviaire chez Egis-rail pour spécifier l'utilisation d'un SAPD dans un PCC, tout en respectant les procédures de régulation en vigueur. Cette vision met en avant une coordination entre les différents agents (opérateurs, agents de surveillance, agents d'information trafic) [37] et le SAPD, permettant une gestion fluide et efficace du trafic ferroviaire. La figure 2.2 illustre les étapes et interactions clés dans un PCC lors de l'utilisation du SAPD.

Lorsqu'un train (étape 1) s'approche d'une station, l'opérateur commence par vérifier l'itinéraire programmé (étape 2). Si la circulation du train se déroule comme prévu, l'opérateur de PCC poursuit simplement le suivi du train (étape 3). Cependant, en cas de problème détecté (étape 4), une procédure spécifique est déclenchée. Dans ce cas, l'opérateur demande l'autorisation de l'agent de surveillance local (étape 5), qui procède à une vérification de son secteur (étape 6). En parallèle, le SAPD analyse les données disponibles et confirme la nature du problème (étape 7). Une fois le problème validé, l'opérateur utilise les commandes des aiguillages et des signaux (étape 8) pour résoudre la situation. Cette action permet également de transmettre les informations nécessaires à l'agent d'information trafic (étape 9), qui est chargé d'informer les voyageurs des perturbations et des éventuels changements.

Les agents de PCC doivent prendre des décisions dans ces situations. Ces décisions incluent :

- Anticiper l'impact du retard à l'arrivée du train perturbateur.
- Contrôler les signaux pour garantir la sécurité et la fluidité du trafic.
- Exploiter les fonctionnalités du SAPD pour obtenir des statuts, des propriétés, et des tendances précises liées au trafic.
- Commander les installations, comme les aiguillages et les systèmes de signalisation.
- Résoudre les conflits liés au respect de l'intervalle ou à la gestion des itinéraires.

Ces décisions, prises en coordination entre l'opérateur, un manager de proximité, et le SAPD, permettent de minimiser les perturbations et de garantir une gestion optimale du trafic ferroviaire. Dans les sections suivantes, nous analyserons plus en détail les différentes décisions et leur impact, afin de mettre en lumière les bénéfices apportés par le SAPD dans un PCC.

2.2.4 Discussion

L'intégration d'un SAPD dans les PCC représente un défi pour les opérateurs. Bien que l'installation technique de ce système soit relativement facile grâce à l'infrastructure existante, notamment les systèmes CBTC et ERTMS, l'acceptation d'un tel outil par les opérateurs constitue un véritable enjeu.

Le principal obstacle réside dans la réticence des opérateurs à adopter un nouvel écran sous la forme d'un SAPD dans leur environnement de travail quotidien. En effet, ces opérateurs possèdent une expertise approfondie dans le domaine ferroviaire et sont habitués à travailler avec des outils qu'ils maîtrisent parfaitement. L'introduction d'un nouveau système peut être perçue comme une intrusion, voire une perturbation de leur routine, car ils peuvent craindre que cela ne complique leurs tâches ou n'entraîne une surcharge cognitive ou une déqualification professionnelle.

Pour surmonter ce défi, il est crucial de concevoir une interface utilisateur intuitive et ergonomique, qui s'intègre de manière fluide dans les outils déjà utilisés par les opérateurs. Cette interface doit être pensée pour s'insérer harmonieusement au sein de leurs écrans quotidiens, afin de minimiser l'impact sur leur routine de travail. L'ajout d'un nouvel écran, qui fonctionne parallèlement aux outils existants, pourrait être une première étape stratégique. L'idée est de permettre aux opérateurs d'utiliser le SAPD progressivement, en complément de leurs outils habituels, sans perturber leurs habitudes établies.

Il est cependant important de noter que, selon les retours des experts exploitation ferroviaire chez Egis-rail, une résistance initiale à l'utilisation du SAPD est à prévoir lors de son déploiement. Cette résistance s'explique par la méfiance naturelle des opérateurs envers tout nouveau système susceptible de modifier leurs méthodes de travail bien maîtrisées. Ils pourraient percevoir le SAPD comme une contrainte supplémentaire, notamment s'il nécessite un effort d'adaptation immédiat ou perturbe leur flux de travail quotidien.

L'objectif principal est de transformer cette perception négative en une acceptation progressive. Pour cela, il est recommandé d'implémenter le SAPD comme un outil complémentaire, placé à côté de leurs écrans habituels. Cette approche permet aux opérateurs de l'utiliser uniquement lorsque cela est nécessaire, en se familiarisant progressivement avec ses fonctionnalités. Avec le temps et un accompagnement approprié, le SAPD pourra être perçu non pas comme une contrainte, mais comme un véritable soutien à leur expertise, leur permettant de prendre des décisions plus rapides, plus précises, et mieux informées.

Derrière le SAPD que nous proposons, se trouvent les procédures de régulation décrites dans cette section, ainsi que l'utilisation des données d'exploitation présentées dans la section précédente. Ces données, qui ont été discutées en détail dans la discussion 2.1.4, jouent un rôle central dans le fonctionnement du SAPD. Elles nécessitent une préparation rigoureuse, notamment en termes de structuration et de nettoyage, avant d'être intégrées dans des modèles d'apprentissage adaptés. Ces modèles permettent de produire des prédictions fiables sur l'état futur des rames et de faciliter l'anticipation des retards. En résumé, le SAPD repose sur une synergie entre les procédures décrites ici et les données d'exploitation analysées auparavant, pour offrir un soutien décisionnel performant et aligné sur les exigences opérationnelles des PCC.

La section suivante présente les modèles d'apprentissage susceptibles d'être intégrés dans notre SAPD pour gérer les retards, en se concentrant sur des éléments essentiels à l'analyse des données d'exploitation, ainsi que sur les processus de prédiction à mettre en œuvre pour optimiser la gestion du trafic ferroviaire. Elle aborde les technologies et méthodologies existantes, les modèles d'apprentissage automatique adaptés à ce domaine, ainsi que leur application pratique dans le contexte des PCC, en les comparant à des travaux similaires portant sur l'IA dans les PCC. L'objectif est de fournir une vue d'ensemble des approches modernes utilisées pour améliorer la précision des prédictions et soutenir la prise de décision opérationnelle.

2.3 Intelligence artificiel en vue de l'intégration d'un SAPD dans PCC

Dans la plupart des travaux cités ci-dessous, la problématique de l'utilisation de modèles d'apprentissage dans les PCC pour résoudre les défis liés à la gestion ferroviaire repose sur l'exploitation d'un ensemble de données spécifiques adaptées à chaque problématique. Ces problématiques incluent, entre autres, l'optimisation des TTM, la gestion des retards, et la régulation des flux de voyageurs... Ces données nécessitent un prétraitement minutieux dans un schéma relationnel afin de décrire les problématiques et les données associées à chaque enjeu à gérer. Plusieurs approches sont employées pour résoudre ce type de problèmes, notamment des modèles d'apprentissage automatique, des techniques de deep-learning, des formules mathématiques, et des graphes.

Cependant, une question clé demeure : ces travaux existants sont-ils en mesure de fournir une assistance efficace à notre SAPD, en offrant aux opérateurs des prédictions basées sur des modèles d'intelligence artificielle (IA) dans un délai suffisamment court pour leur permettre de gérer efficacement le trafic ferroviaire ?

Cette section examine les travaux scientifiques les plus pertinents et explore les modèles d'apprentissage automatique les plus prometteurs, adaptés à notre approche.

2.3.1 Concepts et cadre général de l'intelligence artificielle

L'IA trouve ses origines dans les travaux pionniers d'Alan Turing dans les années 1940 et 1950. En 1950, Turing publie son article fondateur Computing Machinery and Intelligence" dans la revue Mind, où il propose le célèbre test de Turing" pour évaluer l'intelligence d'une machine [38]. Ce test vise à déterminer si un ordinateur peut imiter la pensée humaine au point d'être indiscernable lors d'une conversation.

Le terme "intelligence artificielle" est officiellement introduit en 1956 lors de la conférence de Dartmouth, considérée comme l'acte de naissance de l'IA en tant que domaine de recherche autonome [39]. Cette conférence, organisée par Marvin Minsky et John McCarthy, réunit vingt chercheurs éminents et marque le début de l'IA comme discipline scientifique. Malgré des avancées initiales prometteuses, l'IA connaît une période de stagnation dans les années 1970, souvent appelée "l'hiver de l'IA", due aux limitations technologiques et aux attentes trop élevées [40]. Cependant, les années 1980 voient un renouveau avec l'émergence des systèmes experts, capables de simuler le raisonnement humain dans des domaines spécifiques [41].

Au début du XXIe siècle, l'IA connaît une résurgence spectaculaire grâce aux progrès en puissance de calcul et à la disponibilité de grandes quantités de données. Cette nouvelle ère est dominée par les techniques d'apprentissage automatique et d'apprentissage profond, permettant des avancées significatives dans divers domaines tels que la reconnaissance visuelle, le traitement du langage naturel et la prise de décision automatisée [42]. Aujourd'hui, l'IA est omniprésente et transforme de nombreux secteurs, notamment la santé, les transports et la finance, tout en soulevant des questions éthiques importantes sur son impact sociétal et ses implications à long terme.

L'IA [43] est un domaine scientifique visant à créer des systèmes capables d'accomplir des tâches généralement associées à l'intelligence humaine, comme la reconnaissance visuelle, la compréhension du langage et la prise de décision[41]. Les avancées

récentes en IA, notamment dans l'apprentissage automatique et l'apprentissage profond, ont conduit à des progrès significatifs dans diverses applications, de la conduite autonome[44] à la médecine personnalisée[45]. Les algorithmes d'apprentissage supervisé et non supervisé sont au cœur de ces développements. Des modèles tels que les réseaux neuronaux convolutifs (CNN) et les réseaux récurrents (RNN) offrent des performances remarquables dans le traitement de données complexes[42]. L'optimisation des techniques, comme les approches par essaims particulaires ou les méthodes bayésiennes, a également contribué à améliorer l'efficacité des modèles[46]. Malgré ces avancées, l'IA soulève des questions éthiques importantes, notamment concernant la transparence des algorithmes et les biais potentiels dans les données utilisées[47]. Ainsi, tout en continuant à transformer de nombreux secteurs, l'IA nécessite une réflexion approfondie sur son impact sociétal et les conséquences à long terme de son déploiement.

La figure 2.3 illustre les principales catégories et techniques associées à l'apprentissage automatique et profond. Elle met en évidence les différentes approches utilisées pour résoudre des problèmes variés en intelligence artificielle, en les regroupant en quatre grandes familles : l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage par renforcement et l'apprentissage profond.

L'apprentissage automatique, ou Machine Learning (ML), désigne un domaine de l'intelligence artificielle qui vise à développer des algorithmes capables d'apprendre à partir de données et d'améliorer leurs performances avec le temps sans être explicitement programmés [48]. Les modèles de Machine Learning se basent sur des techniques statistiques et probabilistes pour établir des relations entre des données d'entrée et des cibles, permettant ainsi des tâches telles que la classification, la régression et le clustering.

L'apprentissage supervisé repose sur l'entraînement d'un modèle à partir de données étiquetées (X, y), où X représente les entrées et y les sorties associées. Les principales techniques incluent les régressions pour des sorties continues et les algorithmes comme les machines à vecteurs de support (SVM), les arbres de décision, les forêts aléatoires et les modèles de Gradient Boosting tels que XGBoost et LightGBM [41, 44]. Ces méthodes sont largement utilisées dans la classification, la régression et la prévision.

$$y = f(x)$$

où :

— y est la sortie ou la valeur cible que l'on cherche à prédire.
— f représente la fonction qui modélise la relation entre les entrées et les sorties.
— x désigne les caractéristiques ou variables d'entrée, souvent appelées features.

L'apprentissage non supervisé vise à identifier des structures ou des modèles cachés dans des données non étiquetées, sans disposer de cibles y. Parmi les principales techniques, on trouve le regroupement (*clustering*), comme l'algorithme *k-means*, et la réduction de dimensionnalité, avec des méthodes telles que l'analyse en composantes principales (ACP) ou les auto-encodeurs [42, 44]. Ces approches sont couramment utilisées dans des domaines tels que la segmentation de marché, l'exploration de données ou la détection d'anomalies.

L'apprentissage par renforcement se distingue par le fait que le modèle apprend par essais et erreurs en interagissant avec son environnement. Les techniques incluent le Q-Learning, les Deep Q-Networks et les approches basées sur la politique. Ces méthodes sont particulièrement efficaces dans les environnements où les décisions

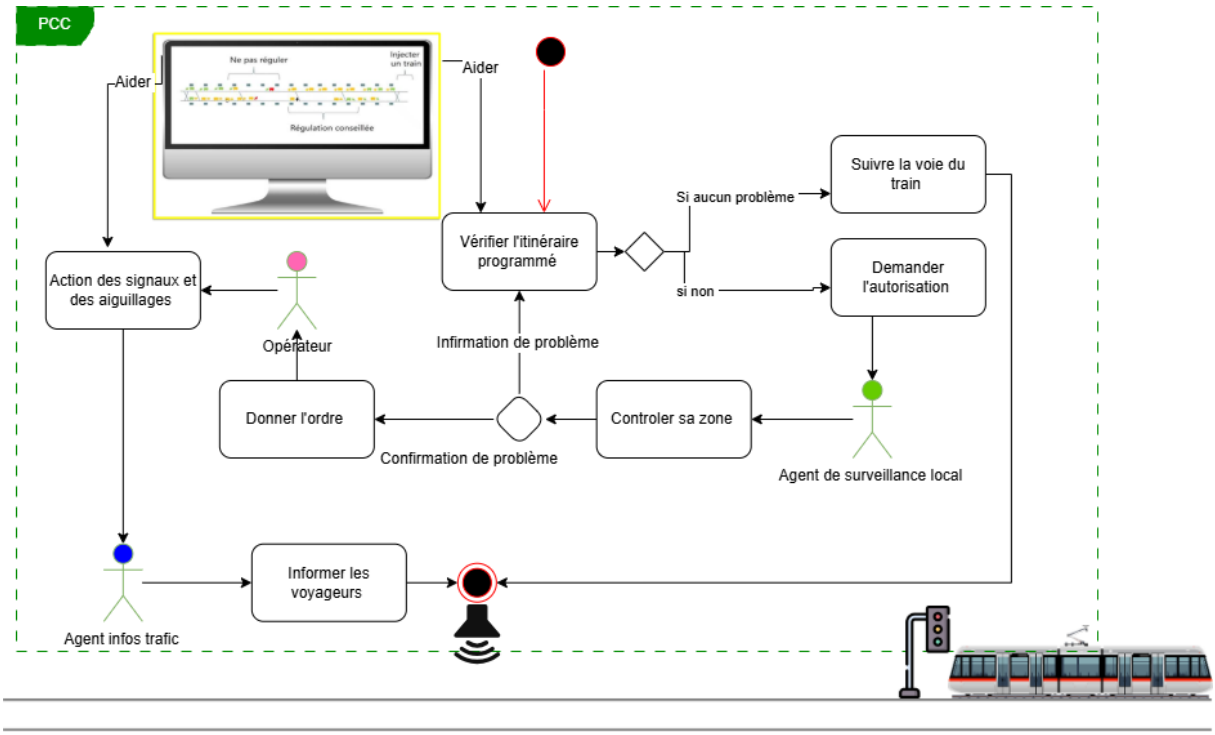


Figure 2.2 – Processus global d'utilisation du SAPD dans un PCC.

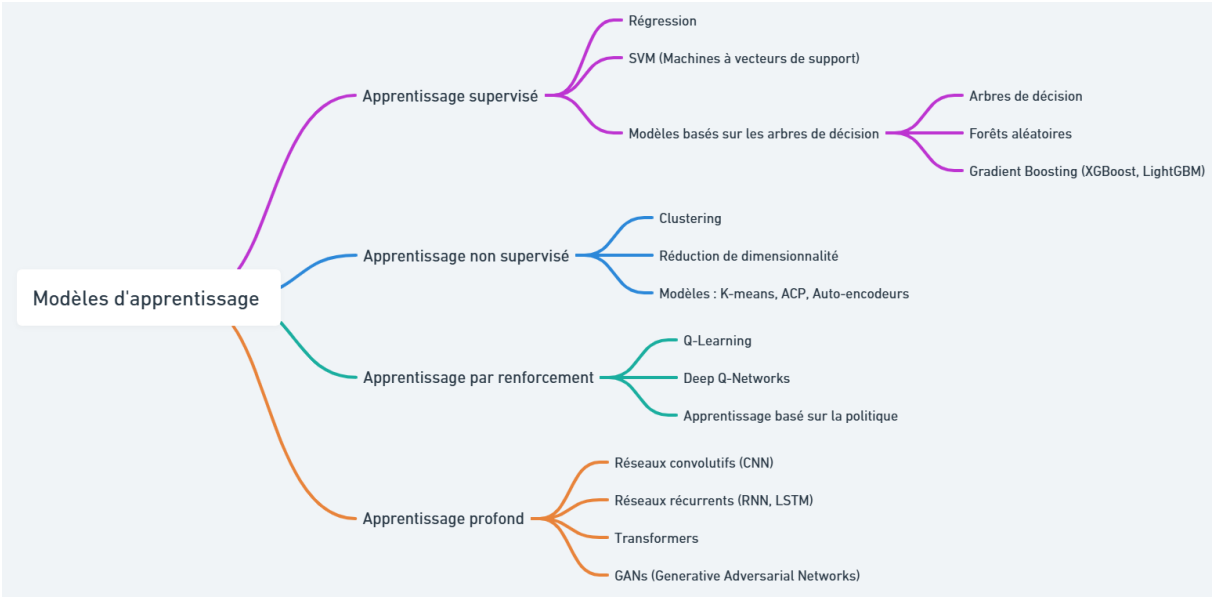


Figure 2.3 – Classification des techniques d'apprentissage automatique et profond, incluant les approches supervisées, non supervisées et par renforcement

successives influencent les résultats, comme les jeux ou la robotique [49, 41].

L'apprentissage profond, ou Deep Learning, est une sous-branche de l'apprentissage automatique qui repose sur l'utilisation de réseaux neuronaux profonds, composés de plusieurs couches hiérarchiques d'apprentissages [50]. L'apprentissage profond englobe des modèles complexes capables de capturer des relations non linéaires directement à partir des données brutes. Les réseaux convolutifs (CNN) sont adaptés à l'analyse d'images, tandis que les réseaux récurrents (RNN, LSTM) sont utilisés pour des données séquentielles. Les transformers, tels que BERT et GPT, dominent le traitement du langage naturel, tandis que les réseaux génératifs adversariaux (GANs) permettent de créer des données synthétiques réalistes [42, 44, 51].

Cette classification reflète l'évolution et la diversité des approches utilisées en apprentissage automatique, en répondant aux besoins spécifiques de divers cas d'application.

2.3.2 Travaux menés dans l'Intelligence Artificielle appliquée au secteur des transports publics

Le domaine de la prédiction dans les applications ferroviaires a connu des avancées significatives ces dernières années, avec une attention particulière portée à l'utilisation d'algorithmes d'apprentissage, en particulier des modèles d'apprentissage profond, pour des prévisions précises. Plusieurs études ont exploré l'application de ces techniques sophistiquées pour prédire dans des scénarios en temps réel. Cependant, une limitation commune observée dans les travaux existants est le caractère instantané des prédictions, ce qui pose des défis pratiques pour les opérateurs ferroviaires.

L'article de Camarazo et al. [52] propose une méthodologie centrée sur l'ingénierie dirigée par les modèles (*Model-Based System Engineering - MBSE*) pour aborder la complexité des systèmes ferroviaires. Cette approche met en avant des outils tels que UML [53], afin de garantir la conformité aux exigences fonctionnelles et de sécurité.

- Analyse des besoins : L'étude des besoins fonctionnels et non-fonctionnels, menée en collaboration avec divers acteurs (constructeurs de matériel roulant, gestionnaires d'infrastructures), sert de base à la conception [52].
- Conceptualisation : L'utilisation d'UML[54, 53] permet de traduire les besoins en modèles conceptuels, facilitant la transition vers une solution logique réutilisable.
- Modélisation de l'architecture : La méthode ARCADIA [55] et l'outil Capella sont employés pour élaborer des architectures logiques et physiques. Cette étape vise à répondre aux contraintes techniques et opérationnelles tout en assurant une modularité [52].
- Déploiement : Les prototypes sont validés et corrigés en s'appuyant sur les modèles produits lors des étapes précédentes.

Les auteurs soulignent également l'importance d'intégrer des données issues des infrastructures ferroviaires (voies, signalisation, télécommunications) et du matériel roulant (trains légers, trains de fret). Ces données sont enrichies grâce à des processus collaboratifs impliquant des ingénieurs de disciplines variées, permettant de répondre aux exigences critiques de sécurité [52].

La thèse de Zehrouni [56] propose une méthodologie centrée sur la conception et la modélisation des lignes ferroviaires en utilisant des modèles relationnels. Cette approche vise à optimiser l'utilisation des ressources critiques tout en répondant aux

contraintes opérationnelles et aux incertitudes du système ferroviaire. Dans un premier temps, la thèse décrit la modélisation des lignes ferroviaires, où chaque ligne est représentée comme un ensemble de relations entre les différentes ressources : infrastructure, engins roulants, et agents de conduite. Les gares, les voies, et les trains sont modélisés sous forme d'entités, tandis que les connexions et les horaires sont définis comme des relations entre ces entités. Ce cadre relationnel permet de capturer les interactions complexes entre les différents éléments du réseau ferroviaire.

Ensuite, la méthodologie inclut une conception des lignes ferroviaires basée sur des critères d'optimisation, tels que la minimisation des coûts, la maximisation de la capacité, et l'amélioration de la robustesse face aux retards. Les décisions de conception prennent en compte :

- Les contraintes de capacité des infrastructures, comme les limites des quais ou des voies.
- Les caractéristiques techniques des engins, incluant leur vitesse, leur capacité et leur compatibilité avec l'infrastructure.
- Les disponibilités et les qualifications des agents de conduite.

Un modèle relationnel [57, 58] est utilisé pour structurer ces informations et permettre leur exploitation par des outils d'optimisation. Ce modèle repose sur une formalisation mathématique où chaque relation (par exemple, train-gare, gare-voie) est traduite sous forme de contraintes linéaires dans un programme d'optimisation. En conclusion, la conception et la modélisation des lignes ferroviaires via des modèles relationnels, tels que décrits (on parle des modèles) dans cette thèse, offrent une approche systématique et efficace pour améliorer la gestion des réseaux ferroviaires tout en répondant aux défis opérationnels et stratégiques.

La thèse de Toqué [59] explore la prévision et la visualisation de l'affluence dans les réseaux de transports en commun en utilisant des méthodes d'apprentissage automatique. Ce travail de doctorat se concentre principalement sur trois cas d'étude : (i) le réseau de métro de Montréal au Canada, (ii) le réseau de métro et de bus de Rennes en France, et (iii) le réseau ferré et de tramway du quartier d'affaires de La Défense à Paris en France. Les données billettiques, issues des systèmes de cartes à puce comme Opus, KorriGo et Navigo, sont utilisées pour modéliser les flux de passagers et améliorer la planification des offres de transport. L'objectif principal est d'anticiper les flux de passagers à court et long terme afin de mieux adapter l'offre de transport à la demande, et de fournir des outils de visualisation pour soutenir les décisions opérationnelles.

La thèse de Rémi Coulaud [60] propose une approche de modélisation et de prévision de l'affluence ainsi que du temps d'échange à l'interface quai/train, avec une application spécifique à la ligne H du réseau Transilien. Le travail repose sur l'exploitation des données issues du Comptage Automatique de Voyageurs Embarqués (CAVE), collectées à l'aide de capteurs infrarouges intégrés aux trains. L'objectif principal est de développer des modèles prédictifs pour : 1. Le nombre de montées, de descentes et la charge à bord des trains. 2. La répartition des échanges voyageurs à quai. 3. Le temps d'échange entre les quais et les trains, afin d'optimiser les horaires et d'éviter les retards en cascade. En utilisant des techniques de classification non supervisée (clustering) et de prévision à court terme, la thèse identifie des typologies de gares et de jours, permettant une meilleure compréhension des flux passagers sur le réseau Transilien. Ces typologies sont ensuite utilisées pour ajuster les plans de transport et soutenir les décisions opérationnelles.

L'article de Markovic et al. [61] propose une méthode basée sur des modèles d'ap-

prentissage automatique pour analyser les retards d'arrivée des trains à la prochaine station, en se concentrant sur un horizon temporel de court terme, en mettant en œuvre la régression par vecteurs de support (SVR) et en la comparant aux réseaux de neurones artificiels (ANN). Cette étude constitue la première application de la SVR dans ce domaine spécifique et démontre sa supériorité en termes de généralisation et de précision sur les données de test. Les données utilisées proviennent des chemins de fer serbes (trains régionaux) et incluent divers facteurs tels que les conditions des infrastructures, la catégorie des trains, le pourcentage de trajet accompli, ainsi que les horaires prévus. Une méthode d'évaluation par experts a été mise en place pour quantifier l'influence des infrastructures sur les retards. L'étude présente également les corrélations entre les retards et les variables, mettant en évidence les facteurs pertinents pour la modélisation des retards dans les approches d'apprentissage automatique.

L'article de Kecman et Goverde [62] propose un modèle adaptatif basé sur les données pour la prédiction en temps réel des événements ferroviaires. Le modèle utilise un graphe d'événements avec des arcs dynamiques [63, 64], permettant de prendre en compte les conflits d'itinéraires et les ajustements dynamiques des temps de parcours en fonction des écarts entre les prévisions et les événements réels. En s'appuyant sur des données historiques, notamment les heures d'arrivée et de départ réelles et prévues, les temps d'occupation des voies, ainsi que les conflits d'itinéraires enregistrés, le modèle prédit avec précision les heures d'arrivée et de départ des trains. Les données historiques incluent également les écarts temporels observés entre les horaires planifiés et les performances réelles, permettant au modèle d'identifier des patterns répétitifs et d'améliorer la précision des prévisions. Ces informations, combinées à des ajustements dynamiques, permettent de capturer les variations dues à la congestion ou aux perturbations sur le réseau. Testé sur le corridor ferroviaire de Leiden à Dordrecht aux Pays-Bas, un des axes les plus fréquentés du pays, le modèle a démontré une capacité de prédiction d'un horizon de 30 minutes, avec une erreur moyenne absolue inférieure à une minute. Cette approche offre des perspectives prometteuses pour la gestion en temps réel des conflits d'itinéraires et l'amélioration de la ponctualité dans les réseaux ferroviaires régionaux.

L'article de Yaghini et al.[65] propose un modèle de prédiction basé sur les réseaux neuronaux artificiels (ANN) [66] pour estimer les retards des trains de passagers dans le réseau ferroviaire iranien. L'étude exploite des données historiques de retards collectées entre 2005 et 2009, après un prétraitement rigoureux pour éliminer les données incohérentes et bruitées. Trois méthodes différentes sont utilisées pour définir les entrées du modèle : codage en nombre réel normalisé, codage binaire, et codage par ensembles binaires. Pour optimiser l'architecture du réseau neuronal, trois stratégies (rapide, dynamique, et multiple) sont comparées. Les performances du modèle sont ensuite évaluées par rapport à des méthodes de prédiction classiques, telles que les arbres de décision et la régression logistique multinomiale, en utilisant des métriques comme le temps d'entraînement, la précision sur les données de test, et la taille du réseau.

L'article de Mou et al. [67] présente un modèle prédictif des retards ferroviaires basé sur un réseau neuronal à mémoire à long terme (LSTM) [68, 69]. En utilisant des données réelles du réseau ferroviaire néerlandais entre Rotterdam Central et Dordrecht, couvrant sept stations sur une période de 66 jours ouvrables (septembre-décembre 2017), le modèle vise à prédire les retards d'arrivée des trains. Les caractéristiques

influençant les retards comprennent le type de train, les retards observés aux stations précédentes, ainsi que les temps de trajet prévus. Comparé à des méthodes traditionnelles comme les forêts aléatoires et les réseaux neuronaux artificiels (ANN), le modèle LSTM a montré des performances supérieures avec une précision de 87,6% pour des erreurs de prédiction inférieures à 30 secondes. Bien que cette approche soit prometteuse pour la gestion opérationnelle et la ponctualité des trains, l'étude reconnaît des limites, notamment le manque d'intégration de données d'infrastructure supplémentaires et la nécessité d'adaptation à d'autres réseaux ferroviaires.

L'étude de Jiang et al. [70] propose un modèle de micro-simulation [71] basé sur les données massives de passagers pour évaluer et optimiser les horaires ferroviaires. En utilisant les données origine-destination collectées par les systèmes automatiques de perception des tarifs (AFC), le modèle simule les interactions entre passagers et trains pour identifier les problèmes liés à la surcharge et aux temps d'attente. Appliquée à une ligne ferroviaire à Shanghai, la méthode a permis de réduire significativement le nombre de passagers en attente et les temps d'attente cumulés en ajustant les horaires pendant les heures de pointe. Bien que prometteuse, cette approche n'intègre pas encore les retards opérationnels des trains, ce qui représente une piste pour notre recherche.

L'article de Berger et al. [72] présente un modèle stochastique [73] innovant pour la propagation des retards dans les réseaux des trains régionaux. Ce modèle utilise un graphe d'événements pour représenter l'horaire des trains et intègre des éléments clés tels que les politiques d'attente pour les correspondances, les profils de temps de trajet variables, et le potentiel de rattrapage des retards. Il est conçu pour s'adapter à un scénario en ligne, permettant des mises à jour rapides du statut des trains. Le modèle calcule les distributions des temps d'arrivée et de départ pour l'ensemble du réseau, démontrant une efficacité remarquable en propageant les retards sur tout le réseau ferroviaire allemand. Les expérimentations montrent que le modèle est suffisamment rapide pour des applications pratiques en temps réel, bien que la qualité des prévisions nécessite un ajustement des distributions de temps de trajet utilisées. Ce modèle offre des applications potentielles dans l'information des voyageurs en temps réel, la gestion des retards, et l'analyse de la stabilité des horaires planifiés, représentant une avancée significative dans la modélisation stochastique des réseaux de transport ferroviaire.

L'article de Dong et al. [74] propose une méthode de prédiction à court terme des paramètres de trafic [75], tels que les débits et les vitesses moyennes, dans un réseau d'autoroutes urbaines, en utilisant des modèles multivariés d'espace d'état[76]. Cette approche introduit des modèles spécifiques pour les conditions de trafic congestionné CSS et non congestionné NSS[77]. Les modèles exploitent une combinaison de données temporelles (données historiques proches) et spatiales (informations des segments voisins) pour capturer les dynamiques complexes du trafic routier. Ces variables sont intégrées dans les modèles NSS et CSS, conçus pour s'adapter aux conditions spécifiques du trafic, qu'il soit fluide ou congestionné. Cette étude souligne l'importance des modèles multivariés d'espace d'état pour prédire le trafic routier en temps réel, en combinant efficacement les caractéristiques temporelles et spatiales afin d'améliorer la précision des prévisions et d'aider à une gestion proactive des réseaux routiers urbains.

L'article de Gorman [78] présente une méthode d'estimation statistique innovante pour analyser les retards dus à la congestion dans le transport ferroviaire de marchandises aux États-Unis. En exploitant des données réelles collectées auprès de la compagnie BNSF Railway, l'étude met en lumière les principaux facteurs contribuant

aux retards et propose un modèle économétrique capable de prédire avec précision le temps total de parcours des trains. Parmi les facteurs clés identifiés figurent les croisements de trains, les dépassements, la variabilité des espacements entre trains, ainsi que l'influence des temps de circulation des trains précédents. Ces variables permettent de modéliser de manière détaillée les retards liés à la congestion. Le modèle économétrique développé démontre une performance remarquable, avec une erreur moyenne absolue de seulement 4,6% pour les prévisions mensuelles du temps moyen de parcours des trains, soulignant son potentiel pour améliorer la planification et la gestion des réseaux ferroviaires.

L'article de Corman et Meng [79] offre une analyse approfondie des modèles dynamiques et des algorithmes utilisés pour la gestion en temps réel du trafic ferroviaire. Il vise à classer et évaluer les approches actuelles dédiées à la reprogrammation des horaires et à la gestion des retards dans des contextes dynamiques et incertains. Ces approches incluent des solutions en boucle ouverte, où les décisions sont prises sans rétroaction, et des solutions en boucle fermée, qui ajustent les horaires en fonction des données en temps réel. Les techniques présentées sont également catégorisées en stratégies proactives, visant à anticiper les perturbations, et réactives, centrées sur la résolution des problèmes existants. Les prédictions sont effectuées sur un horizon temporel de 30 minutes à 1 heure, offrant une rapidité suffisante pour la gestion opérationnelle.

L'article de Flier et al. [80] propose une méthode algorithmique pour identifier les dépendances systémiques des retards dans les réseaux ferroviaires, causées par les conflits de ressources ou les correspondances maintenues entre trains. En exploitant les données des chemins de fer suisses (SBB) couvrant environ 2 300 points d'exploitation sur une période de deux mois en 2008, les auteurs analysent les retards primaires et secondaires pour détecter les relations récurrentes de dépendance. Deux types de dépendances sont identifiés : les dépendances d'attente, où les retards sont dus à des correspondances maintenues, et les dépendances de blocage, causées par l'utilisation partagée de ressources telles que les voies ou les plateformes. Les algorithmes développés permettent de détecter ces relations sans hypothèses préalables sur les distributions des retards et révèlent que des ajustements mineurs dans les horaires peuvent atténuer les retards secondaires. Bien que l'absence d'informations explicites sur les causes des retards limite la précision des dépendances identifiées, cette méthode offre une approche robuste pour améliorer la gestion des horaires ferroviaires grâce à l'analyse systématique des retards.

L'article de Rui Shi et al. [81] propose une méthode combinant l'algorithme XGBoost [82] et une optimisation bayésienne (Bayesian Optimization) [83] pour prédire les retards à l'arrivée des trains dans les réseaux ferroviaires à grande vitesse (HSR) en Chine. Cette méthode exploite des données opérationnelles et d'infrastructure pour améliorer la précision des prédictions, notamment dans des cas d'événements anormaux. Les résultats montrent une meilleure performance du modèle par rapport à d'autres approches, avec des erreurs réduites (RMSE et MAE) [84] sur deux lignes HSR chinoises.

L'article de Xu Bao et al. [46] propose une méthode hybride combinant Extreme Learning Machine (ELM) [85] et l'optimisation par essaims particulaires (Particle Swarm Optimization - PSO) [86] pour prédire les retards des trains dans les réseaux ferroviaires à grande vitesse (HSR). Cette approche se base sur neuf caractéristiques spatio-temporelles (comme le numéro de train, le code de station et les temps de parcours

réels et planifiés) pour améliorer la précision des prédictions. Les données utilisées proviennent d'une section de la ligne Wuhan-Guangzhou (W-G) HSR, couvrant 15 stations sur une distance de 1096 km, avec plus de 400 000 points de données collectés entre octobre 2018 et avril 2019. Les résultats montrent que le modèle ELM-PSO surpasse quatre modèles de référence (KNN [87], CatBoost [88], GBDT [89], et Lasso [90]) en termes de précision (MAE, RMSE et R-carré) [84] et de capacité à gérer les grandes quantités de données.

L'article de Wen et al. [91] propose un modèle basé sur le réseau LSTM (Long Short-Term Memory) pour prédire les retards d'arrivée des trains en utilisant les données opérationnelles réelles du réseau ferroviaire néerlandais. En s'appuyant sur les séries temporelles des retards et leurs facteurs d'influence, le modèle est comparé à d'autres approches, notamment les forêts aléatoires (Random Forest - RF) et les réseaux neuronaux artificiels (Artificial Neural Network - ANN). Les résultats montrent que le modèle LSTM offre des performances supérieures, avec une meilleure précision et une capacité accrue à capturer les relations complexes entre les retards et les facteurs spatio-temporels. Cette méthode contribue à améliorer la robustesse des horaires ferroviaires et à soutenir les décisions opérationnelles.

L'article de Marsh et Bearfield [92] propose un modèle causal [93] pour analyser et gérer les risques liés aux incidents ferroviaires de faible conséquence, tels que les glissades, les chutes et les passagers coincés dans les portes lors de l'embarquement ou du débarquement. Ce modèle, basé sur les réseaux bayésiens, vise à identifier les facteurs causaux influençant ces incidents et à estimer l'impact des interventions de gestion des risques. En s'appuyant sur des données opérationnelles collectées à l'échelle nationale via le système SMIS (Safety Management Information System), l'approche permet de profiler les risques spécifiques à chaque station et d'évaluer l'efficacité potentielle des mesures correctives. Le modèle intègre également des données contextuelles, telles que la saison, l'heure de la journée, et les caractéristiques des infrastructures, pour affiner les prédictions et soutenir les décisions locales en matière de sécurité.

L'article de Sobrie et al. [94] propose une comparaison entre un système basé sur des règles (Swiss Rail Control System - RCS) et un modèle de deep learning (LSTM encoder-decoder) [69] pour la prédiction des retards ferroviaires en Belgique. Le modèle LSTM est conçu pour capturer les interactions complexes à l'échelle spatio-temporelle, utilisant des données incluant des caractéristiques opérationnelles, des retards passés, et des informations sur les stations voisines. Les résultats montrent que le modèle LSTM surpasse le système RCS en réduisant l'erreur quadratique moyenne (RMSE) [84] de 18%, avec des améliorations marquées dans les contextes complexes tels que les zones densément peuplées ou les périodes de pointe. Implémenté comme preuve de concept dans les centres de contrôle d'Infrabel, ce modèle démontre la pertinence des approches basées sur le deep learning [50] pour la gestion en temps réel du trafic ferroviaire.

L'article de Yu et al. [95] propose un modèle innovant, nommé "Spatial–Temporal Bi-directional Long Short-Term Memory (ST-BiLSTM)", pour prédire les retards ferroviaires. Ce modèle intègre un bloc de convolution spatio-temporelle (ST-Conv) permettant de capturer les dépendances spatiales et temporelles à travers des graphes convolutifs et des convolutions causales. Il inclut également un bloc Bi-Att combinant un LSTM bidirectionnel et un mécanisme d'attention, ce qui permet de pondérer les informations critiques et d'améliorer la précision des prédictions. Les données utilisées dans l'étude proviennent de l'API Darwin, le moteur d'information officiel du réseau fer-

roviaire britannique. Elles contiennent des métriques de service, telles que les horaires de départ et d'arrivée, ainsi que des détails sur les trajets, notamment les identifiants des trains, les horaires planifiés et les horaires réels. L'analyse porte sur des trajets reliant "Didcot Parkway" et "London Paddington" sur les années 2016 et 2017. Les résultats expérimentaux montrent que le modèle ST-BiLSTM surpasse les approches classiques, telles que la régression linéaire, le perceptron multicouche (MLP) et les réseaux convolutionnels spatio-temporels (STGCN).

Notre article de CHTIOUI et al. [96] propose une approche innovante pour la prédiction des retards dans les systèmes ferroviaires urbains en temps réel, en utilisant l'algorithme XGBoost optimisé par des techniques bayésiennes. Ce travail, issu de notre thèse, s'inscrit dans le cadre des recherches menées pour améliorer la gestion des retards et optimiser la ponctualité des transports en commun. L'étude met en avant une méthodologie basée sur l'intégration de données multi-attributs, incluant des données temporelles, contextuelles et historiques. Après une phase de prétraitement minutieuse, le modèle atteint une précision de 95 % sur un horizon de prédiction à court terme, démontrant son efficacité pour fournir des prédictions robustes en environnement opérationnel.

2.3.3 Travaux menés dans les méthodes d'évaluation et de validation des modèles

Les méthodes d'évaluation et de validation des modèles présentées dans les travaux analysés reposent sur des métriques standardisées [84] telles que le RMSE (Root Mean Squared Error) 2.1 et le MAE (Mean Absolute Error) 2.2 pour quantifier l'écart entre les prédictions et les observations, avec pour objectif leur minimisation. En complément, le coefficient de détermination R^2 2.3 est utilisé pour mesurer la proportion de variance expliquée par les modèles présentés dans la figure 2.4, une valeur élevée confirmant leur pertinence et leur fiabilité.

$$\text{RMSE} = \sqrt{\frac{1}{p}\sum_{l=1}^{p}(y_l - \hat{y}_l)^2}, \tag{2.1}$$

$$\text{MAE} = \frac{1}{p}\sum_{l=1}^{p}|y_l - \hat{y}_l|, \tag{2.2}$$

$$R^2 = 1 - \frac{\sum_{l=1}^{p}(y_l - \hat{y}_l)^2}{\sum_{l=1}^{p}(y_l - \bar{y})^2}, \tag{2.3}$$

où :

— $\mathbf{p}$ est le nombre total d'observations.
— $\mathbf{y_l}$ est la valeur réelle de la variable cible pour l'observation l.
— $\hat{\mathbf{y}}_\mathbf{l}$ est la valeur prédite pour l'observation l.
— $\bar{\mathbf{y}}$ est la moyenne des valeurs réelles $\mathbf{y_l}$.

En complément des métriques courantes comme le RMSE, le MAE et le R^2, certains articles proposent des approches supplémentaires pour valider et évaluer les modèles. Par exemple, Markovic et al.[61] comparent les performances de la régression par vecteurs de support (SVR) et des réseaux de neurones artificiels (ANN), mettant en évidence la supériorité de la SVR en termes de précision et de généralisation. De leur côté, Berger et al. [72] analysent les distributions des erreurs pour mieux comprendre

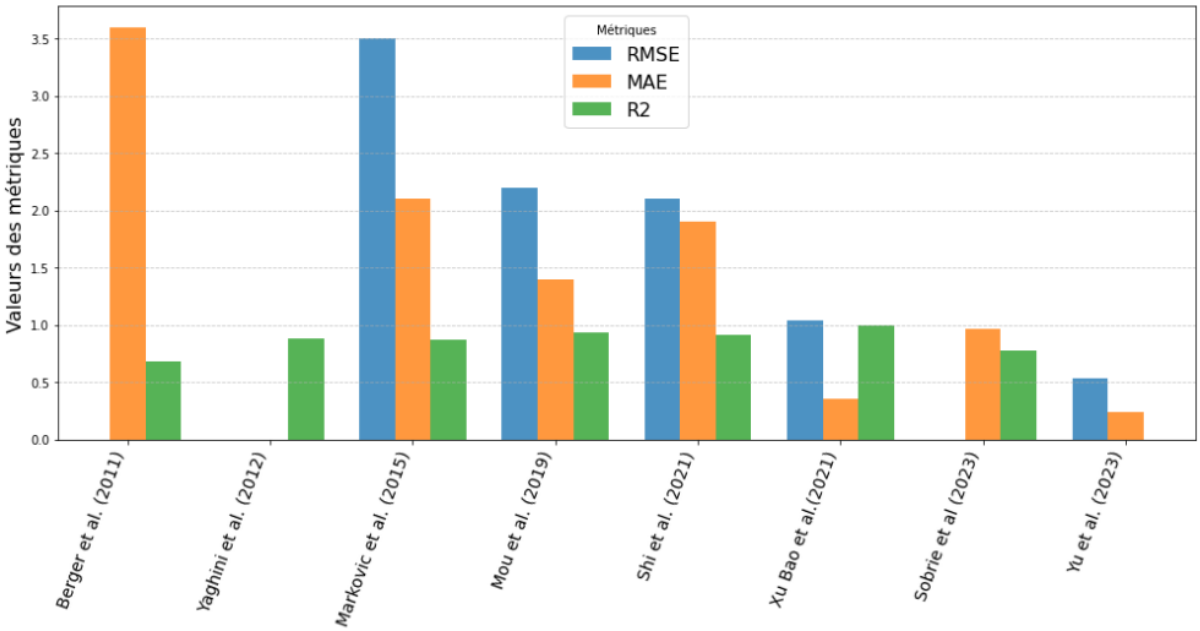


Figure 2.4 – Comparaison des performances des modèles pour différents articles en termes de métriques RMSE, MAE et R^2 ([72], [65], [61], [67], [46], [81], [94], [95]).

les écarts entre les prédictions et les observations, en particulier dans les scénarios de propagation des retards [97].

La validation croisée [98], combinée à des techniques d'optimisation est essentielle pour garantir la robustesse des modèles sur des ensembles de données indépendants. Elle est utilisée dans plusieurs travaux, notamment par Xu Bao et al.[46] pour leur modèle hybride ELM-PSO.

Certaines études adoptent une validation qualitative. Par exemple, Markovic et al. [61] utilisent une évaluation par des experts pour quantifier l'impact des infrastructures sur les retards, tandis que Kecman et Goverde [62] testent leurs modèles sur un corridor ferroviaire réel (Leiden-Dordrecht) pour valider leur applicabilité pratique. Enfin, des analyses comparatives temporelles, comme celles réalisées par Yaghini et al.[65], examinent le temps d'entraînement des modèles pour évaluer leur efficacité dans des environnements contraints par le temps. Ces méthodologies variées permettent d'assurer une évaluation rigoureuse des modèles, en tenant compte de leur précision, de leur robustesse et de leur applicabilité dans des scénarios réels.

Prédire les retards dans les systèmes de métro ou de tramway représente généralement un défi majeur en raison de la complexité de la collecte et de la disponibilité des données provenant de diverses sources, impliquant des formats multiples, y compris des données statiques et temporelles. La tâche devient complexe car elle nécessite de comprendre et d'intégrer des informations hétérogènes pour développer des modèles de prédiction robustes. La diversité des formats de données, couplée à la nature dynamique du trafic dans les systèmes de transport public, rend la modélisation des retards à la fois passionnante et complexe. Cette complexité nécessite des approches innovantes et des algorithmes avancés pour exploiter pleinement le potentiel des données disponibles et fournir des prédictions précises en temps réel.

Le tableau 2.1 présente une synthèse des travaux exposés précédemment. Il met en lumière les types d'infrastructures utilisés ainsi que les méthodologies appliquées, afin d'obtenir une vision globale de ce qui pourrait être adapté à notre thèse pour

la modélisation relationnelle des réseaux publics et la prédiction des retards dans les transports publics. Le premier défi réside dans le fait que la plupart des travaux étudiés dans cette thèse se concentrent principalement sur les réseaux ferroviaires régionaux ou à grande vitesse. En revanche, notre objectif est d'intégrer l'intelligence artificielle (IA) dans les postes de contrôle ferroviaire des réseaux de transport public, notamment les métros et les tramways. Enfin, le tableau présente les contributions spécifiques de chaque travail (articles) et comment elles ont été mobilisées dans la construction de notre approche.

La modélisation des données ferroviaires sous forme de modèles relationnels est cruciale pour gérer les liens entre les différents ensembles de données. Cette approche facilite l'identification des variables et des facteurs influençant les performances des lignes ferroviaires. Les travaux de Camarazo et al. [52] et de Zehrouni [56] apportent des perspectives essentielles à ce sujet. D'une part, Camarazo et al. [52] mettent en avant une méthode de modélisation des systèmes ferroviaires. Ils expliquent comment les interactions complexes entre les infrastructures, les agents humains et les composants matériels peuvent être représentées à l'aide de modèles conceptuels comme UML [54],[53]. Ces outils permettent de visualiser les relations entre les éléments critiques tels que les voies, la signalisation, et les agents de conduite, tout en garantissant la conformité avec les exigences du système. Ces travaux soulignent également l'importance de l'ingénierie dirigée par les modèles pour améliorer la cohérence et la compréhension des systèmes ferroviaires.

D'autre part, Zehrouni [56] se concentre sur la planification robuste et intégrée des ressources ferroviaires, en identifiant les entités clés du système ferroviaire et leur rôle dans la modélisation. Son approche propose un modèle d'optimisation qui intègre les principales entités du système, telles que les infrastructures (voies, gares, et équipements de signalisation), les engins ferroviaires (trains, wagons) et les agents humains (conducteurs, contrôleurs, et gestionnaires). Ces éléments essentiels forment la base d'un modèle relationnel permettant de représenter les interactions complexes entre les différents acteurs et composants.

Ces deux contributions s'alignent sur notre approche de modélisation relationnelle en fournissant des cadres conceptuels pour intégrer les données d'exploitation, les flux voyageurs et les caractéristiques des infrastructures. Elles permettent également de poser les bases pour une analyse systémique des dépendances entre les variables, un aspect essentiel pour la prédiction des retards dans les transports publics.

Table 2.1 – Synthèse des travaux sur l'intelligence artificielle appliquée aux transports publics

Travaux	**Problématique**	**Infrastructure ferroviaire**	**Utilisation dans notre approche**
Camarazo et al. [52]	Modélisation des systèmes ferroviaires	Système ferroviaire en général	-Aider à construire le modèle relationnel
Zehrouni [56]	La conception et la modélisation des lignes ferroviaires	Les réseaux régionaux	-Aider à construire le modèle relationnel
Jiang et al. [70]	Simulation des interactions entre les passagers et les trains	Le réseau public	-Intégration prévue des flux de voyageurs dans le système (perspective)
Berger et al. [72]	Propagation des retards dans les réseaux ferroviaires	Les réseaux régionaux	-Analyse de retard -Utilisation des données d'exploitation
Markovic et al. [61]	Prédiction des retards à l'arrivée des trains	Les réseaux régionaux	-Études comparatives -Intégration des facteurs liés aux infrastructures ferroviaires
Dong et al. [74]	Prédiction à court terme des paramètres de trafic	Les réseaux urbains	-Utilisation des données historiques proches
Gorman [78]	Analyse des retards	Les réseaux publics	-Analyse des facteurs contribuant aux retards
Corman et Meng [79]	La reprogrammation des horaires et la gestion des retards	Trafic routier urbain	-Horizon de prédiction à court terme
Yaghini et al.[65]	Estimer les retards des trains de passagers	Les réseaux régionaux	-Prétraitement des données -Analyse comparative -Utilisation des retards historiques

Travaux	Problématique	Infrastructure	Utilisation dans notre approche
Mou et al. [67]	Réduire les retards à l'arrivée des trains	Les réseaux régionaux	-Analyse comparative -Utilisation des retards historiques aux stations précédentes
Flier et al. [80]	Identifier les dépendances systémiques des retards	Les réseaux régionaux	-Analyse de la dépendance du retard avec l'infrastructure
Kecman et Goverde [62]	Prédiction des heures d'arrivée et de départ des trains	Les réseaux régionaux	-Utilisation des données infrastructurelles -Horizon de prédiction court terme
Rui Shi et al. [81]	Prédire les retards à l'arrivée des trains	Les réseaux à grande vitesse	- Analyse comparative (XGBoost) -Optimisation bayésienne des hyperparamètres. -Données d'exploitations
Xu Bao et al. [46]	Prédire les retards à l'arrivée des trains	Les réseaux à grande vitesse	-Données d'exploitations
Wen et al. [91]	Prédire le retard d'arrivée des trains	Les réseaux régionaux	-Données en séries temporelles -Caractéristiques des trains, -Analyse comparative (ANN,arbres)
Marsh et Bearfield [92]	Analyse et la gestion des risques liés aux incidents ferroviaires	Les réseaux régionaux	-Données contextuelles : saison, heure et caractéristiques des infrastructures
Rémi Coulaud [60], Toqué [59]	Prévision de l'affluence voyageurs	Les réseaux publics	-Aider à construire le modèle relationnel -Utilisation des données de flux voyageurs (perspective)
Sobrie et al. [94]	Prédiction des retards	Les réseaux régionaux	-Analyse comparative(deep learning) -Exploitation des données des stations voisines, -Analyse des heures de pointe et de creuses.

Travaux	Problématique	Infrastructure	Utilisation dans notre approche
Yu et al. [95]	Prédiction des retards	Les réseaux régionaux	-Analyse comparative(deep learning)
Chtioui et al. [96]	Prédiction des retards	Les réseaux publics	-Données d'exploitation -Données contextuelles -Fenêtre glissante sur le passé -Horizon de prédiction court terme

Le premier défi consiste à identifier les données et les facteurs pertinents pouvant être utilisés pour entraîner nos modèles d'apprentissage. Plusieurs articles, tels que [72] et [61], fournissent des idées claires sur les ensembles de données nécessaires pour les premiers tests d'entraînement. Ces travaux mettent en avant l'importance des données d'exploitation, notamment les TTPs et les TTRs. Ces données constituent une base essentielle pour comprendre les écarts entre les performances prévues et celles observées. En plus des données d'exploitation, il est également crucial d'intégrer des facteurs clés qui influencent les retards ferroviaires. Ces facteurs incluent les conditions des infrastructures, telles que l'état des voies et les capacités, ainsi que les caractéristiques spécifiques des lignes ferroviaires. Parmi ces caractéristiques, on peut citer les types d'infrastructures (régionales, grandes vitesses, urbaines), les vitesses des trains, les stations, et les distances entre les stations. Ces éléments permettent de modéliser avec précision les relations complexes entre les différents paramètres opérationnels.

Par ailleurs, plusieurs articles, comme [74], [79] et [65], insistent sur l'importance des données historiques. Ces travaux soulignent que l'analyse des performances des rames dans les stations précédentes, ainsi que les interactions entre une rame donnée et les rames prédécesseurs ou successeurs, joue un rôle clé pour comprendre et prédire les retards. Ces données historiques aident également à identifier les dépendances systémiques des retards, offrant ainsi une perspective précieuse pour optimiser la gestion des flux ferroviaires. En résumé, les contributions issues de ces travaux nous aident à structurer nos ensembles de données et à établir des relations pertinentes entre les facteurs influençant les retards ferroviaires. Ces approches permettent de poser les bases d'une modélisation robuste et d'entraîner nos modèles avec des données pertinentes et complètes.

L'intégration des données contextuelles dans les systèmes de modélisation est essentielle pour améliorer la compréhension des facteurs influençant les retards ferroviaires. Marsh et Bearfield [92] présentent une approche structurée de modélisation causale des incidents ferroviaires, intégrant des données telles que la saison, l'heure de la journée, et les caractéristiques des infrastructures. Ces données permettent non seulement d'analyser l'indépendance des retards vis-à-vis de la météo, mais également d'évaluer l'impact des jours de la semaine, des jours fériés et des périodes de l'année sur la performance des trains. L'accent est également mis sur l'utilisation de modèles bayésiens pour mieux comprendre les facteurs contributifs aux retards, facilitant ainsi la gestion proactive des risques.

Par ailleurs, Sobrie et al. [94] soulignent l'importance de l'analyse des heures de pointe et des heures creuses pour évaluer l'impact des variations spatiotemporelles sur les retards. Leur étude sur le réseau ferroviaire belge montre que les périodes de pointe augmentent considérablement la complexité du réseau et la probabilité de propagation des retards. En intégrant ces informations dans un modèle relationnel, il devient possible de capturer les interactions dynamiques entre les stations, les trains, et les facteurs temporels. Ces deux approches se complètent dans le cadre de notre recherche, où l'objectif est d'intégrer des données contextuelles et spatiotemporelles dans un modèle relationnel pour les systèmes de transport public. Cela permet d'identifier les dépendances critiques entre les variables et de proposer des solutions adaptées pour réduire les retards et optimiser les opérations.

L'horizon temporel est une notion critique dans les systèmes de transport ferroviaire, particulièrement lorsqu'il s'agit de prédire les retards dans des infrastructures

distinctes. Les auteurs Kecman et Goverde [62] et Rui Shi et al. [81] mettent en avant un horizon temporel à court terme, souvent limité à 30 minutes. Ce dernier est spécifiquement appliqué dans des réseaux ferroviaires régionaux et couvre généralement la station suivante. Dans ce contexte, les informations des trains à la station précédente sont utilisées pour anticiper les retards à la station suivante, une démarche logique dans des systèmes où les flux de passagers et la configuration des infrastructures favorisent une forte corrélation entre ces deux entités.

Cependant, cette approche présente des limites significatives dans le cadre de notre recherche. Notre étude se concentre sur les réseaux de transport public, tels que les métros et tramways, où l'horizon de 30 minutes peut inclure jusqu'à 10 stations ou plus. Par ailleurs, dans ce type de réseau, les intervalles de passage sont beaucoup plus courts, autour de 2 minutes 30 secondes en moyenne, ce qui complexifie l'application directe des modèles développés pour des réseaux régionaux ou à grande vitesse. Ces contraintes rendent nécessaire une réévaluation de l'horizon temporel afin d'adapter les modèles aux spécificités du transport public.

2.3.4 Contributions existantes relatives aux modèles d'apprentissage

La figure 2.5 illustre la contribution des modèles identifiés à partir de notre étude bibliographique. Ces modèles ont été sélectionnés en fonction des éléments essentiels extraits de chaque article analysé, afin d'être testés dans le cadre de notre approche. Ils sont classés en trois grandes catégories principales : les méthodes statistiques [99], les méthodes à base d'arbres de décision [100], et les méthodes d'apprentissage profond [50].

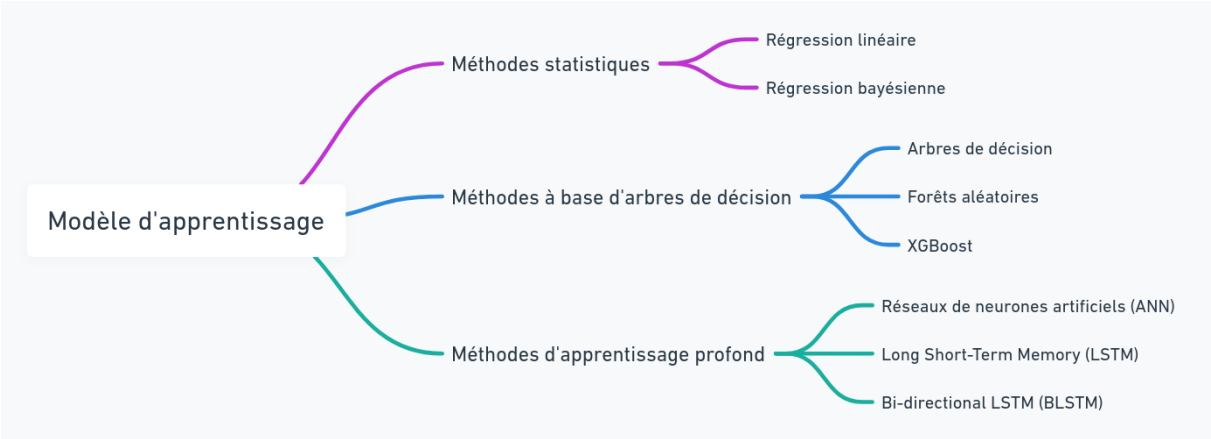


Figure 2.5 – Classification des méthodes d'apprentissage issues des contributions des travaux étudiés.

L'objectif principal de cette classification et des tests associés est d'identifier l'approche qui offre les meilleures performances pour répondre à notre problématique. En intégrant ces modèles dans notre cadre expérimental, nous visons à sélectionner la méthode la plus adaptée à la prédiction des retards dans les réseaux de transport public.

2.3.4.1 Méthodes statistiques

Le choix de commencer par les premiers tests avec des modèles statistiques a été fait dans le but d'évaluer la qualité des variables et d'analyser leur importance. Ces méthodes offrent l'avantage d'un temps de calcul rapide et d'une interprétabilité accrue des résultats, contrairement aux méthodes d'apprentissage profond. Par exemple, les modèles de régression permettent d'identifier l'importance relative des variables d'entrée en étudiant les paramètres associés. Nous présentons ici deux méthodes statistiques couramment utilisées dans les études d'analyse de la mobilité : la régression linéaire et le modèle bayesien naïf.

Régression linéaire [101] est une méthode statistique utilisée pour modéliser la relation entre une variable dépendante (cible) et une ou plusieurs variables indépendantes (prédicteurs). Elle cherche à ajuster une ligne droite (ou un hyperplan pour des données multidimensionnelles) qui minimise la différence entre les valeurs observées et les valeurs prédites. Cette méthode est largement utilisée pour des tâches de prédiction et d'analyse de tendances, en raison de sa simplicité et de son interprétabilité.

La régression bayésienne [102] repose sur une approche probabiliste pour estimer les paramètres d'un modèle de régression. Plutôt que de calculer un ensemble unique de coefficients comme dans la régression linéaire ordinaire, elle considère chaque coefficient comme une variable aléatoire dotée d'une distribution de probabilité. Cette méthode permet d'incorporer directement l'incertitude des paramètres dans le modèle, offrant ainsi des prédictions avec des intervalles de confiance. En appliquant une régularisation intégrée, la régression bayésienne permet de contrôler la complexité du modèle et d'éviter le surapprentissage en ajustant automatiquement les hyperparamètres pendant l'entraînement. Ces propriétés en font un outil robuste pour des données présentant des variabilités ou lorsque les données disponibles sont limitées, tout en facilitant l'interprétation des résultats à travers une estimation probabiliste des coefficients

2.3.4.2 Méthodes à base d'arbres de décision et d'optimisation bayésienne

Les méthodes à base d'arbres de décision [100] sont fréquemment utilisées pour traiter à la fois des données linéaires et non linéaires, ce qui est pertinent dans notre cas, où les relations entre les variables sont souvent complexes et difficiles à modéliser avec des approches plus simples. Ces modèles sont appréciés non seulement pour leur capacité à gérer des données hétérogènes, mais aussi pour leur explicabilité. En effet, ils permettent d'identifier l'importance de chaque variable d'entrée grâce à des indicateurs comme la feature importance. Nous présentons ici trois modèles basés sur des arbres de décision : l'arbre de décision [103], la forêt aléatoire [104], et XGBoost [82]. Ces méthodes varient en termes de complexité et de performance, mais partagent un point commun : Elles sont construites à partir d'arbres de décision.

Un arbre de décision [103] est un modèle prédictif qui segmente les données en sous-groupes en fonction de certaines variables. Il fonctionne en divisant les données à chaque nœud en fonction d'une condition spécifique (par exemple, si une variable est supérieure à une certaine valeur). À chaque nœud de l'arbre, la donnée est répartie selon ces règles de division, jusqu'à ce qu'elle atteigne une feuille terminale, qui représente une prédiction finale (pour la classification ou la régression). Cette approche

permet une grande flexibilité pour capturer les relations complexes entre les variables, tout en restant facilement interprétable, car les décisions sont représentées sous forme de chemin dans l'arbre, comme illustré dans la figure2.6.

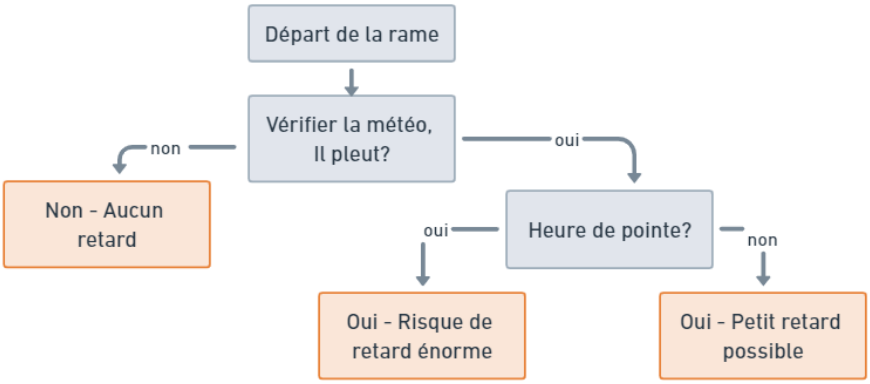


Figure 2.6 – Schéma d'un arbre de décision dont l'objectif est de prévoir le retard

L'optimisation bayésienne (BO) [83] est une méthode efficace pour optimiser les hyperparamètres d'un modèle, en particulier pour les modèles basés sur des arbres de décision. Comme nous le savons, pour ces types de modèles, il est nécessaire de définir plusieurs hyperparamètres avant de procéder à l'entraînement du modèle. BO équilibre l'exploration et l'exploitation de l'espace des hyperparamètres afin de trouver la meilleure combinaison qui minimise l'erreur quadratique moyenne (RMSE).

Les paramètres, notés θ, définissent les plages potentielles pour chaque hyperparamètre des modèles basés sur des arbres de décision. L'algorithme 1 présente le principe de cette optimisation. Initialement, θ_{best} est défini comme une fonction de moyenne nulle, puis il est mis à jour de manière itérative en optimisant l'acquisition $\alpha(\theta)$ pour déterminer quels hyperparamètres doivent être évalués. Le processus d'optimisation équilibre exploration et exploitation en calculant l'amélioration espérée, puis en évaluant la fonction objectif avec la négation de cette amélioration. Les hyperparamètres optimaux sont mis à jour en fonction des résultats de l'optimisation. L'algorithme 1 s'adapte pour explorer et exploiter l'espace des hyperparamètres θ_{best}, avec pour objectif de minimiser la valeur négative de la RMSE pour chaque modèle. Il est simplement nécessaire de définir les hyperparamètres spécifiques à chaque modèle basé sur des arbres de décision. Il est à noter que, dans l'ensemble des modèles basés sur les arbres de décision, certains hyperparamètres se retrouvent dans certains modèles et sont absents dans d'autres, comme présenté dans la tableau 2.2.

$$B(\theta) = -\text{RMSE} \tag{2.4}$$

où $\theta = \{\texttt{max_depth}, \texttt{min_samples_split}, \texttt{min_samples_leaf}, \texttt{max_features}, \texttt{n_estimators}, subsample, learning_rate, colsample_bytree\}$, sont les hyperparamètres à optimiser.

— **Profondeur maximale (max_depth)** : Définit la profondeur maximale des arbres. Plus l'arbre est profond, plus le modèle peut s'adapter aux données, mais cela augmente le risque de surapprentissage.

- **Nombre minimal d'échantillons pour diviser un nœud (min_samples_split)** : Nombre minimal d'échantillons requis pour diviser un nœud.
- **Nombre minimal d'échantillons par feuille (min_samples_leaf)** : Nombre minimal d'échantillons dans une feuille terminale.
- **Nombre maximal de caractéristiques (max_features)** : Nombre maximal de caractéristiques à considérer pour chaque division de nœud.
- **Nombre d'estimateurs (n_estimators)** : Nombre total d'arbres dans une forêt (Random Forest) ou de modèles dans un ensemble (comme XGBoost).
- **Sous-échantillonnage (subsample)** : Fraction des échantillons d'entraînement utilisés pour chaque arbre dans XGBoost.
- **Taux d'apprentissage (learning_rate)** : Contrôle la vitesse d'apprentissage dans XGBoost. Un taux d'apprentissage plus faible permet d'augmenter la précision, mais nécessite plus d'itérations.
- **Fraction des colonnes par arbre (colsample_bytree)** : Fraction des colonnes (caractéristiques) utilisées pour former chaque arbre dans XGBoost, permettant d'introduire plus de diversité entre les arbres.

Algorithme 1 : Algorithme d'optimisation bayésienne pour le réglage des hyperparamètres

Data : θ^{espace} (espace des hyperparamètres), X (ensemble de données d'entraînement)
Result : Meilleure combinaison d'hyperparam'etres θ^{meilleur}

$\theta^{\text{espace}} \leftarrow (\text{défini en fonction du type de modèle})$;
$B(\theta) \leftarrow -\text{RMSE}$;
$\theta^{\text{meilleur}} \leftarrow 0$;
while $\theta^{meilleur}$ *n'a pas convergé* **do**
  $\alpha(\theta) \leftarrow \mathbb{E}[\max(B(\theta^{\text{meilleur}}) - B(\theta), 0)]$; `// Optimiser la fonction d'acquisition`
  $ov \leftarrow -B(\alpha(\theta))$; `// Échantillonner la fonction objectif`
  $\theta^{\text{meilleur}} \leftarrow$ `mise_à_jour`(ov); `// Mettre à jour les paramètres en fonction des résultats de l'optimisation`
  $X \leftarrow$ no; `// Mettre à jour les données avec les nouvelles observations` no
return θ^{meilleur};

Arbres de décision (Decision Tree Regression) [103] sont des modèles prédictifs qui segmentent les données en fonction de caractéristiques spécifiques. À chaque nœud de l'arbre, une variable est sélectionnée pour diviser les données de manière à minimiser l'erreur de prédiction dans les sous-groupes formés. Le modèle continue à diviser les données jusqu'à atteindre un seuil d'homogénéité ou un nombre minimal d'échantillons dans chaque feuille. L'avantage principal des arbres de décision est leur interprétabilité. Chaque division correspond à une décision claire qui peut être suivie pour comprendre comment le modèle arrive à sa prédiction finale. Cependant, un arbre de décision unique peut être sujet au surapprentissage (overfitting), car il peut trop s'adapter aux particularités du jeu de données d'entraînement, ce qui réduit sa capacité de généralisation sur de nouvelles données. L'application de notre algorithme

Table 2.2 – Exploration des hyperparamètres par modèle de machine learning

Hyperparameters	Decision Tree	Random Forest	XGBoost
max_depth	✓	✓	✓
min_samples_split	✓	✓	
min_samples_leaf	✓	✓	
max_features	✓	✓	
n_estimators		✓	✓
subsample			✓
learning_rate			✓
colsample_bytree			✓

1 est essentielle pour optimiser les hyperparamètres de cet arbre de décision. Nous sélectionnons uniquement les hyperparamètres dédiés à ce type de modèle, comme indiqué dans le tableau 2.2.

Forêts aléatoires (Random Forest Regression) [104] sont une extension des arbres de décision qui visent à améliorer la performance et à réduire le surapprentissage. Une forêt aléatoire est composée de plusieurs arbres de décision indépendants, créés à partir de différents sous-ensembles des données d'entraînement (échantillonnage avec remise) et en sélectionnant de manière aléatoire un sous-ensemble de variables à chaque division. La prédiction finale est obtenue en faisant la moyenne des prédictions des différents arbres. Cette approche réduit la variance et rend le modèle plus robuste et général. Les forêts aléatoires sont efficaces pour capturer des relations complexes entre les variables tout en maintenant une bonne capacité de généralisation. Elles offrent également la possibilité de calculer l'importance des variables (feature importance), ce qui permet d'identifier les caractéristiques les plus influentes dans le modèle. L'application de notre algorithme 1 est essentielle pour optimiser les hyperparamètres de cet arbre de décision. Nous sélectionnons uniquement les hyperparamètres dédiés à ce type de modèle, comme indiqué dans le tableau 2.2.

Extreme Gradient Boosting (XGBoost) [82] est une méthode de boosting qui combine plusieurs modèles (comme les arbres de décision) de manière séquentielle pour former un modèle puissant. À chaque itération, XGBoost ajuste les erreurs commises par les prédictions précédentes en pondérant plus fortement les exemples mal prédits. Cela permet au modèle d'apprendre progressivement à mieux corriger les erreurs et à améliorer la précision des prédictions.

XGBoost est particulièrement performant pour les tâches de régression et est connu pour sa rapidité et son efficacité. Il intègre des techniques comme la régularisation pour éviter le surapprentissage, ainsi que des mécanismes d'optimisation avancés, ce qui en fait un modèle très performant dans de nombreuses compétitions de machine learning. Il est souvent utilisé lorsque la complexité des relations entre les variables est élevée et que des performances maximales sont requises. L'application de notre algorithme 1 est essentielle pour optimiser les hyperparamètres de cet arbre de décision. Nous sélectionnons uniquement les hyperparamètres dédiés à ce type de modèle, comme indiqué dans le tableau 2.2.

2.3.4.3 Méthodes d'apprentissage profond

L'apprentissage profond a démontré son efficacité dans la prédiction liée aux données d'exploitation, notamment avec les récentes avancées technologiques. Il est particulièrement adapté aux problèmes complexes liés aux séries temporelles, grâce à des architectures telles que les réseaux de neurones à couches multiples (ANN) [66], les LSTM [69] et les modèles BiLSTM [105]. Ces modèles excellent dans la capture des dépendances temporelles et des motifs dans les données, ce qui est crucial pour des prédictions fiables dans les séries temporelles associées aux données d'exploitation. Le principe fondamental que ces modèles exploitent est que le passé impacte le présent, qui à son tour influence le futur, et c'est précisément ce type de dépendance que ces modèles cherchent à capturer.

Les hyperparamètres dans les modèles d'apprentissage profond sont des paramètres qui ne sont pas appris directement par le modèle pendant l'entraînement, mais qui doivent être définis avant le processus d'apprentissage. Chaque modèle d'apprentissage profond (ANN, LSTM, BiLSTM, etc.) possède plusieurs hyperparamètres spécifiques qui influencent la performance et la capacité du modèle à généraliser aux nouvelles données. Comme indiqué dans le tableau 2.3, ces hyperparamètres incluent :

$$\mathcal{H} = \{\texttt{learning_rate}, \texttt{batch_size}, \texttt{neurons}, \texttt{layers}, \texttt{epochs}, \texttt{activation}, \texttt{dropout}, \texttt{recurrent_dropout}, \texttt{init}, \texttt{lstm_activation}, \texttt{l2_reg}, \texttt{max_norm_value}\},$$

sont les hyperparamètres à optimiser.

- **Le taux d'apprentissage (learning rate)** : Contrôle la taille des ajustements apportés aux poids du modèle à chaque étape de l'entraînement. Un taux trop élevé peut entraîner une convergence rapide mais inexacte, tandis qu'un taux trop faible peut ralentir l'apprentissage.
- **La taille du batch (batch size)** : Nombre d'exemples utilisés pour mettre à jour les poids du modèle en une seule itération. Les petites tailles de batch peuvent entraîner un apprentissage plus rapide, mais moins stable, tandis que de plus grandes tailles augmentent la stabilité mais ralentissent l'entraînement.
- **Le nombre d'époques (epochs)** : Nombre de fois que l'ensemble des données d'entraînement passe entièrement à travers le réseau. Un nombre élevé d'époques permet d'améliorer l'apprentissage mais peut causer un surajustement aux données d'entraînement.
- **Les fonctions d'activation (activation)** : Contrôlent la manière dont les neurones transmettent l'information entre les couches. Des fonctions comme ReLU ou Tanh introduisent de la non-linéarité et permettent d'apprendre des relations complexes dans les données.
- **Le taux de régularisation (dropout)** : Une technique utilisée pour éviter le surapprentissage en désactivant aléatoirement un pourcentage de neurones pendant l'entraînement. Cela empêche le modèle de trop dépendre de certaines connexions spécifiques.
- **Le dropout récurrent (recurrent dropout)** : Similaire au taux de régularisation classique, mais appliqué dans les réseaux récurrents comme LSTM ou BiLSTM pour éviter le surapprentissage dans les connexions récurrentes.
- **L'initialisation des poids (init)** : Méthode utilisée pour initialiser les poids du réseau, qui définit une plage de valeurs adaptée pour un entraînement efficace.

Le paramètre `units` contient des informations cruciales concernant la structure du réseau, y compris le nombre de couches et le nombre de neurones dans chaque couche.

— **Le nombre de couches (layers)** : Les couches supplémentaires permettent au modèle d'apprendre des relations plus abstraites, mais un nombre trop élevé de couches peut entraîner un surapprentissage.
— **Le nombre de neurones par couche (neurons)** : Affecte la capacité du modèle à apprendre des représentations complexes. Plus de neurones permettent de capter des détails plus fins, mais peuvent rendre le modèle plus coûteux en calculs et susceptible de surapprentissage.
— **Les fonctions d'activation LSTM (lstm_activation)** : Spécifique aux couches LSTM, cette fonction d'activation contrôle l'activation des cellules LSTM et peut être différente des fonctions d'activation utilisées dans d'autres couches du réseau.

Table 2.3 – Exploration des hyperparamètres par modèle d'apprentissage profond

Hyperparamètres	ANN	LSTM	BiLSTM
batch_size	✓	✓	✓
epochs	✓	✓	✓
optimizer	✓	✓	✓
learning_rate	✓	✓	✓
dropout_rate	✓	✓	✓
layers	✓	✓	✓
init	✓	✓	✓
activation	✓	✓	✓
lstm_activation			✓

La méthode de **recherche par grille (grid search)** [106] consiste à tester plusieurs combinaisons d'hyperparamètres pour trouver la meilleure configuration. Cette approche permet d'optimiser les performances du modèle pour un ensemble de données donné en explorant toutes les combinaisons possibles d'hyperparamètres spécifiés, comme montrée dans l'algorithme 2.

Réseaux de neurones artificiels (ANN) [66] sont parmi les premières architectures développées en apprentissage profond. Ils sont adaptés à une grande variété de problèmes, mais ne capturent pas spécifiquement les dépendances temporelles entre les données. Cependant, l'ensemble de données que nous avons préparé intègre les notions de 'passé', 'présent' et 'futur' sous forme de colonnes, ce qui permet d'examiner le comportement des ANN par rapport aux autres modèles plus adaptés aux séries temporelles. De plus, nous avons utilisé la méthode de *Grid Search*, comme illustré dans l'algorithme 2, afin d'optimiser les hyperparamètres. Bien que les ANN ne soient pas explicitement conçus pour gérer la notion temporelle, cette approche nous permet de trouver la configuration optimale et d'évaluer leur performance sur ces données structurées.

Algorithme 2 : Algorithme de Grid Search pour l'optimisation des hyperparamètres

Data : $\mathcal{H}$ (espace des hyperparamètres), X (ensemble de données d'entraînement)
Result : Meilleure combinaison d'hyperparamètres θ^{meilleur}

$\mathcal{H} \leftarrow$ définir l'espace des hyperparamètres (ex : taux d'apprentissage, taille du batch, etc.);
$\mathcal{H}^{\text{meilleur}} \leftarrow 0$; `// Initialiser la meilleure combinaison d'hyperparamètres`
score$^{\text{meilleur}} \leftarrow 0$; `// Initialiser le meilleur score`
foreach *combinaison d'hyperparamètres* $\theta \in \mathcal{H}$ **do**
    $M_\theta \leftarrow$ construire et configurer le modèle avec θ; `// Créer le modèle avec les hyperparamètres donnés`
    $M_\theta \leftarrow$ entraîner M_θ sur X; `// Entraîner le modèle sur les données d'entraînement`
    $score_\theta \leftarrow$ évaluer M_θ sur les données de validation;
    if $score_\theta >$ *score*$^{\textit{meilleur}}$ **then**
        score$^{\text{meilleur}} \leftarrow score_\theta$;
        $\mathcal{H}^{\text{meilleur}} \leftarrow \theta$; `// Mettre à jour la meilleure combinaison d'hyperparamètres`
return $\mathcal{H}^{\text{meilleur}}$;

Long Short-Term Memory (LSTM) [69] est une variante des réseaux de neurones récurrents qui surmonte le problème de la mémoire à court terme des RNN classiques. Les LSTM sont capables de conserver des informations sur de longues séquences, ce qui est essentiel lorsque des événements passés lointains influencent les prédictions futures. L'utilisation de notre algorithme 2, nous a permis d'optimiser les hyperparamètres et ainsi de maximiser la capacité du modèle à capturer l'influence temporelle. Ce modèle reflète donc l'importance du passé dans la prédiction des événements futurs. L'idée est de tester, dans un premier temps, le modèle LSTM sur les données sans les colonnes représentant le passé, mais en appliquant une fenêtre glissante pour capturer les informations passées. Ensuite, le modèle sera testé sur l'ensemble des données d'entraînement, afin de comparer les performances obtenues avec chaque approche.

Bi-directional LSTM (BiLSTM) [105] sont une extension des LSTM qui permettent de traiter les séquences temporelles dans les deux directions, capturant ainsi à la fois l'impact du passé sur le présent et celui du futur sur le passé. Cette capacité bidirectionnelle offre une compréhension plus complète des relations temporelles dans les données. Nous avons utilisé l'algorithme 2 pour optimiser les hyperparamètres, ce qui permet au modèle de capturer de manière plus efficace les interactions complexes entre le passé, le présent et le futur dans les séries temporelles. De plus, nous avons comparé cette approche avec notre méthodologie consistant à intégrer explicitement les informations du passé dans les données, afin d'améliorer les résultats sur les données d'exploitation.

2.3.5 Discussion

Notre analyse sur l'ensemble des travaux étudiés jusqu'à présent nous permet de construire une méthodologie visant à répondre à deux problématiques essentielles : 'comment intégrer un système d'aide à la prise de décision (SAD) dans un Poste de Commandement Centralisé (PCC)' et 'comment résoudre la problématique des retards dans les réseaux de transport public ferroviaire'. Notre piste consiste à développer des modèles d'apprentissage pour prédire les heures de départ des rames dans les stations, afin de faciliter la gestion opérationnelle dans les PCC.

La figure 2.7 et le tableau 2.4 représentent les différentes catégories associées à l'analyse des travaux étudiés, identifiées par des abréviations en majuscules pour plus de clarté. Les lettres utilisées correspondent aux éléments suivants : R pour "Retard", RFP pour "Réseau ferroviaire public", DE pour "Données d'exploitation", DC pour "Données contextuelles", DH pour "Données historiques", HLT pour "Horizon long terme", HCT pour "Horizon à court terme", HS+1 pour "Prédiction sur la station suivante", et HS+f pour "Prédiction sur plusieurs stations". Ces catégories permettent de visualiser la répartition des études selon leurs caractéristiques principales.

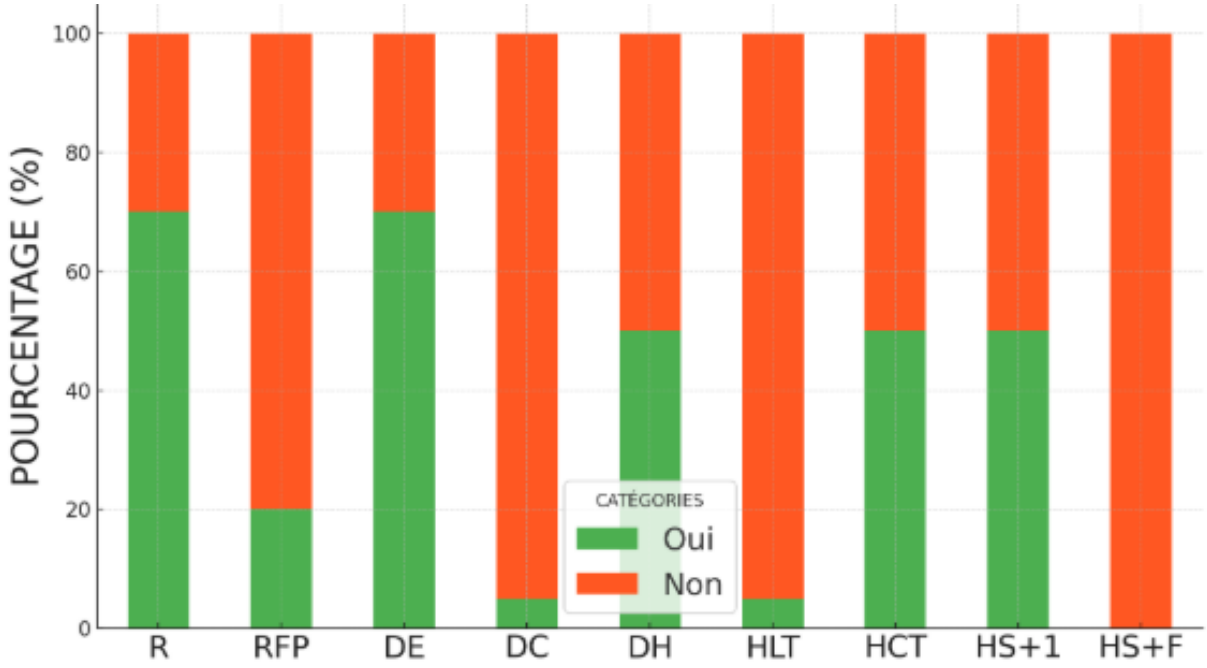


Figure 2.7 – Répartition des contributions des articles sur les problématiques ferroviaire.

Le tableau 2.4 présente les contributions des articles étudiés en lien avec la problématique des retards. Ces contributions sont analysées en fonction de plusieurs éléments essentiels pour notre méthodologie, notamment les données utilisées, les horizons temporels considérés, et les objectifs spécifiques des travaux. Ces éléments sont également pertinents pour la gestion des retards dans les PCC, avec pour objectif d'améliorer la prise de décision dans ces environnements critiques.

Les articles étudiés montrent une diversité de contributions aux problématiques liées aux réseaux ferroviaires. Concernant la problématique du retard (R), 70 % des articles abordent explicitement cette question, soulignant son importance pour la gestion et la réduction des retards dans les études récentes. Cependant, 30 % des articles ne

traitent pas directement ce sujet, reflétant des priorités de recherche alternatives ou complémentaires.

Table 2.4 – Comparaison des travaux liés à la problématique des retards dans les réseaux ferroviaires.

Travaux	**R**	**RFP**	**DE**	**DC**	**DH**	**HLT**	**HCT**	**HS+1**	**HS+f**
Camarazo et al. [52]	×	×	✓	×	×	×	×	×	×
Zehrouni [56]	×	×	✓	×	×	×	×	×	×
Jiang et al. [70]	×	✓	×	×	×	×	×	×	×
Berger et al. [72]	✓	×	✓	×	✓	×	×	×	×
Markovic et al. [61]	✓	×	✓	×	✓	×	✓	✓	×
Dong et al. [74]	✓	×	×	×	✓	×	×	×	×
Gorman [78]	✓	✓	✓	×	✓	✓	×	×	×
Corman & Merg [79]	✓	×	×	×	×	×	✓	✓	×
Yaghini et al. [65]	✓	×	✓	×	✓	×	✓	✓	×
Mou et al. [67]	✓	×	✓	×	✓	×	✓	✓	×
Flier et al. [80]	✓	×	✓	×	✓	×	×	×	×
Kecman & Goverde [62]	✓	×	✓	×	×	×	✓	✓	×
Rui Shi et al. [81]	✓	×	✓	×	✓	×	✓	✓	×
Xu Bao et al. [46]	✓	×	×	×	×	×	✓	✓	×
Wen et al. [91]	✓	×	×	×	×	×	✓	✓	×
Marsh et Bearfield [92]	×	×	×	✓	×	×	×	×	×
Rémi Couluad [60]	×	✓	✓	×	×	×	×	×	×
Toqué [59]	×	✓	✓	×	×	×	×	×	×
Sobrie et al. [94]	✓	×	✓	×	✓	×	✓	✓	×
Yu et al. [95]	✓	×	✓	×	✓	×	✓	✓	×
Chtioui et al. [96]	✓	✓	✓	✓	✓	×	✓	✓	✓

Pour le réseau ferroviaire public (RFP), seuls 20 % des articles se concentrent sur ce type de réseau, tandis que 80 % explorent d'autres infrastructures ou ne précisent pas cette catégorie. Cela met en évidence un manque d'études dédiées aux défis spécifiques des réseaux publics.

En ce qui concerne les types de données utilisées, 70 % des articles intègrent des données d'exploitation (DE), souvent en temps réel, alors que seulement 5 % exploitent des données contextuelles (DC), comme les horaires ou les conditions environnementales. Les données historiques (DH) sont utilisées dans 50 % des cas, indiquant une intégration plus fréquente mais encore partielle de ces informations pour enrichir les analyses et modèles prédictifs.

Pour les horizons temporels, 50 % des articles abordent des prédictions à court terme (HCT), tandis que seuls 5 % considèrent des prévisions à long terme (HLT). De plus, 50 % des articles incluent des prédictions pour la station suivante (HS+1), mais aucune étude ne propose des prédictions sur un ensemble de stations (HS+f). Ces résultats montrent une concentration des recherches sur des prédictions locales et immédiates, laissant des opportunités pour des approches plus globales et intégrées.

Cette analyse nous offre une vision globale des données et des approches nécessaires pour intégrer un SAD dans notre méthodologie. Nous proposons de nous

concentrer sur les données historiques, capturant les informations des stations précédentes ainsi que des trains successeurs et prédécesseurs. Ces données seront exploitées via une fenêtre glissante pour capturer les tendances historiques et les relations temporelles.

Par ailleurs, notre méthodologie mettra l'accent sur un "horizon temporel à court terme", tout en considérant les prédictions sur plusieurs stations en aval. Cela permettra de donner aux opérateurs un délai supplémentaire pour exploiter ces prédictions dans leur gestion quotidienne, améliorant ainsi l'efficacité opérationnelle et la prise de décision dans les PCC.

Cette combinaison d'une analyse historique approfondie et d'une vision élargie des stations suivantes représente une approche innovante, répondant aux lacunes identifiées dans les contributions existantes tout en s'alignant sur les besoins opérationnels des réseaux ferroviaires publics.

2.4 Conclusion

Ce chapitre a permis d'établir une revue détaillée de l'état de l'art sur les problématiques liées à la gestion et à la prédiction des retards dans les réseaux ferroviaires, en mettant l'accent sur l'intégration de SAPD dans les PCC. À travers une analyse approfondie, plusieurs approches méthodologiques et techniques ont été identifiées, offrant des perspectives novatrices pour répondre aux défis opérationnels spécifiques aux réseaux ferroviaires.

Dans un premier temps, nous avons souligné l'évolution des infrastructures ferroviaires et des systèmes de supervision, comme les CTC et les ATS, qui représentent des fondations solides pour l'intégration de solutions prédictives. Ces systèmes, bien que performants, montrent des limites face aux exigences croissantes en termes de réactivité et de gestion proactive des perturbations.

Ensuite, les travaux analysés ont mis en lumière les avancées réalisées dans la modélisation des retards et la conception de solutions prédictives, en particulier grâce aux modèles d'apprentissage automatique et profond. Ces modèles exploitent des données hétérogènes – historiques, contextuelles et opérationnelles – pour anticiper les perturbations et améliorer la prise de décision. Cependant, des lacunes persistent, notamment concernant l'application de ces approches aux réseaux de transport public comme les tramways et métros, où les besoins opérationnels diffèrent des réseaux régionaux ou à grande vitesse.

Enfin, les contributions des travaux étudiés nous ont permis de définir les bases méthodologiques nécessaires à l'élaboration d'un SAPD adapté. En intégrant des données contextuelles et historiques, nous proposons de développer des modèles prédictifs capables de répondre aux besoins spécifiques des PCC liés au retard, tout en tenant compte des particularités des réseaux de transport public.

Le chapitre suivant présente notre méthodologie, conçue pour répondre aux limitations identifiées dans ce chapitre. Cette approche vise à développer un système de prédiction des retards dans le transport public à court terme. Elle s'appuie sur la mobilisation complète des données d'exploitation et l'utilisation d'une fenêtre glissante dans le passé, permettant de capturer les relations entre le passé, le présent et le futur. L'objectif final est d'intégrer ce modèle dans un prototype d'interface homme-machine (IHM), destiné à être installé dans un PCC. Ce système a pour vocation d'améliorer la gestion opérationnelle et la prise de décision en temps réel.

3

Méthodologie : système d'aide à la prise de décision

Sommaire

2.1 Transport ferroviaire urbain 17

2.1.1 Histoire du ferroviaire urbain 18

2.1.2 Contrôle de trafic centralisé (CTC) 19

2.1.3 Timetable Management (TTM) dans les systèmes ferroviaires . 20

2.1.4 Discussion 21

2.2 Systèmes d'aide à la prise de décision dans un PCC 22

2.2.1 Présentation générale des systèmes d'aide à la prise de décision (SAPD) 22

2.2.2 Intégration du SAPD dans un PCC 23

2.2.3 Une vision améliorée de la régulation du trafic ferroviaire grâce au SAPD 25

2.2.4 Discussion 25

2.3 Intelligence artificiel en vue de l'intégration d'un SAPD dans PCC 27

2.3.1 Concepts et cadre général de l'intelligence artificielle 27

2.3.2 Travaux menés dans l'Intelligence Artificielle appliquée au secteur des transports publics 30

2.3.3 Travaux menés dans les méthodes d'évaluation et de validation des modèles 36

2.3.4 Contributions existantes relatives aux modèles d'apprentissage 43

2.3.5 Discussion 51

2.4 Conclusion 53

Après avoir présenté l'état de l'art sur le système ferroviaire, sa relation avec les modèles d'apprentissage et d'IA, ainsi que le positionnement scientifique de notre méthodologie par rapport aux approches existantes, proposé dans le chapitre 2, nous nous intéressons désormais à la méthodologie à mettre en œuvre pour valider le travail sur le SAPD. L'objectif est d'orienter la prise de décision dans le PCC vers une approche novatrice, basée sur les modèles d'IA.

Dans le but d'optimiser et d'orienter les prises de décision dans le cadre de l'exploitation ferroviaire, nous avons proposé une méthodologie mobilisant un système d'aide à la prise de décision sous la forme d'un écran supplémentaire, conçu pour fonctionner en synergie avec les opérateurs de PCC. Inspiré des écrans déjà utilisés dans ces postes, ce système se distingue par sa capacité à anticiper les retards grâce à des prédictions avancées, apportant une assistance précieuse dans les situations critiques. Ce système utilise les données historiques des lignes ferroviaires et s'appuie sur des modèles d'apprentissage automatique sélectionnés à la suite d'une étude comparative, tenant compte des contraintes de temps et de budget. Ces modèles fournissent des prévisions à court terme sur les retards quotidiens pour chaque rame dès son arrivée en station, permettant une gestion proactive et adaptée du trafic. Comme détaillé dans le chapitre 2, notre étude met en évidence un déficit significatif d'intégration de l'intelligence artificielle dans les PCC. Cela s'explique principalement par le manque de données disponibles et par la réticence des experts à adopter des systèmes de décision automatisés. Pour répondre à ces enjeux, notre approche dite non intrusive ne vise pas à remplacer les opérateurs, mais à leur offrir un outil complémentaire capable d'apporter un soutien ciblé et efficace lorsque cela est nécessaire.

La figure 3.1 illustre la conception de notre écran, représentant l'interface destinée à être intégrée dans les PCC. Ce prototype a été conçu pour visualiser en temps réel les sorties de notre modèle d'apprentissage automatique, en exploitant les données fournies par Egis Rail et les attentes de leurs opérateurs PCC. Il permet de suivre la circulation des rames et de prédire si celles-ci seront en retard ou en avance dans un futur proche. Les résultats des prédictions sont affichés de manière intuitive grâce à un code couleur et des symboles, facilitant ainsi l'interprétation par les opérateurs. Le symbole "R" indique un retard, tandis que "A" signale une avance.

Suivant les exigences remontées, les couleurs associées aux prédictions ont été définies comme suit :

- **Retards :**
  - **Vert** avec un symbole "R" : Retard $\leq$ 5 minutes.
  - **Orange** avec un symbole "R" : 5 minutes $<$ Retard $\leq$ 10 minutes.
  - **Rouge** avec un symbole "R" : Retard $>$ 10 minutes (retard critique). La couleur rouge clignote pour attirer l'attention de l'opérateur.
- **Avances :**
  - **Vert** avec un symbole "A" : Avance $\leq$ 5 minutes.
  - **Orange** avec un symbole "A" : 5 minutes $<$ Avance $\leq$ 10 minutes.
  - **Rouge** avec un symbole "A" : Avance $>$ 10 minutes. La couleur rouge clignote également pour signaler l'anomalie.

Cette interface offre une visualisation claire et ergonomique des prédictions pour chaque rame. En simplifiant l'interprétation des données, elle facilite une prise de décision rapide et efficace par les opérateurs, leur permettant d'intervenir rapidement pour réguler le trafic et minimiser les perturbations. Les prévisions des retards affichées sur l'écran supplémentaire 3.1 reposent sur deux notions temporelles clés : la fenêtre de passé et l'horizon de prédiction à court terme. Ces deux dimensions temporelles permettent d'analyser les données historiques tout en anticipant les retards futurs.

La **fenêtre de passé** se concentre sur l'analyse des données historiques d'une rame. Elle capture les relations entre la rame étudiée et ses prédécesseurs ou successeurs sur la ligne ferroviaire. L'objectif principal est d'examiner l'impact des retards passés et leurs effets consécutifs, en identifiant les schémas répétitifs ou les anomalies dans le

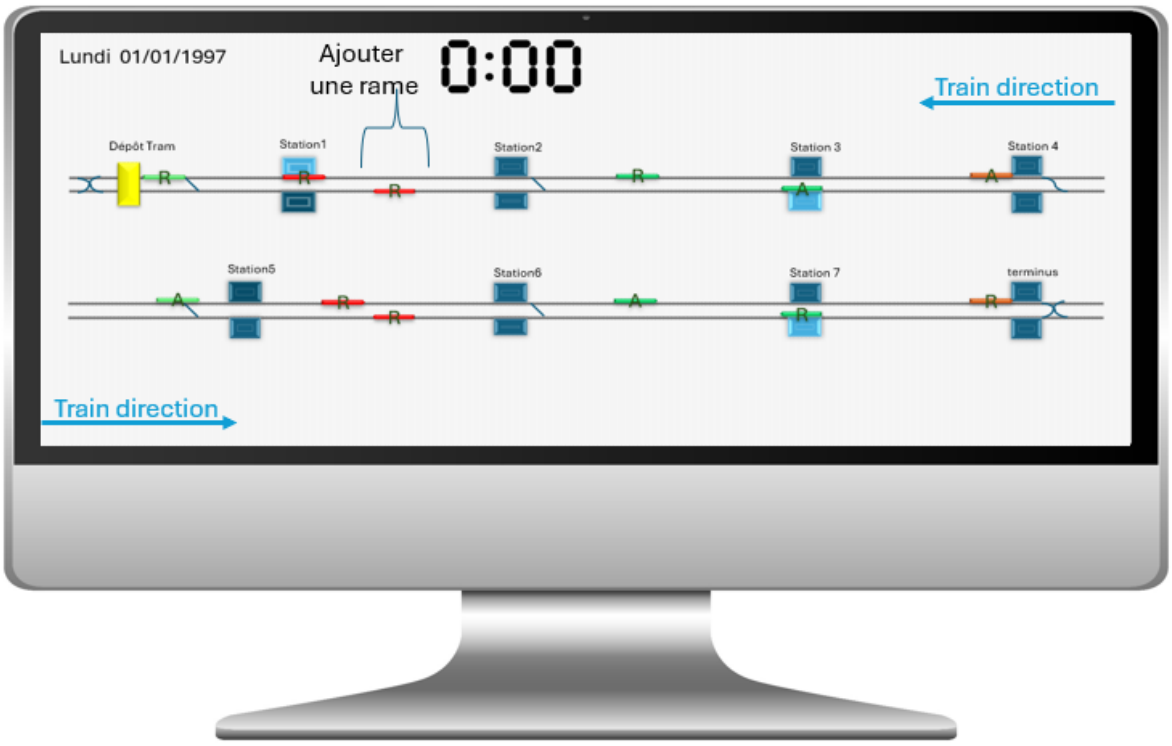


Figure 3.1 – Conception de l'écran prototype destiné à être intégré dans les Postes de Commande Centralisée (PCC), illustrant l'affichage des retards en temps réel et les interactions possibles pour la gestion des incidents

comportement des rames. Cette approche permet d'exploiter les données historiques pour mieux comprendre les dynamiques du trafic ferroviaire et de prévoir la demande en fonction des tendances observées.

L'**horizon de prédiction à court terme** se concentre sur le futur immédiat de la rame étudiée. Il s'appuie sur son emplacement actuel et projette son positionnement dans un futur proche, en tenant compte des retards récents et des estimations pour les prochaines stations de son parcours. Cette approche intègre également des caractéristiques contextuelles pour affiner les prévisions, telles que le type de jour (vacances, jours fériés, week-ends, etc.) ou encore les conditions spécifiques du réseau, comme des travaux ou des incidents signalés.

L'objectif principal du système est d'anticiper les perturbations potentielles, qu'il s'agisse de retards ou d'avances imprévues, afin d'optimiser la planification des services de transport. En tirant parti à la fois de l'analyse historique et des prévisions en temps réel, ce système propose des estimations précises et adaptées aux besoins opérationnels. Cette double perspective permet d'ajuster les horaires et les ressources en temps réel, tout en tenant compte des situations atypiques, comme des incidents ou des fluctuations de la demande. Grâce à cette approche combinée, la gestion des retards est significativement améliorée, offrant une capacité d'intervention rapide et proactive. Cela contribue non seulement à la régularité du trafic, mais également à une meilleure satisfaction des usagers en minimisant les impacts des perturbations sur l'ensemble du réseau.

Dans ce chapitre, nous détaillons la méthodologie adoptée pour la mise en œuvre de notre système d'aide à la décision destiné à optimiser la gestion des retards ferro-

viaires. Nous commençons par présenter l'approche proposée, qui comprend le développement et la validation du modèle, suivi de la collecte et de l'analyse des données nécessaires. La section sur les données couvre l'introduction générale, les sources spécifiques utilisées pour les données d'exploitation, et la description détaillée des données collectées. Nous aborderons également la modélisation relationnelle, qui vise à améliorer l'aide à la décision en intégrant les données de manière cohérente. Enfin, nous détaillerons le prétraitement des données, l'analyse des retards et la sélection des modèles d'apprentissage automatique.

Le plan du chapitre est présenté comme suit : La section 3.1, démarche proposée, introduit la méthodologie générale adoptée pour le développement et la validation du système d'aide à la décision. La section 3.2, sources de données, présente les différents ensembles de données exploités dans ce projet et l'évolution de la collecte des jeux de données , incluant le tramway et le RER, et la modélisation relationnelle, qui aborde l'organisation des données en un modèle relationnel visant à améliorer l'aide à la décision. Dans cette section, des sous-sections détaillent le processus de documentation, les visites en centre de contrôle, et la création du modèle relationnel. La section 3.3, prétraitement des données, explique les étapes appliquées aux données brutes, comme le nettoyage, la correction des incohérences temporelles et des valeurs aberrantes, ainsi que l'enrichissement et la transformation des variables. Ensuite, la section 3.4, analyse des données, explore les caractéristiques des données et leur pertinence pour la modélisation prédictive. La section 3.5, conception des caractéristiques pour la modélisation prédictive, détaille la mise en forme des données pour la prévision des retards, en incluant l'analyse des retards passés et l'anticipation des rames. Enfin, la section 3.3, approches d'apprentissage, présente les méthodes utilisées, incluant les méthodes statistiques, les arbres de décision, et les réseaux de neurones, en insistant sur l'utilisation de XGBoost et d'autres techniques pour une analyse comparative approfondie. Enfin, la section 3.8, analyse de sensibilité, examine la robustesse de la méthodologie en évaluant la sensibilité des performances des modèles aux variations des infrastructures.

3.1 Démarche proposée

La problématique des retards quotidiens, identifiée et définie lors de la phase d'état de l'art en collaboration avec les experts d'Egis Rail, constitue le point de départ de notre étude. Après avoir analysé les différentes approches et solutions existantes mentionnées dans le chapitre 2, nous avons développé une méthodologie adaptée pour concevoir un système applicable en conditions réelles. La figure 3.2 illustre une méthodologie combinée qui repose sur deux axes complémentaires : un mode hors ligne et un mode en ligne. Ces deux modes s'appuient sur une analyse approfondie de l'influence des données temporelles, incluant le passé, le présent et le futur. Cette démarche permet de définir un horizon de prédiction à court terme, optimisé pour anticiper les retards et faciliter la prise de décision en temps réel.

Le mode hors ligne débute par la récupération des données historiques couvrant une période suffisante pour permettre une analyse approfondie et l'apprentissage. La phase 1 de la figure 3.2 présente la construction d'un modèle relationnel. Cette étape repose sur l'utilisation d'un diagramme de classes qui modélise de manière structurée les relations entre les variables. Une analyse détaillée des données issues des infrastructures Tram et RER a été réalisée pour identifier les variables ayant un impact direct

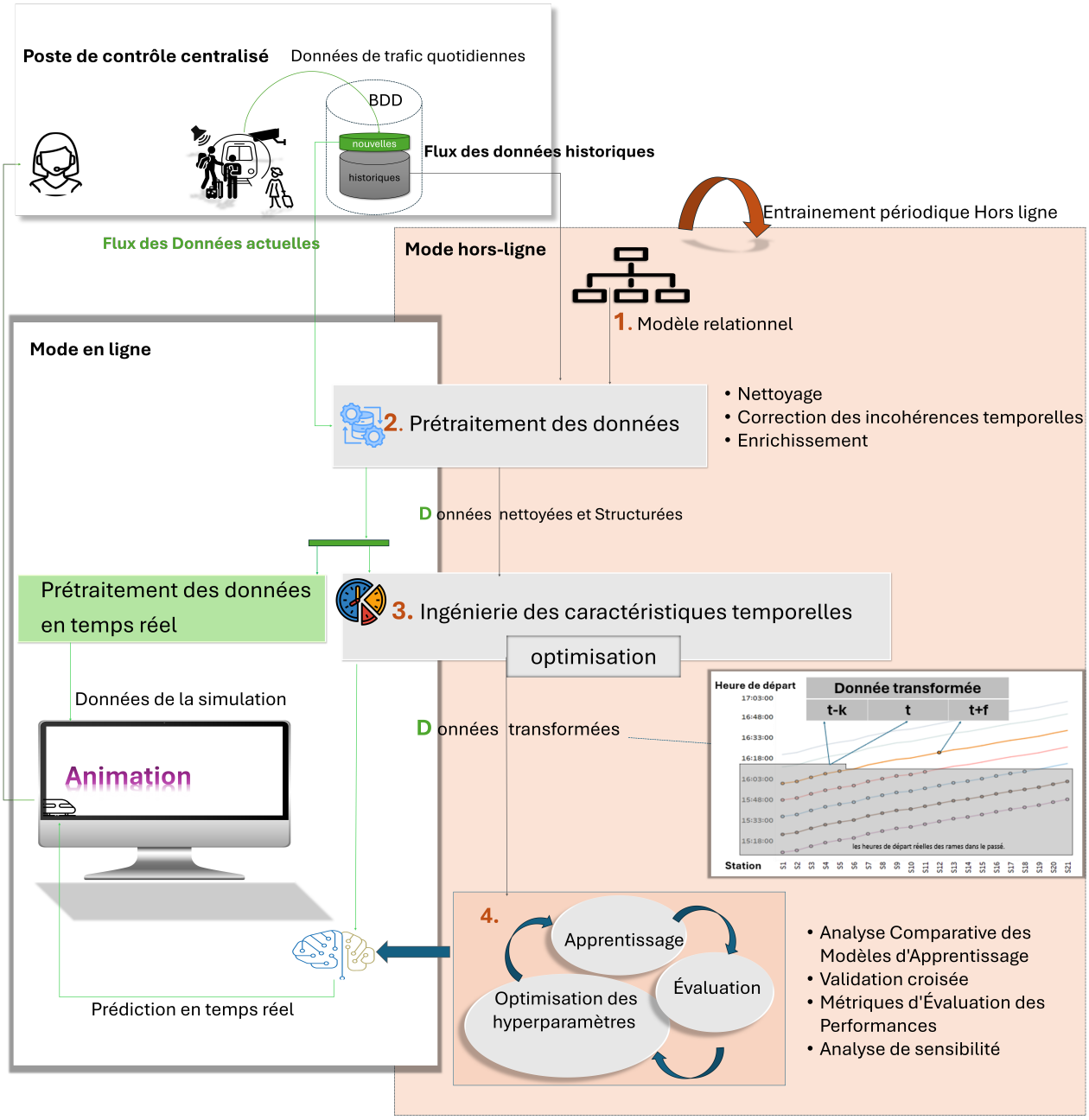


Figure 3.2 – Méthodologie pour l'intégration de l'IA dans les postes de contrôle ferroviaire, de l'analyse des données à l'évaluation des modèles de prédiction des retards.

sur les retards. Ce modèle relationnel a été validé avec des experts ferroviaires d'Egis Rail afin de faciliter les discussions et d'illustrer les relations entre les variables et leur rôle dans les modèles d'apprentissage. Par exemple, une variable telle que "l'heure de départ" correspond à une entité spécifique, issue de le tableau TTP dans la base de données.

La phase 2 est dédiée au prétraitement des données historiques. Ce processus comprend des étapes importantes, notamment le nettoyage des données, la gestion des valeurs manquantes, la correction des incohérences temporelles, et l'enrichissement des données à l'aide d'informations externes telles que les données météorologiques.

Dans la phase 3, une ingénierie des caractéristiques temporelles est appliquée pour transformer les données et extraire des informations exploitables par les modèles d'apprentissage. Cette étape repose sur trois dimensions temporelles principales, illustrées dans la figure 3.3. La première dimension, le passé, analyse les comportements historiques des rames sur une ligne donnée. En effet, les heures de départ réelles des rames ayant déjà quitté chaque station sont utilisées pour modéliser les schémas de retard ou d'avance. Ces données historiques permettent d'identifier des patterns récurrents ou des anomalies, fournissant un contexte important pour les prédictions futures. La deuxième dimension, le présent, se concentre sur l'état actuel de la rame au moment de son départ d'une station. Cela inclut les heures de départ réelles des rames en temps réel, qui sont essentielles pour évaluer les conditions opérationnelles actuelles. Ces variables servent également de point de référence pour ajuster les prédictions futures. La troisième dimension, le futur, comprend les variables prédictives à court terme, telles que l'heure de départ prévue pour les prochaines stations. Ces données permettent de simuler les comportements attendus des rames en fonction des conditions actuelles et des schémas historiques identifiés. Par exemple, comme montré dans la figure 3.3, les heures de départ prévues pour le futur sont associées aux stations à venir, offrant une projection dynamique du trafic ferroviaire. Une optimisation des caractéristiques est ensuite réalisée pour sélectionner les variables les plus pertinentes et réduire la complexité des modèles.

La phase 4 consiste en une étude comparative de plusieurs modèles d'apprentissage, incluant des approches de machine learning et de deep learning. Chaque modèle est optimisé à l'aide de techniques spécifiques d'optimisation des hyperparamètres. L'évaluation des performances s'appuie sur des métriques classiques comme le coefficient de détermination (R^2) et la racine de l'erreur quadratique moyenne (RMSE), ainsi qu'une métrique spécifique, appelée $P(\%)$, développée avec les opérateurs pour valider la fiabilité des modèles. Une analyse de sensibilité est également menée pour évaluer l'impact des facteurs clés sur la précision des prédictions, tout en testant la robustesse du modèle sur un autre jeu de données.

Le mode en ligne démarre avec le début de la journée d'exploitation. Comme l'accès direct à une base de données de PCC n'est pas disponible, les données historiques sont divisées en deux parties : une pour simuler les données historiques du mode hors ligne, et une autre pour servir de données instantanées simulées en temps réel. À l'arrivée de nouvelles données actuelles, celles-ci subissent le même prétraitement que celui appliqué dans le mode hors ligne afin de garantir leur cohérence. Une interface utilisateur ergonomique a été développée pour intégrer le modèle sélectionné, offrant une utilisation fluide dans les PCC. Parallèlement, une animation en temps réel a été mise en place pour simuler les départs et déplacements des rames, tout en gérant

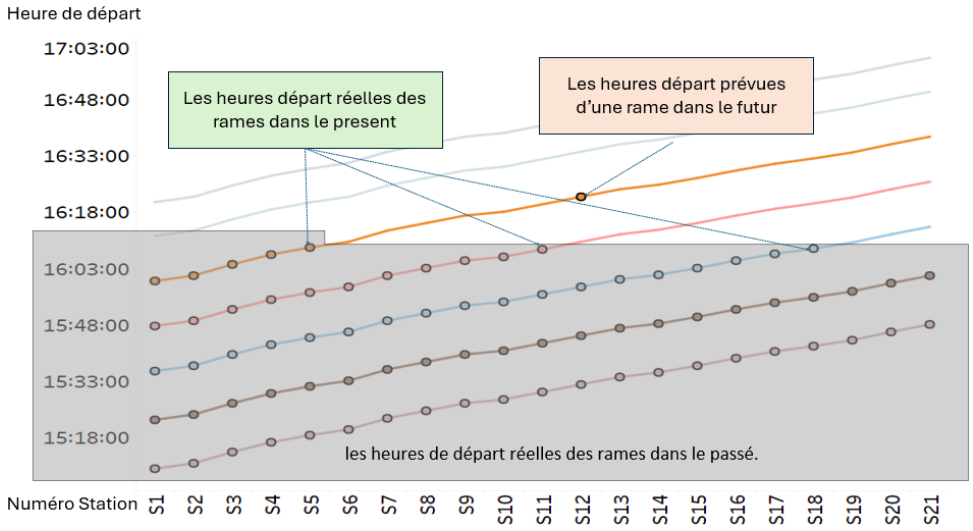


Figure 3.3 – Interaction des heures de départ des rames : relation entre les départs passés, présents, et futurs

les exceptions et les erreurs potentielles. Cette animation combine les informations historiques et actuelles pour générer des prédictions à court terme adaptées aux nouveaux ensembles de données.

La section suivante présente les données d'exploitation disponibles dans les postes de contrôle ferroviaire. Ces données sont organisées à l'aide d'un modèle relationnel structuré, présenté dans la phase 1 de la méthodologie (voir figure 3.2). Ce modèle organise les informations de manière systématique afin d'identifier les relations entre les différentes entités impliquées, telles que les rames, les stations, et les horaires.

3.2 Approche relationnelle pour la modélisation des données ferroviaires

Chaque jour, des millions de lignes de données ferroviaires sont stockées dans les bases de données des PCC. Elles proviennent de diverses sources, telles que les capteurs embarqués sur les rames, les systèmes de signalisation, les caméras de surveillance, et les dispositifs de suivi des conditions météorologiques. Par exemple, les informations sur la position des trains, les horaires de départ et d'arrivée, les retards, ou encore l'état des infrastructures ferroviaires sont collectées en temps réel. Dans un scénario typique, lorsqu'une rame quitte une station, les données relatives à son départ, à sa vitesse et à son itinéraire sont immédiatement enregistrées. Si un retard survient, les informations correspondantes, y compris les raisons du retard (panne technique, conditions météorologiques défavorables, etc.), sont également capturées et intégrées dans la base de données. Les opérateurs de PCC utilisent ces données pour assurer la gestion du trafic : résolution des pannes et des incidents, gestion de l'affluence des voyageurs, garantie de la sécurité des passagers à bord des trains et sur les quais, suivi des horaires réels des départs et des arrivées, gestion des retards excessifs, et bien d'autres tâches encore.

Dans notre étude, nous avons travaillé avec deux ensembles de données d'exploitation liés aux horaires prévus et réels, respectivement appelés TTP (Timetable Prévu)

et TTR (Timetable Réel). Ces ensembles de données contiennent des informations essentielles sur la circulation des rames sur leur ligne ferroviaire. Le TTP fournit les horaires planifiés pour chaque rame, incluant les heures de départ et d'arrivée prévues à chaque station, tandis que le TTR enregistre les horaires réels observés, prenant en compte les éventuels retards ou modifications d'itinéraires. En comparant ces deux ensembles de données, nous pouvons analyser les écarts entre les prévisions et la réalité, ce qui est crucial pour la modélisation des retards et l'optimisation des décisions dans les situations opérationnelles.

La figure 3.4 illustre les ressources fournies par Egis Rail, notamment les ensembles de données. En complément, des fichiers internes ont été partagés, jouant un rôle essentiel pour approfondir notre compréhension des différents types de données disponibles ainsi que de l'infrastructure ferroviaire. Par ailleurs, des opérateurs expérimentés ont contribué à affiner notre perception des besoins opérationnels liés à la gestion des retards, ce qui a grandement aidé à l'élaboration du modèle relationnel.

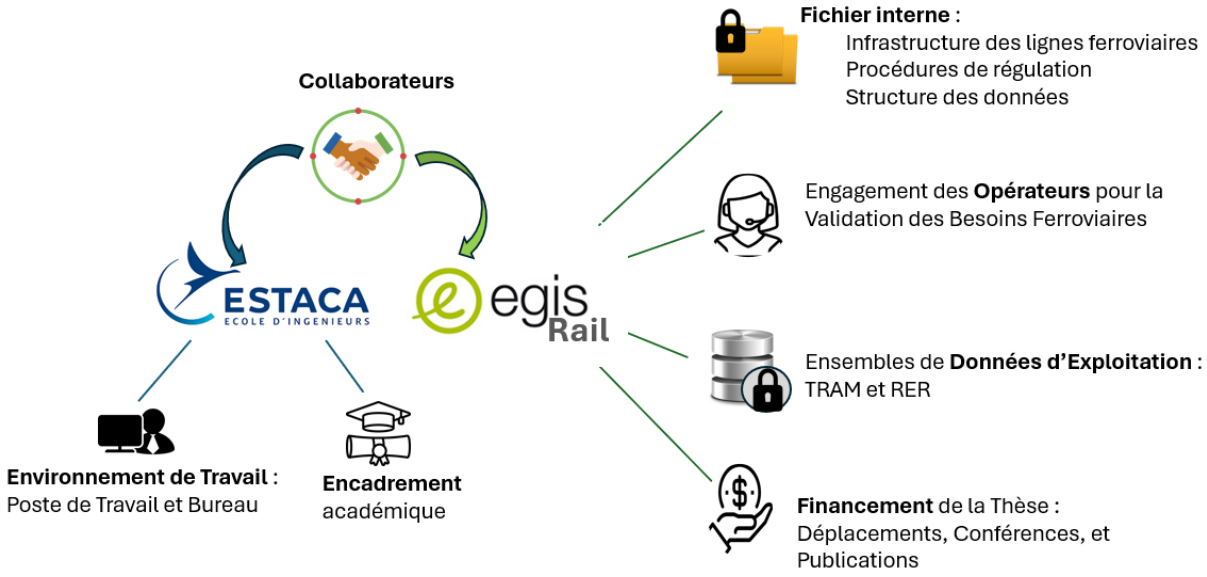


Figure 3.4 – Cartographie des ressources utilisées : synergie entre l'environnement académique (Estaca) et industriel (Egis-rail) au cœur de la thèse

L'organisation des données est cruciale avant de commencer toute démarche sur les exigences d'un système ferroviaire, en particulier pour développer un modèle capable de traiter des phénomènes complexes comme les retards ou les interruptions. Parmi les étapes essentielles figure la modélisation des données, qui doit être réalisée avant même le nettoyage ou le prétraitement. En effet, sans une compréhension claire de la provenance des données et des relations entre les différentes entités, il devient difficile de définir précisément les besoins et les fonctionnalités à intégrer. Le modèle relationnel se présente comme l'une des méthodes les plus adaptées à cette situation. Avec un ensemble de données initial limité, décrivant par exemple TTP et TTR, les retours d'expertise des opérateurs, ainsi que les fichiers internes d'Egis décrivant l'infrastructure et la base de données interne d'un PCC, il devient plus aisé de réaliser une modélisation globale. Cela permet de regrouper toutes les variables d'une ligne sur une journée, facilitant ainsi l'identification des besoins, la gestion des données et la mise en place des fonctionnalités clés pour une gestion efficace du trafic ferroviaire.

Le diagramme de classe [53] est notre choix pour structurer et modéliser les

données dans le cadre de notre projet ferroviaire. Ce type de diagramme permet de représenter les entités clés du système, leurs attributs, ainsi que les relations entre elles, offrant une vue d'ensemble claire et précise. En capturant les interactions entre les trains, les stations, les lignes, les horaires, et les postes de commande, le diagramme de classe nous aide à mieux comprendre la complexité des données et à définir les exigences du système. Grâce à cette modélisation, nous pouvons non seulement clarifier les dépendances entre les différentes entités, mais aussi anticiper les besoins en données et les fonctionnalités à intégrer. Ce choix garantit une approche structurée et efficace pour aborder les défis liés à la gestion du trafic ferroviaire et à l'optimisation des retards.Cette section décrivant notre modèle relationnel est tirée d'un article que nous avons présenté lors de la "7th International Conference on Intelligent Traffic and Transportation (2023)", tenue du 18 au 20 septembre 2023.

3.2.1 Méthode de travail

La construction de notre modèle relationnel s'est déroulée en trois phases principales, menant au développement d'un modèle cohérent et adapté à une ligne ferroviaire. La figure 3.5 illustre ces trois étapes essentielles de notre méthode.La première phase consiste en une analyse documentaire, qui repose sur l'examen approfondi d'articles scientifiques et techniques similaires. Cette étape vise à identifier les types de données utilisées dans les algorithmes existants pour représenter les contraintes d'un système ferroviaire, en particulier celles liées aux retards, aux capacités des infrastructures et aux flux de passagers. L'analyse de l'état de l'art, qui se réfère à l'état actuel des connaissances et des pratiques dans le domaine, nous permet d'extraire des éléments pertinents pour notre contexte. En simplifiant des ensembles de données complexes grâce à cette étude, nous pouvons les rendre plus adaptés à une modélisation précise.

La deuxième phase, fondée sur des échanges approfondis avec les opérateurs des centres de contrôle, a permis d'obtenir des retours d'information précieux sur la pertinence et l'applicabilité des modèles relationnels. Ces discussions avec les experts du centre de contrôle ferroviaire ont joué un rôle crucial dans la validation des modèles à partir de cas réels, en bénéficiant de l'expertise de ceux qui maîtrisent parfaitement le système. Grâce à ces échanges répétés, nous avons pu extraire des informations bien plus riches que celles obtenues en nous basant uniquement sur des sources externes. Cette compréhension approfondie des données et de leur contexte a considérablement enrichi la phase de modélisation, rendant les modèles plus adaptés à des applications concrètes. La validation par les centres de contrôle garantit également que ces modèles pourront être facilement mis en œuvre dans des scénarios réels, en particulier dans l'intégration d'un système d'aide à la décision. Enfin, la troisième phase constitue un processus itératif visant à affiner notre compréhension des données et à formaliser un diagramme relationnel, représenté dans le rectangle jaune de la figure 3.5, facilitant ainsi une interprétation claire et structurée des informations, comme expliqué dans la section précédente.

La modélisation relationnelle des données relatives aux chemins de fer métropolitains consiste à organiser les informations d'un système ferroviaire en un ensemble de tables, chacune représentant un aspect spécifique du système. Ces tables sont reliées entre elles par des éléments de données communs, permettant ainsi d'établir des relations entre les différentes entités. Par exemple, une table peut contenir des données sur les stations, telles que leur nom et leur emplacement, tandis qu'une autre

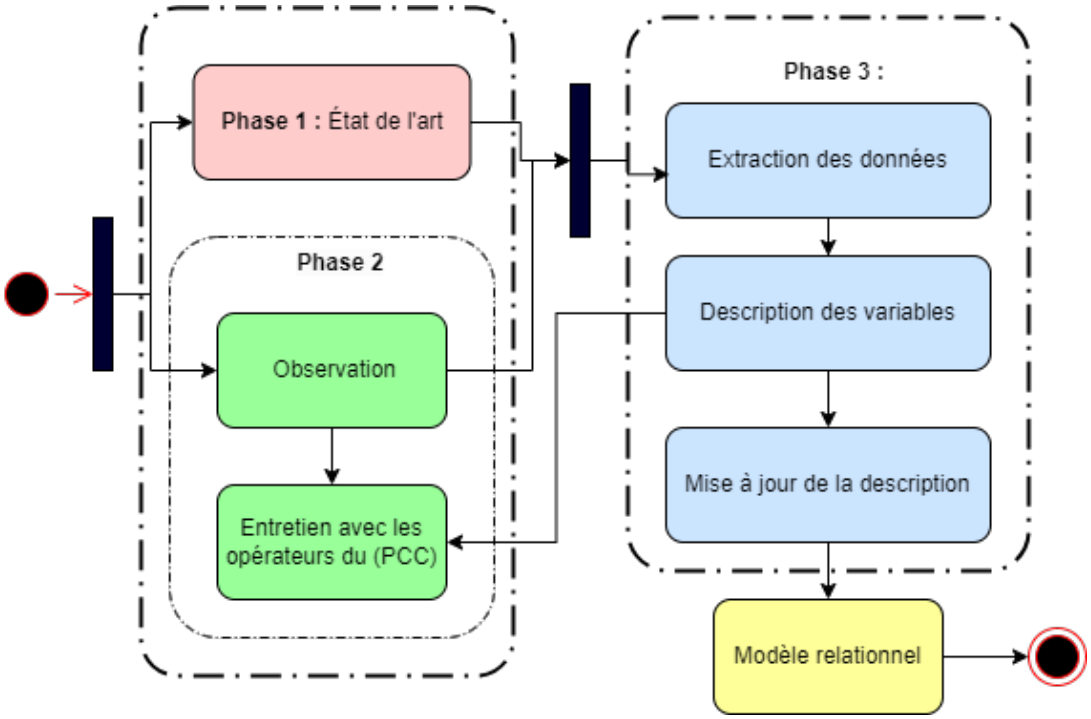


Figure 3.5 – Illustration du processus de modélisation des données relationnelles

peut répertorier les rames circulant sur le réseau, avec leurs horaires et leurs capacités. En reliant ces tables via un élément commun, comme l'identifiant de la station, il devient possible de répondre à des questions complexes, telles que : À quelle heure une rame arrive-t-elle en gare ? Ce type de modélisation facilite la gestion des informations et permet de naviguer efficacement dans des données volumineuses pour extraire des connaissances pertinentes.

3.2.2 Création du modèle relationnel

La modélisation des données commence par l'identification des concepts clés du système, lesquels sont ensuite représentés sous forme de classes. Chaque classe possède un nom, des attributs (données) et peut implémenter des méthodes via des interfaces, définissant ainsi la structure et le comportement du système modélisé. Dans notre cas, nous avons créé un modèle pour un système de ferroviaire, avec plusieurs classes basées sur les différentes phases présentées dans la figure 3.5. Chaque classe est définie par des attributs spécifiques tels que l'identifiant de la station, l'identifiant du train, ainsi que les heures de départ planifiées et réelles. Cette section introduit un modèle relationnel à travers un diagramme de classes, afin de présenter les variables et leurs relations dans un format clair et lisible.

La figure 3.5 illustre les étapes clés de notre démarche méthodologique pour la modélisation relationnelle des données d'exploitation ferroviaire. Cette progression s'articule autour de trois phases successives : documentation, visites au centre de contrôle, et formalisation du modèle relationnel.

La première phase, intitulée « Documentation », repose sur l'analyse de l'état de l'art, combinée à l'utilisation de documents internes fournis par Egis Rail. Cette phase a permis d'identifier les entités initiales et les relations clés pour la construction d'un diagramme de classes de base. Les informations issues d'articles académiques ont été

particulièrement utiles pour définir des variables spécifiques, telles que la station, la ligne, la voie, et les rames. Des variables supplémentaires, comme celles liées à la billetterie et aux systèmes de caméra, ont également été intégrées grâce aux documents internes. Cependant, en raison de limitations en matière de données, ces aspects nécessitent des approfondissements futurs pour compléter l'analyse des charges de voyageurs.

La deuxième phase, « Visites au centre de contrôle », a consisté en des observations approfondies dans les centres de contrôle ferroviaires, complétées par des entretiens avec des opérateurs experts du domaine. Cette démarche pratique a permis non seulement de mieux comprendre le système, mais aussi de recueillir des données en temps réel et de bénéficier d'une interaction directe avec les experts, enrichissant ainsi notre modèle relationnel grâce à leurs retours professionnels. En observant directement les opérations des centres de contrôle, nous avons pu vérifier, ajuster et affiner la phase 1 afin de l'aligner avec les processus et opérations réels. afin de la valider par les experts des centres de contrôle des métros. Cette version révisée intègre de nouvelles classes de calcul, telles qu'Inter-station, Information et Intervalle, ainsi que des méthodes supplémentaires pour les classes existantes. Par ailleurs, la classe Jour a été ajoutée afin de prendre en compte l'influence des différents types de jours sur le système, incluant les événements, les conditions météorologiques et le jour de la semaine. Les experts des centres de contrôle ont souligné l'importance de ces variables, qui peuvent influer sur les retards et la congestion des voyageurs sur les quais.

La troisième phase, « Modèle relationnel », consiste en une reformulation du diagramme de classes visant à améliorer sa clarté et à séparer les sous-systèmes "Exploitation" et "Flux de voyageurs". Cette réorganisation sert de base au développement d'un système d'aide à la décision. Le modèle de classes permet ainsi de mieux identifier les dépendances entre variables pour aborder des problématiques spécifiques, telles que les retards et la gestion des flux de voyageurs. En structurant les classes en deux ensembles distincts, le diagramme offre une vue plus précise des relations au sein de chaque sous-système. Ce modèle constitue un pilier essentiel pour les phases suivantes, facilitant la mise en place d'outils d'aide à la décision dans divers aspects du système ferroviaire. En effet, grâce à cette structuration rigoureuse du modèle de classes, il devient plus aisé de reconnaître les relations et interactions entre les différentes variables. Cette approche permet ainsi de prendre des décisions éclairées en exploitant les informations contenues dans le modèle.

Le diagramme de classes présenté à la figure 3.6 illustre l'ensemble des classes et interfaces d'un système ferroviaire de la ligne de ferroviaire, réparti en deux ensembles distincts : l'un dédié à l'exploitation et l'autre aux flux de voyageurs. Cette division vise à simplifier le calcul des variables propres à chaque système, améliorant ainsi la précision globale. La classe "Exploitation" est responsable de la collecte de toutes les données relatives au trafic, telles que "ligne, station rame, voie, jour, TTP et TTR", qui constituent les données brutes du centre de contrôle. Les classes Intervalle, Inter-station et Information sont, quant à elles, dédiées au calcul des données manquantes dans le jeu de données brut, telles que les retards, les intervalles et les trajets précédents. Leur utilité a été validée lors des visites auprès des opérateurs des centres de contrôle des métros, comme expliqué à la figure 3.5.

D'un autre côté, les classes liées aux flux de voyageurs contiennent des informations concernant le nombre de personnes présentes sur les quais et dans les trains, grâce à des systèmes de caméras, des dispositifs de comptage et le poids des rames. Cela

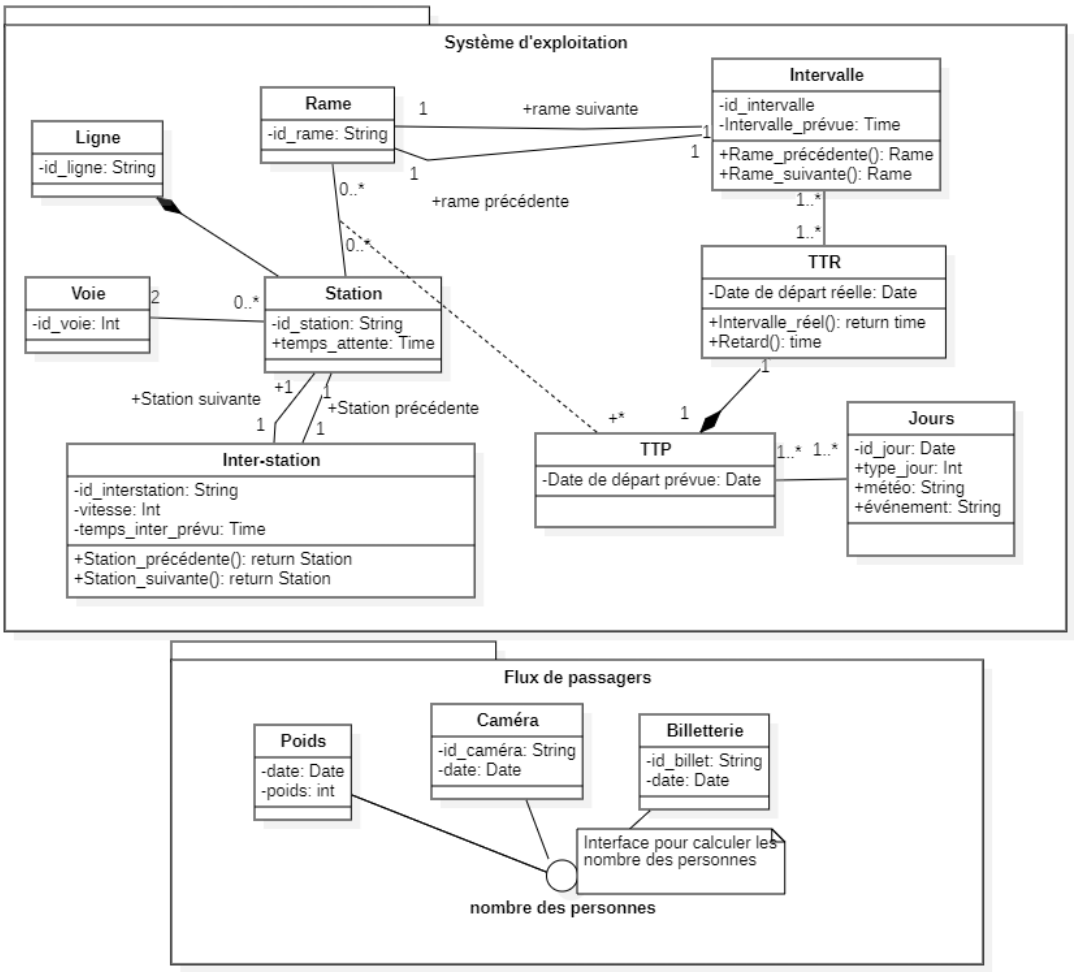


Figure 3.6 – Diagramme de classes final pour les données opérationnelles et de flux de passagers

permettra une meilleure compréhension des flux de voyageurs et aidera à anticiper les problèmes potentiels pouvant survenir en cas de forte affluence. Toutes les classes sont interconnectées, que ce soit directement ou indirectement. Par exemple, la station est reliée à la rame, à la voie, à l'intervalle, au TTP, au TTR, aux flux de voyageurs, et ainsi de suite pour chaque classe. Cette méthodologie permettra une gestion plus efficace du trafic ferroviaire en facilitant l'analyse et le traitement des données relatives à l'exploitation et aux flux de voyageurs. En séparant les ensembles de classes en différents sous-systèmes, nous pouvons plus facilement identifier les ensembles de données nécessaires pour chaque phénomène, qu'il s'agisse des retards ou des charges de voyageurs. Grâce à une simple observation du diagramme, il devient possible de spécifier quelles données sont requises pour calculer chaque variable associée à chaque phénomène. Cela simplifie grandement le travail ultérieur visant à identifier les paramètres influençant un phénomène avant de passer à la phase de modélisation par apprentissage.

Chaque opération représentée dans ce diagramme définit un ensemble d'opérations mathématiques réalisées pour calculer des variables telles que les retards réels et précédents, ainsi que les intervalles entre les trains successifs. Ces liens et opérations simplifient la lecture et la compréhension des données ferroviaires, ce qui permet de définir les données essentielles pour chaque processus d'apprentissage (comme la prédiction des retards et des flux de voyageurs pour nos futures recherches) et de visualiser les données afin d'aider les opérateurs à prendre les bonnes décisions au sein des centres de contrôle.

3.2.3 Cas d'usage

L'intégration de données réelles, obtenues en collaboration avec un PCC, garantit la fiabilité des modèles. Ces données incluent les numéros de TTP, TTR et de rame, les noms des stations, les quais et le poids des rames, comme indiqué dans le tableau 3.1. Elles permettent une adaptation aux situations concrètes, améliorant ainsi la compréhension des phénomènes et la précision des recommandations.

le tableau 3.1 présente un aperçu simplifié des données brutes récupérées pour une ligne ferroviaire typique, avec les caractéristiques suivantes, données incluent le nombre de rame n et de stations m , ainsi que les heures de départ prévues et réelles pour chaque passage d'une rame dans une station. De plus, le nombre de passagers présents dans les trains et sur les quais est estimé à l'aide du système de pesée des rames. Ce processus de modélisation des données joue un rôle crucial en facilitant la manipulation des variables et en offrant une visualisation claire de leur utilisation.

La modélisation relationnelle, comme le montre la figure 3.6, est une technique couramment utilisée dans la conception de bases de données. Elle est particulièrement bien adaptée à la gestion de jeux de données complexes avec de nombreuses informations interconnectées. En décomposant un grand ensemble de données en tables plus petites et gérables, tout en établissant des relations entre elles, cette méthode permet de réaliser des requêtes et des analyses plus efficaces.

La projection des données présentées dans le tableau 3.1 sur notre modèle relationnel a permis d'enrichir les informations disponibles et d'établir des liens entre différentes variables. Ce processus transforme les données brutes issues de le tableau 3.1 en données enrichies, présentées dans les tableaux 3.2, 3.3 et 3.4.

Après cette projection sur le diagramme de classes, nous obtenons des classes ini-

Table 3.1 – Exemple de données brutes pour n stations et m trains obtenues à partir du centre de contrôle avant l'utilisation du modèle relationnel expliqué en haut de cette section.

Heure de départ prévue	Heure de départ réelle	Voie	Rame	Station	Poids du train (tonnes)
12 :05 :00	12 :05 :56	1	t_1	s_1	280
12 :08 :00	12 :08 :52	1	t_1	s_2	175
...	...	...	...	...	...
14 :54 :00	14 :59 :00	1	t_{n-1}	s_{m-1}	130
15 :00 :00	15 :03 :33	1	t_n	s_m	100

tialement vides qui doivent être complétées avec des classes de calcul et des méthodes adaptées. Par exemple, pour chaque paire de rames consécutives, un ensemble de données nommé "Intervalle" est défini afin de calculer les intervalles prévus et réels. De plus, pour chaque triplet de stations, une classe Inter-station est introduite pour gérer les stations précédentes et suivantes associées à chaque rame. Une interface dédiée au poids des rames permet également de générer un ensemble de données utilisé pour calculer et afficher le nombre de passagers à bord de chaque rame.

Ce modèle a permis d'extraire de nouveaux attributs, détaillés dans les tableaux 3.2, 3.3 et 3.4, tout en clarifiant les relations entre les variables. Parmi les résultats obtenus, l'équation 3.1 a permis de calculer les intervalles prévus entre deux trains consécutifs, comme illustré dans le tableau 3.2. L'équation 3.3 a ensuite permis de déterminer les retards des trains en comparant les horaires de départ prévus et réels, ces résultats étant présentés dans le tableau 3.3. Enfin, l'équation 3.4 a été utilisée pour estimer le nombre de passagers à bord des rames et sur les quais. Toutefois, en raison des exigences de confidentialité imposées par le centre de contrôle, nous ne pouvons pas divulguer les détails spécifiques de cette dernière équation.

Table 3.2 – Exemple de données pour la station s_3 et m trains après projection des données brutes présentées dans le tableau 3.1 sur le modèle relationnel.

Heure de départ prévue	Heure de départ réelle	Voie	Rame	Retard (s)	Intervalle prévu (s)	Intervalle réel (s)	Nombre de personnes par rame
12 :14 :00	12 :16 :12	1	t_1	132	540	440	13
12 :22 :00	12 :24 :52	1	t_2	172	480	520	7
...	...	...	...	...	...	...	...
14 :04 :00	14 :03 :19	1	t_n	-41	240	252	20

Table 3.3 – Exemple de données pour la rame numéro 2 dans m stations après projection des données brutes présentées dans le tableau 3.1 sur le modèle relationnel.

Heure de départ prévue	Heure de départ réelle	Voie	Station	Retard (s)	Intervalle prévu (s)	Intervalle réel (s)	Nombre de personnes par rame
12 :22 :00	12 :24 :52	1	s_3	172	480	520	7
12 :22 :00	12 :26 :15	1	s_4	255	480	512	10
...	...	...	...	...	...	...	...
12 :43 :00	12 :48 :37	1	s_m	337	300	332	20

Table 3.4 – Exemple de données pour m stations et n rames après projection des données brutes présentées dans le tableau 3.1 sur le modèle relationnel.

Heure de départ prévue	Heure de départ réelle	Voie	Rame	Station	Retard (s)	Intervalle prévu (s)	Intervalle réel (s)	Nombre de personnes par rame
12 :10 :00	12 :10 :26	1	t_1	s_3	26	0	0	10
12 :14 :00	12 :14 :05	1	t_2	s_1	5	540	489	35
...	...	...	...	...	...	...	...	...
14 :54 :00	14 :59 :00	1	t_{m-1}	s_{n-1}	300	300	344	30
15 :00 :00	15 :03 :33	1	t_n	s_m	153	300	186	20

3.2.4 Représentation mathématique des données d'exploitation XOP

La modélisation du diagramme de classe présentée dans la figure 3.6 a été réalisée suivant deux objectifs principaux. Le premier objectif est de visualiser les liaisons entre les différentes variables, permettant ainsi d'identifier celles qui doivent être étudiées ou calculées. Ce diagramme offre une vue d'ensemble structurée des relations entre les variables, ce qui est essentiel pour comprendre les interdépendances et pour définir les priorités lors de l'analyse des données. Le deuxième objectif est de faciliter les discussions avec les experts du domaine ferroviaire. En effet, au lieu de présenter des formules mathématiques complexes qui peuvent être difficiles à appréhender, ce diagramme constitue un outil visuel efficace regroupant toutes les composantes pertinentes dans un schéma clair et compréhensible. Cette représentation simplifiée permet d'engager des discussions constructives avec les spécialistes ferroviaires, en favorisant une meilleure compréhension et collaboration autour des problématiques à résoudre. Enfin, après cette étape de modélisation et de discussion, nous passons à une explication mathématique plus détaillée pour décrire le problème des retards, en exprimant les relations entre les variables sous forme de formules. Cette étape vise à formaliser le problème d'apprentissage automatique en exploitant l'ensemble des données disponibles.

Nous considérons une rame d'intérêt t_i qui vient d'arriver à la station s_j. Le problème consiste à calculer le retard au départ et l'intervalle prévu de t_i à la station s_j, sachant que t_{i-1} est la rame précédent. Pour calculer les retards et les intervalles à l'aide des méthodes définies dans notre diagramme de classes, nous nous appuyons sur la projection des données brutes présentées dans le tableau 3.1 sur le diagramme de classes afin d'obtenir les données calculables présentées dans le tableau 3.4, définies comme suit :

Soit $x_{i,j} \in XOP$ où $XOP = (T, S, V, TTP, TTR, AI, A, D, PI, P, AT, Y, NP)$, l'ensemble global des données regroupant différentes catégories de données opérationnelles, chaque catégorie étant identifiée par une lettre spécifique. En utilisant cette notation, chaque élément de l'ensemble XOP est associé à une catégorie spécifique d'informations liées aux rames, aux stations, à la planification et aux heures de départ, où :

- t_i : Numéro de la rame pour chaque jour, $T = t_1, t_2, \ldots, t_n$.
- s_j : Numéro de la station, $S = s_1, s_2, \ldots, s_m$.
- v : Le sens de circulation de la rame, $V = \{0, 1\}$ avec 0 pour l'aller et 1 pour le retour.
- pt_{t_i,s_j} : Heure de départ prévue de t_i à la station actuelle s_j ($pt_{t_i,s_j} \in TTP$).

— rt_{t_i,s_j} : Heure de départ réelle de t_i à la station actuelle s_j ($rt_{t_i,s_j} \in TTR$).
— ai_{s_j} : Temps moyen prévu entre la station actuelle s_j et la station précédente s_{j-1} (secondes).
— a_{s_j} : Temps moyen passé à la station actuelle s_j (secondes).
— d : Jour de la semaine.
— pi_{t_i,s_j} : Intervalle de temps prévu entre deux trains consécutifs t_i et t_{i+1} (secondes).

$$pi_{t_i,s_j} \in PI \,|\, pi_{t_i,s_j} = pt_{t_i,s_j} - pt_{t_{i+1},s_j} \tag{3.1}$$

— p_{t_i,s_j} : Trajet suivant de t_i entre s_j et s_{j+1}.

$$p_{t_i,s_j} \in P \,|\, p_{t_i,s_j} = pt_{t_i,s_{j+1}} - pt_{t_i,s_j} \tag{3.2}$$

— at : Température moyenne pour une journée.
— y_{t_i,s_j} : Temps d'écart réel du train t_i à la station actuelle s_j, indiquant la différence entre l'heure de départ réelle et l'heure de départ prévue (secondes).

$$y_{t_i,s_j} \in Y \,|\, y_{t_i,s_j} = rt_{t_i,s_j} - pt_{t_i,s_j} \tag{3.3}$$

— nb_{t_i} : Nombre de personnes par rame.

$$nb_{t_i} \in NP \,|\, nb = ******** \tag{3.4}$$

Le modèle relationnel proposé permet une meilleure compréhension des relations au sein d'une ligne ferroviaire, comme illustré dans notre cas d'utilisation spécifique. Il démontre la robustesse de notre modèle en influençant indirectement le temps de travail. Au lieu de sélectionner aléatoirement des données pour chaque problème d'apprentissage, tel que les retards et les flux de passagers, nous pouvons identifier directement les connexions et les variables pertinentes pour des prédictions ciblées. De plus, les méthodes proposées dans le diagramme nous permettent de calculer les variables manquantes, telles que les retards, les intervalles et les flux de passagers, à partir des données brutes disponibles. Ce modèle constitue la base de la prochaine phase, l'approche hybride, qui combine l'apprentissage automatique.

La section suivante présente la phase 2 de notre méthodologie, illustrée dans la figure 3.2, qui détaille les étapes essentielles de prétraitement des données d'exploitation. Cette phase vise à garantir une base solide pour un apprentissage automatique efficace et des modèles prédictifs fiables. Les étapes décrites comprennent le nettoyage des données, la gestion des valeurs manquantes, la correction des incohérences et la transformation des variables pour les adapter aux exigences des algorithmes de machine learning.

3.3 Prétraitement des données

La phase 2 de notre méthodologie (3.2) se concentre sur le prétraitement des données, une phase indispensable où chaque jeu de données est structuré, nettoyé et préparé pour les analyses et l'apprentissage automatique. Ce processus garantit des bases solides pour obtenir des résultats fiables et exploitables. Nous avons utilisé notre diagramme de classes (3.6) pour illustrer les relations entre les variables et décrire la problématique spécifique des retards à étudier. Bien que ces jeux de données couvrent

en détail des informations telles que les horaires, les retards et d'autres aspects opérationnels, notre étude s'est focalisée sur le phénomène des retards pour développer un modèle adapté. Après avoir intégré les données brutes dans notre modèle relationnel, comme mentionné dans la section précédente, nous avons identifié les variables à calculer ou à enrichir à l'aide des fonctions décrites dans la figure 3.6.

Avant d'aborder la phase de modélisation par apprentissage automatique, un prétraitement rigoureux des données est indispensable pour garantir une base exploitable. Dans cette section, nous décrivons les étapes de prétraitement appliquées aux jeux de données, comme illustré dans la figure 3.7. Ces étapes préparent les données pour la phase suivante, où elles seront utilisées pour résoudre la problématique des retards et évaluer l'influence de chaque paramètre sur les retards quotidiens.



Figure 3.7 – Étapes de prétraitement des données pour l'analyse et l'apprentissage automatique

3.3.1 Nettoyage des données

La première phase de notre prétraitement des données montré dans la figure 3.7 est axée sur le nettoyage des erreurs et des problèmes survenus lors de la récupération des données. Ce processus comprend la correction des formats de date et d'heure afin d'assurer leur cohérence, ainsi que la vérification et la correction des identifiants mal écrits pour les stations et les rames. L'objectif est de garantir que les données utilisées dans nos analyses soient exemptes d'erreurs, ce qui est essentiel pour la fiabilité des résultats. Ce nettoyage minutieux permet de s'assurer que les données sont correctement formatées et prêtes à être utilisées dans les étapes suivantes du traitement et de la modélisation.

La figure 3.8 illustre le processus de nettoyage des données, qui se décompose en plusieurs étapes clés. Ce processus constitue une phase essentielle, réalisée sur les deux ensembles de données reçus. L'objectif est de garantir que chaque ligne et chaque

colonne de ces ensembles respecte un format cohérent. Cela concerne particulièrement les formats de date et les identifiants des stations et des rames. Pour les dates, nous veillons à ce que toutes les valeurs suivent le format standardisé « aaaa-mm-jj hh :mm :ss ». Ce format correspond à l'année (aaaa), suivie du mois (mm), puis du jour (jj), et est complété par l'heure (hh), les minutes (mm), et les secondes (ss). Cela permet d'assurer une uniformité des données temporelles, facilitant leur traitement et analyse. Par exemple, une date initialement formatée comme "2018-6-3 08 :05 :09" sera corrigée en "2018-06-03 08 :05 :09", afin de correspondre au format attendu.

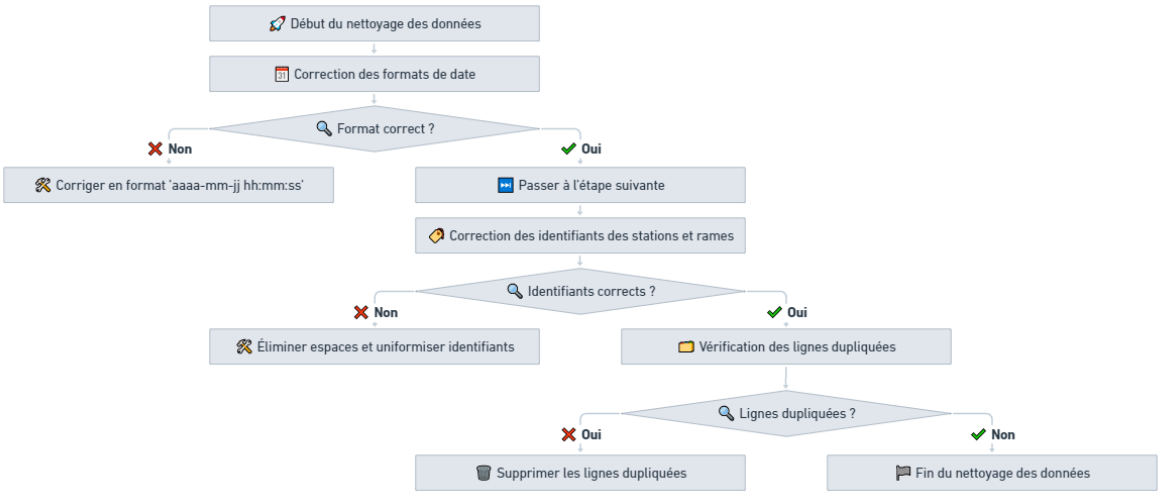


Figure 3.8 – Algorithme de nettoyage des données pour les formats de date et les identifiants

Concernant les identifiants des stations et des rames, nous effectuons un contrôle rigoureux afin de corriger les erreurs potentielles de saisie. Ces derniers peuvent inclure des espaces mal placés, des lettres manquantes ou incorrectes, ainsi que des fautes de frappe courantes. Par exemple, un identifiant saisi comme "Station A" sera corrigé en "StationA" pour éliminer les espaces inutiles et uniformiser les données. Ces étapes de nettoyage garantissent que les données sont exemptes de ces erreurs.

La duplication des lignes de données est un phénomène bien connu dans le domaine de la science des données. Dans notre cas, nous avons identifié deux types de duplication. La première duplication identifiée concerne l'heure de départ réelle d'une rame dans plusieurs stations. En effet, plusieurs stations ont la même heure de départ réelle pour la même rame, bien que les heures de départ prévues soient différentes, comme illustré par les cellules colorées en jaune dans le tableau 3.5. Par exemple, la rame numéro 12 est partie à la même heure réelle, soit 10 :09 :44, depuis trois stations différentes (stations s_1, s_2, et s_3), alors que les heures de départ prévues pour ces stations étaient respectivement 09 :35 :00, 09 :37 :12, et 09 :38 :51. Ce phénomène de duplication pose un problème d'interprétation, car il suggère que la rame était simultanément à plusieurs endroits au même moment, ce qui est évidemment impossible. Cette incohérence peut résulter d'une erreur dans les données ou d'une mauvaise captation des horaires. La solution proposée pour résoudre ce problème consiste à supprimer tout le trajet effectué par cette rame pendant cette période. Cela signifie qu'il faut retirer toutes les lignes de données associées à ce train, depuis la première station jusqu'au terminus, pour cette période spécifique. Ainsi, les données restantes seront plus cohérentes et exploitables pour une analyse correcte.

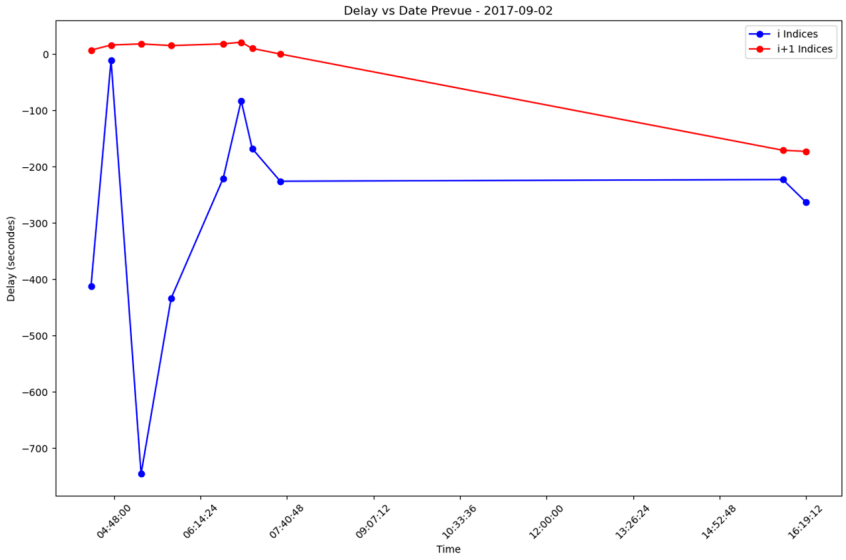


Figure 3.9 – Duplication des retards de la même rame par rapport à l'heure de départ prévue : comparaison des courbes bleues et rouges.

Le deuxième type de duplication concerne spécifiquement la première station de l'ensemble de données, en particulier lorsque le dépôt des rames est situé à proximité immédiate de la station de départ. Cela crée un problème pour une même rame, par exemple la rame Tn, qui semble effectuer son départ à plusieurs reprises depuis la première station s_1, qui est à la fois un garage et une station de départ. Cette duplication se manifeste par une différence d'une seconde entre les horaires de départ prévus et les horaires enregistrés comme motré dans le tableau 3.6. Cela génère des données en apparence contradictoires, où une même rame peut être marquée comme étant à la fois en avance et en retard. Ce phénomène est particulièrement fréquent le matin, entre 04h20 et 08h00, avec une fréquence élevée de duplications des départs. En revanche, l'après-midi, ce phénomène se produit avec une fréquence plus faible. Ce type de duplication peut compliquer l'analyse des données et fausser les résultats, nécessitant une attention particulière lors de la phase de nettoyage des données. Après une discussion avec les opérateurs du PCC, nous avons décidé de supprimer la première ligne de chaque duplication, représentée par la courbe bleue montrée dans la figure 3.9.

La partie suivante présente le travail réalisé sur la correction des incohérences temporelles identifiées dans un ensemble de données de type séries temporelles. Elle met l'accent sur les majeures erreurs rencontrés dans les données d'exploitation.

3.3.2 Correction des incohérences temporelles

La figure 3.10 présente la deuxième phase de notre prétraitement des données d'exploitation 3.7, qui consiste à corriger les incohérences temporelles susceptibles de survenir dans les données, comme c'est le cas dans notre étude. La récupération des données peut entraîner divers problèmes liés aux informations temporelles. Dans cette section, nous nous concentrerons sur la correction des incohérences temporelles liées aux départs des rames, afin de garantir la synchronisation de chaque trajet sur la ligne ferroviaire.

La correction des incohérences temporelles que nous avons effectuée se concentre

Table 3.5 – Exemple de duplication des heures de départ réelles pour une même rame à différentes stations.

Date prévue	Date réelle	Rame	Station
2017-09-01 08 :12 :19	2017-09-01 08 :18 :48	12	s_1
2017-09-01 09 :33 :00	2017-09-01 10 :09 :29	12	s_2
2017-09-01 09 :35 :00	2017-09-01 10 :09 :44	12	s_3
2017-09-01 09 :37 :12	2017-09-01 10 :09 :44	12	s_4
2017-09-01 09 :38 :51	2017-09-01 10 :09 :44	12	s_5
2017-09-01 09 :40 :19	2017-09-01 10 :09 :44	12	s_6
2017-09-01 09 :42 :00	2017-09-01 10 :09 :44	12	s_7
2017-09-01 09 :43 :00	2017-09-01 10 :09 :44	12	s_8
2017-09-01 09 :45 :22	2017-09-01 10 :09 :44	12	s_9
2017-09-01 09 :47 :00	2017-09-01 10 :09 :44	12	s_{10}
2017-09-01 09 :49 :00	2017-09-01 10 :09 :44	12	s_{11}
2017-09-01 09 :51 :00	2017-09-01 10 :09 :44	12	s_{12}
2017-09-01 09 :53 :00	2017-09-01 10 :09 :44	12	s_{13}
2017-09-01 09 :56 :00	2017-09-01 10 :09 :44	12	s_{14}
2017-09-01 09 :57 :00	2017-09-01 10 :09 :44	12	s_{15}
2017-09-01 09 :59 :00	2017-09-01 10 :09 :44	12	s_{16}
2017-09-01 10 :01 :26	2017-09-01 10 :09 :44	12	s_{17}
2017-09-01 10 :04 :15	2017-09-01 10 :09 :44	12	s_{18}
2017-09-01 10 :07 :00	2017-09-01 10 :09 :44	12	s_{19}
2017-09-01 10 :09 :00	2017-09-01 10 :10 :55	12	s_{20}

Table 3.6 – Exemple de duplication des départs réels pour une même rame à la première station.

Date prévue	Date réelle	Rame	Station	Retard (s)
2017-09-01 04 :25 :00	2017-09-01 04 :23 :22	t_{16}	s_1	-98
2017-09-01 04 :25 :01	2017-09-01 04 :25 :22	t_{16}	s_1	21
2017-09-01 04 :45 :00	2017-09-01 04 :27 :36	t_{16}	s_1	-1044
2017-09-01 04 :45 :01	2017-09-01 04 :45 :09	t_{16}	s_1	8
2017-09-01 05 :00 :00	2017-09-01 04 :56 :57	t_{16}	s_1	-183

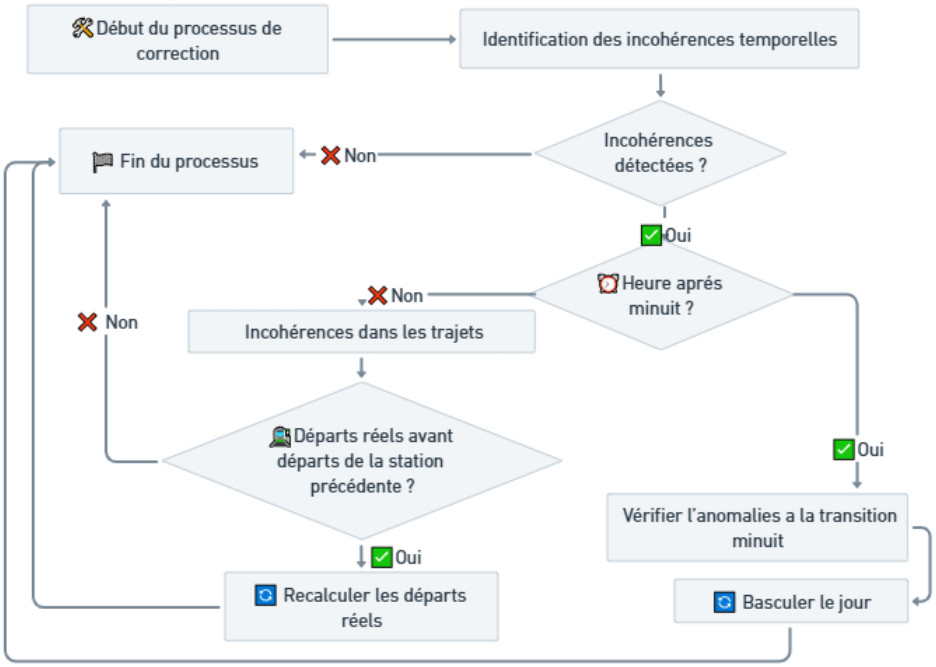


Figure 3.10 – Algorithme de correction des incohérences temporelles affectant les dates de départ des rames

sur deux problèmes majeurs liés aux données d'exploitation, qui sont dissimulés sous la gestion des valeurs manquantes. En effet, dans les données, nous n'avons pas de cases vides, mais nous rencontrons deux anomalies liées à des cases qui ne respectent pas le système. Le premier problème est une anomalie caractérisée par une avance croissante des horaires jusqu'à $23h$. Cette anomalie devient particulièrement évidente lors du passage de $23h$ à minuit, où une erreur de calcul provoque une incohérence dans les horaires. En effet, le système conserve la même date (jour) même après être passé à minuit, ce qui crée une confusion dans la gestion des jours. Cette mauvaise gestion de la transition entre les jours, spécifiquement à l'heure de minuit, perturbe les calculs ultérieurs et peut fausser les analyses basées sur les horaires. le tableau 3.7 illustre une anomalie temporelle liée au passage de $23h$ à minuit, où le système conserve incorrectement le même jour après minuit, entraînant des retards négatifs importants. Par exemple, pour la station s_3, la date prévue est $pt_{t_3,s_3} = "02/01/2017 à 23 : 55"$, et la date réelle est $rt_{t_3,s_3} = "00 : 00"$ le lendemain, ce qui produit un retard erroné de "-86146" secondes. Ce problème se répète pour d'autres enregistrements.

La figure 3.11 présente la deuxième anomalie détectée, illustrée par l'exemple du trajet de la rame t_{16} entre certaines stations, caractérisé par une incohérence notable. En effet, dans certains cas, l'heure de départ réelle de la rame depuis la station s_{i+1} se produit avant son départ réel de la station précédente s_i. Où les points violets représentent les heures de départ prévues et les croix (bleues ou jaunes) représentent les heures de départ réelles. Cette incohérence temporelles se répète fréquemment dans les deux ensembles des données, affectant l'ensemble des rames sur la ligne . Par exemple, dans le figure 3.11, on peut observer que plusieurs stations montrent des départs réels identiques, malgré des heures de départ prévues différentes. Cela donne l'impression que le train quitte simultanément plusieurs stations ou même qu'il est déjà parti de la station suivante avant d'avoir quitté la précédente. Ces incohérences

Table 3.7 – Exemple de données d'exploitation avec incohérences temporelles (passage de 23h à minuit)

Date prévue	Date réelle	Rame	station	Retard (s)
02-01-2017 23 :55	02-01-2017 00 :00	t_3	s_3	-86146
02-01-2017 23 :57	02-01-2017 00 :00	t_8	s_4	-86161
02-01-2017 23 :58	02-01-2017 00 :02	t_5	s_7	-86161
02-01-2017 23 :58	02-01-2017 00 :01	t_9	s_8	-86285
02-01-2017 23 :59	02-01-2017 00 :01	t_2	s_9	-86260
02-01-2017 23 :59	02-01-2017 00 :03	t_{15}	s_{10}	-86200

apparaissent surtout le matin, entre $"04h20"$ et $"08h00"$, et sont moins fréquentes l'après-midi.

L'algorithme 3.10 présente les deux solutions apportées pour corriger les anomalies détectées liées à l'incohérence temporelle des heures de départ réelles des rames. La première anomalie, liée à minuit, consiste à basculer le jour dès que l'heure dépasse minuit, comme indiqué dans l'équation 3.5, et illustré dans le tableau 3.8. Cette correction permet d'assurer la cohérence temporelle et de rectifier les retards dans les analyses futures.

La solution proposée pour corriger l'anomalie liée consiste à recalculer les départs réels plutôt qu'à supprimer simplement ces itinéraires absurdes. L'équation 3.6 utilise une moyenne pondérée des retards observés dans les stations précédentes pour estimer un départ réaliste à chaque station. La formule présentée permet ainsi de recalculer l'heure de départ réelle à partir des retards observés dans les stations précédentes et des heures de départ prévues. En ajustant les données de cette manière, il devient possible de corriger les incohérences tout en préservant la cohérence temporelle des trajets des rames. Après l'application de notre équation 3.6, la figure 3.12 illustre une nette amélioration de la cohérence entre les heures de départ prévues et réelles. Contrairement à la situation initiale où des incohérences notables étaient observées, la correction appliquée a permis de réaligner ces horaires, garantissant ainsi une synchronisation plus fidèle des départs. Cette approche démontre que les heures de départ sont devenues cohérentes après l'application de notre solution, évitant ainsi la nécessité de supprimer des lignes de données entières. Cela permet de préserver un maximum d'informations tout en assurant la fiabilité des données, ce qui est essentiel pour des analyses et des prévisions précises.

$$rt'_{t_i,s_j} = rt_{t_i,s_j} + 1\,\text{jour} \tag{3.5}$$

$$rt'_{t_i,s_j+1} = \frac{1}{i}\sum_{k=1}^{i}(y_{t_i,s_k} + pt_{t_i,s_j+1}) \tag{3.6}$$

Où rt_{t_i,s_j} représente la date initiale , et rt'_{t_i,s_j} est la date corrigée.

La partie suivante présente le travail réalisé sur la gestion des valeurs aberrantes dans un ensemble de données d'exploitation. L'accent est mis sur les propositions des

Table 3.8 – Même exemple de le tableau 3.7 après correction des incohérences temporelles (passage de 23h à minuit)

Date prévue	Date réelle	Rame	station	Retard (s)
02-01-2017 23 :55	03-01-2017 00 :00	t_3	s_3	300
02-01-2017 23 :57	03-01-2017 00 :00	t_8	s_4	180
02-01-2017 23 :58	03-01-2017 00 :02	t_5	s_7	122
02-01-2017 23 :58	03-01-2017 00 :01	t_9	s_8	121
02-01-2017 23 :59	03-01-2017 00 :01	t_2	s_9	61
02-01-2017 23 :59	03-01-2017 00 :03	t_{15}	s_{10}	63

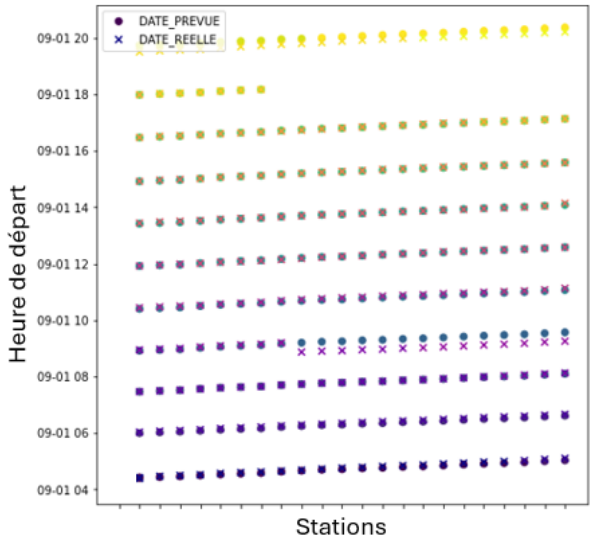


Figure 3.11 – Incohérence des heures de départ : comparaison des heures de départ prévues et réelles pour chaque station.

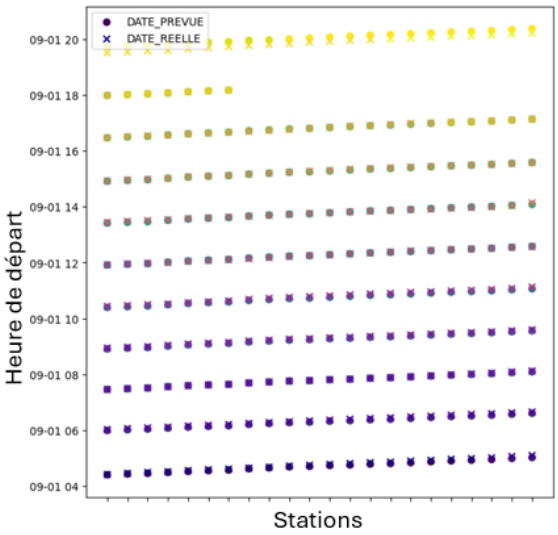


Figure 3.12 – Évolution des heures de départ prévues et réelles par station : après correction des incohérences

opérateurs pour gérer ces valeurs.

3.3.3 Gestion des valeurs aberrantes

La troisième phase, décrite dans la figure 3.7, concerne le traitement des valeurs aberrantes présentes dans les données d'exploitation. Ce phénomène est bien connu dans le domaine de la science des données et survient lorsque certaines valeurs observées s'écartent fortement de ce qui est attendu, ne suivant pas la logique du système étudié. Dans notre cas, ces valeurs aberrantes se manifestent sous la forme de retards des rames qui ne correspondent pas au comportement normal du réseau ferroviaire. Ces anomalies peuvent résulter de défaillances techniques, d'erreurs de saisie ou de circonstances exceptionnelles et doivent être traitées afin de garantir la qualité des prédictions et des analyses basées sur ces données.

D'après les experts exploitation ferroviaire d'Egis-Rail, l'écart entre l'heure de départ réelle et l'heure prévue pour une rame dans une station est limité à un maximum de 900 secondes (15 minutes) pour les retards et de 360 secondes (6 minutes) pour

les avances. Toutefois, certaines rames peuvent, au cours de la journée, dépasser ces limites. Ainsi, dans notre étude, l'objectif est d'analyser les retards quotidiens en se concentrant sur les situations habituelles, plutôt que sur les cas exceptionnels.

Ce phénomène est apparu dans notre étude de cas avec des retards présentant des valeurs aberrantes, atteignant parfois 1h30, et des avances dépassant 10 minutes, comme illustré dans la figure 3.13 avec une moyenne de 5 min par rame. Ces écarts sont illogiques dans le cadre du fonctionnement normal. Après discussion avec les opérateurs, il a été conclu que ces anomalies résultaient d'une mauvaise récupération des données. Pour résoudre ce problème, nous avons mis en place une gestion des valeurs aberrantes, en traitant spécifiquement l'intervalle [-360 : 900].

Une analyse a été réalisée sur le sujet à l'aide de deux méthodes pour garantir la validité des propositions des opérateurs. La première méthode utilise les quartiles. Nous avons tracé le 1er quartile (Q1), la médiane (Q2), et le 3e quartile (Q3) des retards quotidiens. Les graphiques par la suite montrent une dispersion des retards. Ces quartiles nous permettent de définir les bornes de normalité des retards et d'identifier les jours où des valeurs aberrantes apparaissent, c'est-à-dire les jours où les retards s'éloignent significativement de la distribution habituelle.

Les résultats de cette étude des quartiles ne sont d'ailleurs pas très éloignés des décisions proposées par les opérateurs, qui ont défini un intervalle de normalité entre [-360, 900] secondes pour les retards. Cet intervalle correspond bien aux bornes identifiées par les quartiles, renforçant ainsi la validité des décisions prises sur la base des analyses opératoires. Les jours où les retards dépassant cet intervalle sont considérés comme aberrants et nécessitent une attention particulière.

En complément de cette approche, nous avons appliqué la méthode de l'Isolation Forest, une technique d'apprentissage automatique non supervisé conçue pour détecter les anomalies dans des ensembles de données complexes [104]. Cette méthode identifie les valeurs aberrantes en isolant les observations les plus éloignées des autres. Les valeurs normales sont représentées en bleu et les valeurs aberrantes en rouge par la suite. Les résultats montrent que l'Isolation Forest a détecté plusieurs anomalies éparpillées au-delà de la plage des retards typiques. Cela correspond à des événements inhabituels ou des défaillances du système.

Après discussion avec les opérateurs et validation des propositions à l'aide d'analyses scientifiques, La solution adoptée dans la figure 3.14 consiste à remplacer les valeurs aberrantes par des valeurs comprises dans l'intervalle défini par les opérateurs et validé par l'analyse effectuée à ce sujet. Finalement, l'intervalle retenu est compris entre -360 secondes pour les avances et 1000 secondes pour les retards au lieu de[-360 :900] sur la base des résultats de de l'Isolation Forest. Ainsi, toutes les valeurs dépassant ces bornes sont considérées comme aberrantes et sont remplacées par les valeurs limites correspondantes. Concrètement, les retards supérieurs à 1000 secondes sont ajustés à 1000, tandis que les avances inférieures à -360 secondes sont corrigées à -360. Cette approche a été appliquée aux deux ensembles de données que nous avons reçus, garantissant une cohérence dans le traitement des retards tout en respectant les seuils définis. Cela permet d'éliminer les anomalies extrêmes sans pour autant distordre l'analyse globale des retards, assurant ainsi une meilleure qualité des données pour les futures analyses et modèles prédictifs.

La partie suivante présente le travail réalisé sur l'enrichissement des données d'exploitation à l'aide de données contextuelles.

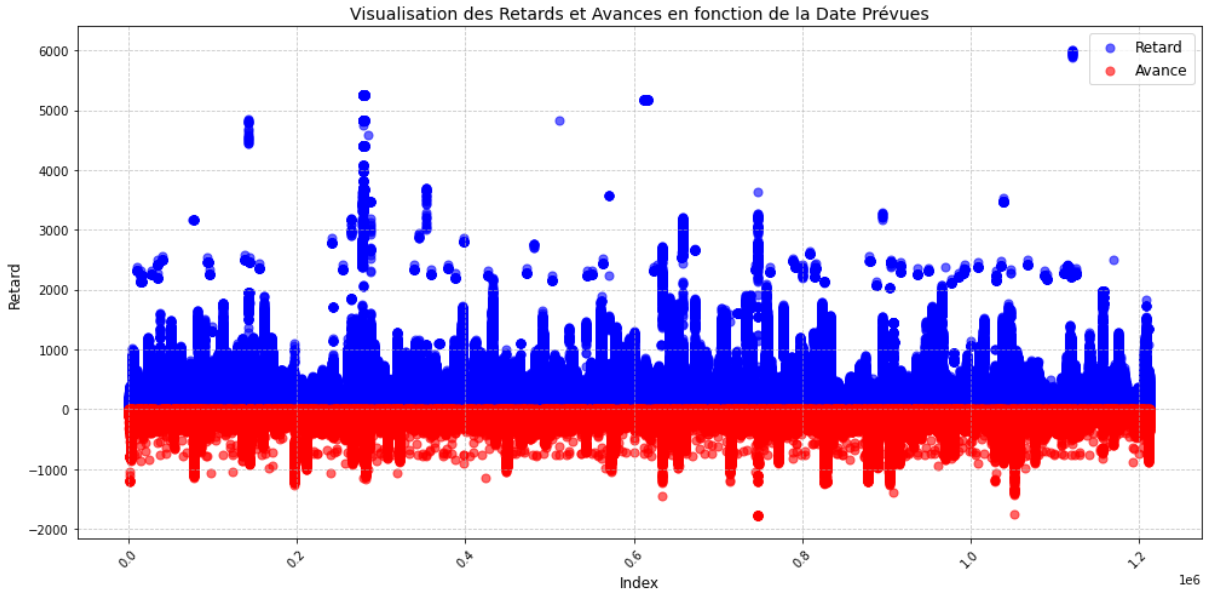


Figure 3.13 – Example d'une distribution des retards observés sur trois périodes annuelles : 2017, 2018 et 2019

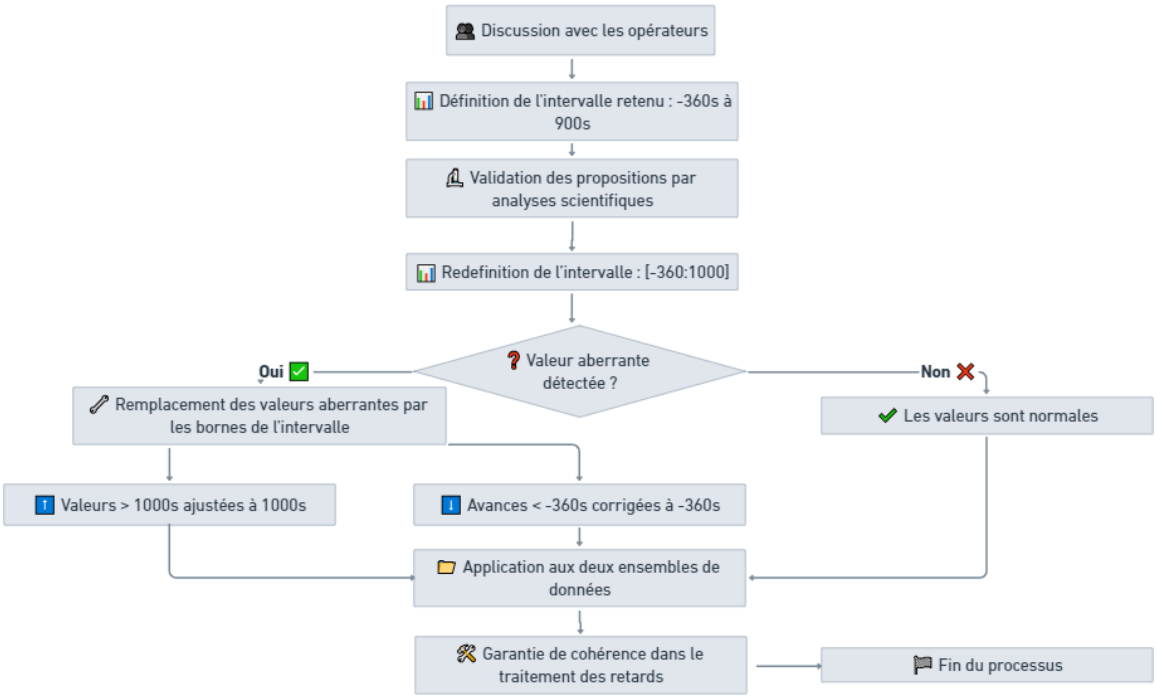


Figure 3.14 – Processus de validation et correction des valeurs aberrantes dans les retards des rames

3.3.4 Enrichissement des données

L'enrichissement des données consiste à ajouter des informations supplémentaires à un ensemble de données existant pour le rendre plus complet et utile. Dans notre contexte, nous travaillons avec des données brutes issues de l'exploitation des lignes de Tram et de RER. Ces données brutes peuvent inclure des informations de base telles que les horaires de départ et d'arrivée des rames figure 4.4, mais elles ne contiennent pas toutes les informations nécessaires pour une analyse approfondie, comme les intervalles de temps prévus et réels entre les départs ou les retards observés.

La figure 3.15 présente la quatrième phase de notre processus de prétraitement, illustrée dans la figure 3.7, qui consiste à enrichir les données en ajoutant de nouvelles colonnes extraites des données brutes d'exploitation. Ces nouvelles informations fournissent des détails supplémentaires sur chaque trajet effectué par une rame au cours d'une journée. Par exemple, les intervalles de temps prévus et réels indiquent le temps qui devrait s'écouler entre le départ de deux rames consécutives (intervalle prévu) et le temps réellement observé (intervalle réel). Ces informations peuvent être calculées à partir des heures de départ des rames dans les données brutes avec l'équation 3.1. De même, un retard est calculé en comparant l'heure de départ prévue et l'heure de départ réelle d'une rame avec l'équation 3.3. Nous avons également calculé le trajet effectué par une rame entre la station s_i et s_{i+1} (trajet suivant)3.2.

Le web scraping [107] est une technique utilisée pour extraire des données à partir de sites web. Dans notre projet, nous avons utilisés cette technique pour obtenir des informations météorologiques liées à chaque jour de notre ensemble de données, comme montré dans la figure 3.15. Par exemple, en accédant à un site de météo, nous pouvons automatiser l'extraction des conditions météorologiques, comme la température, pour chaque journée de notre période d'étude. Ces informations peuvent ensuite être ajoutées à notre ensemble de données enrichi pour analyser l'impact de la météo sur les retards ou d'autres aspects de l'exploitation des lignes de transport.

La partie suivante présente le travail réalisé sur la mise en forme et l'encodage des variables pour l'entraînement des modèles d'apprentissage, afin de garantir une meilleure convergence. L'accent est mis sur les deux types de données : qualitatives et quantitatives.

3.3.5 Mise en forme et encodage des variables pour l'entraînement

Les données sont généralement divisées en deux types : quantitatives et qualitatives. Le premier type, les données quantitatives, représente des informations mesurables et numériques, tandis que les données qualitatives concernent des catégories ou des qualités qui ne peuvent pas être mesurées numériquement, mais qui sont descriptives. Les données quantitatives sont souvent utilisées pour exprimer des quantités ou des valeurs numériques (par exemple, le retard d'une rame, la température ou la distance parcourue), tandis que les données qualitatives peuvent être nominales ou ordinales, comme le type de jour ou les heures de pointe et heures creuses.

La figure 3.7 présente la cinquième phase de notre processus, qui est le principe de la normalisation et de l'encodage des données quantitatives et qualitatives, deux étapes essentielles pour garantir le bon fonctionnement de notre système lors de la phase d'entraînement. Ces mises en formes sont cruciales pour l'entraînement des modèles d'apprentissage automatique, car ces derniers n'acceptent pas les variables qualitatives sous forme de chaînes de caractères. C'est pourquoi nous cherchons à encoder ces

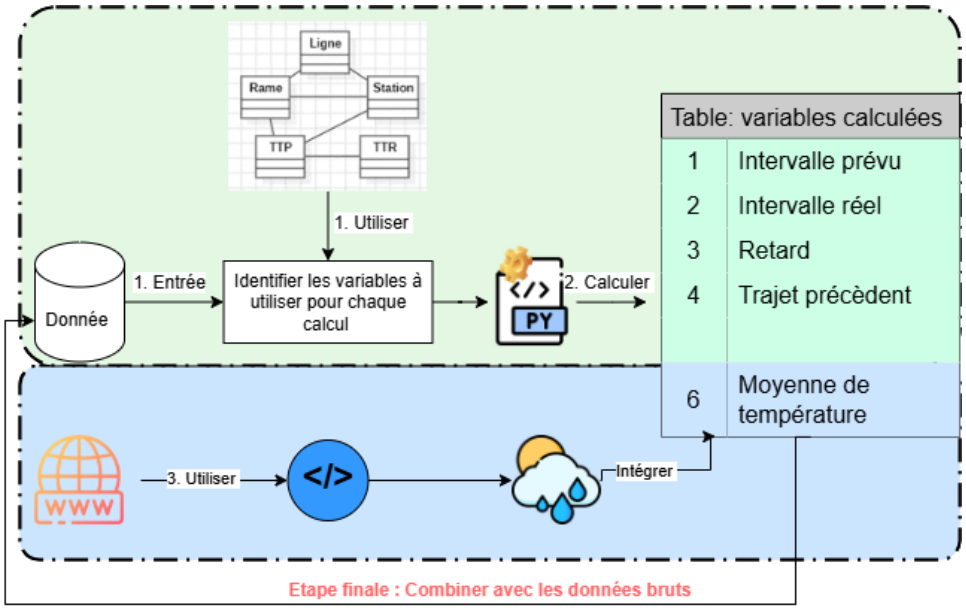


Figure 3.15 – Processus d'enrichissement de la base de données avec des variables calculées et du Web Scraping pour l'analyse des retards ferroviaires

variables en une forme numérique compréhensible par le modèle. La normalisation des données quantitatives permet d'ajuster les valeurs à une échelle commune, ce qui évite que certaines caractéristiques ne dominent le modèle simplement en raison de leur amplitude. Cela améliore la convergence des algorithmes et rend l'apprentissage plus efficace. Le codage des données qualitatives est essentiel pour transformer les catégories non numériques en une forme compréhensible par les algorithmes d'apprentissage automatique.

La normalisation Min-Max est donnée par la formule suivante :

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{3.7}$$

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{3.8}$$

où :

— X est la valeur à normaliser.

— $X_{\min}$ est la valeur minimale dans l'ensemble des données.

— $X_{\max}$ est la valeur maximale dans l'ensemble des données.

Cette méthode permet de transformer les valeurs de X pour qu'elles se situent dans l'intervalle $[0, 1]$.

La figure 3.16 présente le processus de normalisation et d'encodage des données d'exploitation, permettant de catégoriser les données en deux groupes : quantitatifs et qualitatifs. Dans les données d'exploitation disponibles, les variables quantitatives incluent des informations telles que les retards, les intervalles prévus et réels, qui sont calculés à l'aide de l'équation 3.1. Ces variables quantitatives incluent également des valeurs associées aux temps d'attente dans les stations et aux temps inter-stations. Afin d'éviter que des échelles de valeurs différentes ne biaisent les modèles d'apprentissage, il est essentiel de normaliser ces variables. La normalisation permet de ramener les valeurs à une échelle commune, évitant ainsi que certaines caractéristiques ne dominent

le modèle en raison de leur amplitude. À cette fin, nous avons appliqué la fonction de normalisation décrite dans l'équation 3.8, qui ramène toutes les valeurs dans une plage commune (typiquement entre 0 et 1), garantissant ainsi que chaque variable contribue de manière équitable à l'apprentissage du modèle.

D'autre part, les données qualitatives d'exploitation, telles que les noms des stations, les identifiants des rames, le jour de la semaine, la voie, les heures de départ, et le terminus, nécessitent un traitement différent. Étant donné que ces informations sont des catégories non numériques, elles doivent être encodées avant d'être utilisées dans les algorithmes d'apprentissage. Pour cela, nous avons utilisé la méthode "One-Hot Encoding" [108]. Cette technique consiste à transformer chaque catégorie en une série de variables binaires (0 ou 1), où chaque variable représente la présence ou l'absence d'une catégorie spécifique. Par exemple, pour les jours de la semaine, nous avons utilisé un encodage spécifique de manière à créer 7 variables distinctes, chacune correspondant à un jour particulier. Ainsi, lundi est encodé en "0", mardi en "1", mercredi en "2", et ainsi de suite jusqu'à dimanche. Chaque jour est représenté par un "1" dans sa colonne respective et des "0" dans les autres colonnes. Cet encodage permet au modèle de machine learning de traiter efficacement ces informations qualitatives tout en respectant leur nature catégorielle.Cette méthode permet aux algorithmes de machine learning de comprendre et de traiter correctement les variables catégorielles. Ainsi, en combinant la normalisation des variables quantitatives et l'encodage des variables qualitatives, nous avons pu préparer efficacement nos données pour l'apprentissage automatique, assurant ainsi des résultats plus robustes et précis comme montré dans la figure 3.16.

Le prétraitement des données d'exploitation est une phase essentielle pour garantir la qualité des données, et il joue un rôle très important dans la compréhension des données. Dans cette section, nous abordons les problèmes rencontrés lors de la récupération des données, que nous regroupons dans la figure 3.7. Celle-ci illustre les erreurs liées aux formats des variables, comme les formats date-heure, qui posent les problèmes les plus courants dans ce domaine. Un format incorrect des données temporelles peut entraîner la duplication des lignes, un manque de données dû à une incohérence temporelle, ainsi qu'une mauvaise gestion des valeurs aberrantes. Par exemple, une date mal saisie peut impacter le calcul des retards, provoquant parfois des erreurs allant jusqu'à une heure dans certains cas. Bien qu'il existe d'autres problèmes dans les bases de données d'exploitation, ceux liés aux heures sont les plus critiques, car ils peuvent sérieusement affecter les résultats finaux. C'est pourquoi nous avons accordé une attention particulière à ce sujet.

Dans la section suivante 3.4, nous décrivons les étapes essentielles pour analyser les données d'exploitation afin de bien comprendre la nature et l'influence entre ces données avant de passer à la phase d'apprentissage automatique

3.4 Analyse des données

L'analyse des données vise à fournir des résumés synthétiques, souvent sous forme graphique, des séries de valeurs, adaptées à leur nature (qualitatives ou quantitatives) et observées sur une population ou un échantillon [109]. La population représente l'ensemble complet des éléments ou individus étudiés, tandis que l'échantillon est une sous-partie de cette population, sélectionnée pour représenter ses caractéristiques de manière plus pratique et efficace. Si l'on analyse les passages des rames, l'échantillon est limité aux passages des rames entre $"08:00:00"$ et $"09:00:00"$ (heure de pointe).

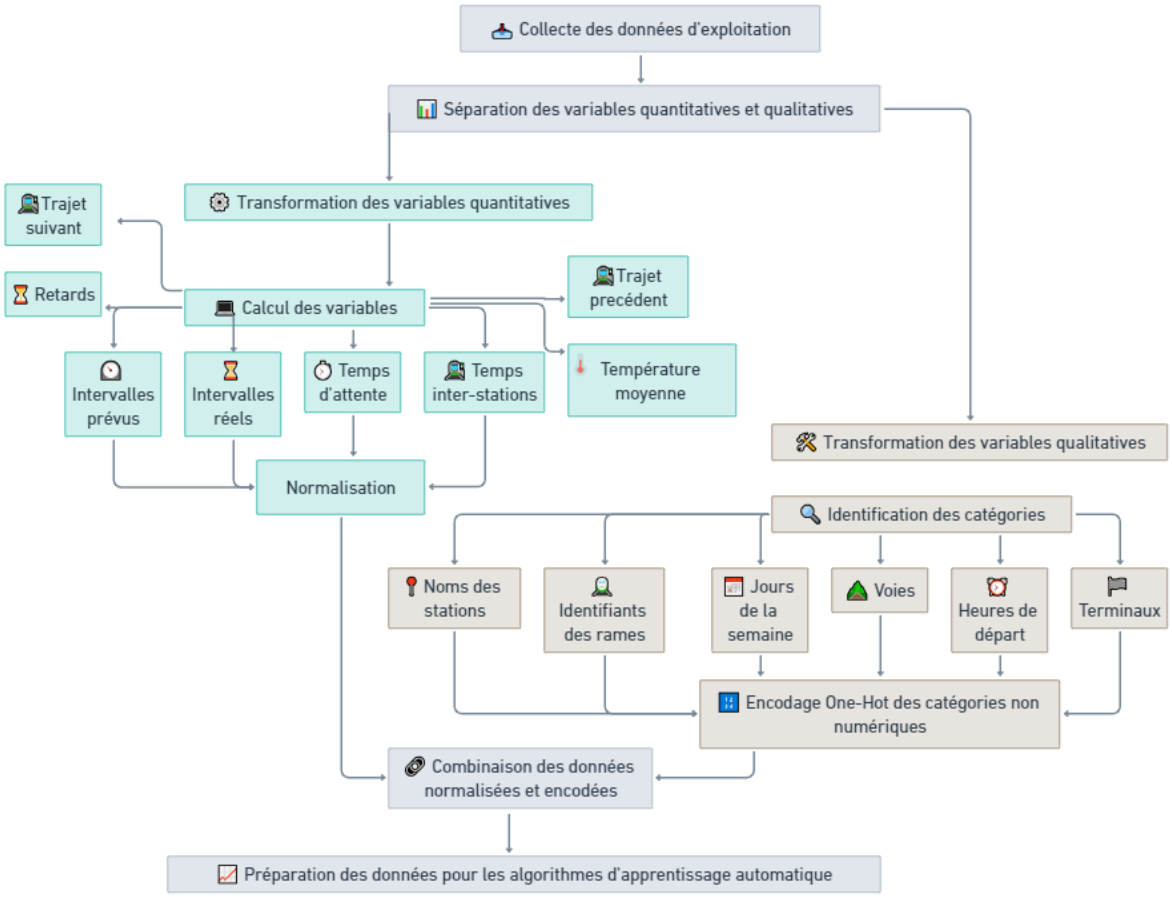


Figure 3.16 – Processus de la transformation des données d'exploitation quantitatives et qualitatives pour l'apprentissage automatique

Ce type d'analyse a pour objectif de déterminer les tendances des données et de traiter les données brutes afin d'en extraire des informations nouvelles, mettant en évidence les aspects les plus pertinents pour leur intégration dans un modèle d'apprentissage. Il s'agit également de comprendre la nature des données en fonction des variables présentes dans l'ensemble du jeu de données, afin d'évaluer l'impact de chaque variable sur l'ensemble des données. D'autre part, l'analyse des données vise aussi à détecter les anomalies au sein d'un échantillon, afin d'identifier les données non représentatives ou celles présentant des cas exceptionnels qui n'apportent aucune valeur ajoutée pour le reste du travail.

La figure 3.17 présente l'étape d'analyse des données appliquée dans le cadre de l'analyse des données d'exploitation, où elle permet de révéler les tendances présentes dans les données. Cette analyse joue un rôle essentiel pour comprendre l'influence des retards. Chaque ensemble de données nécessite un processus structuré d'étapes pour analyser ces tendances. Il est impératif d'étudier les multiples facettes de l'analyse sur les deux types de données et de définir les relations entre les variables présentées dans les données d'exploitation. Cela inclut une analyse approfondie de l'importance de ces variables pour les modèles d'apprentissage automatique utilisés.

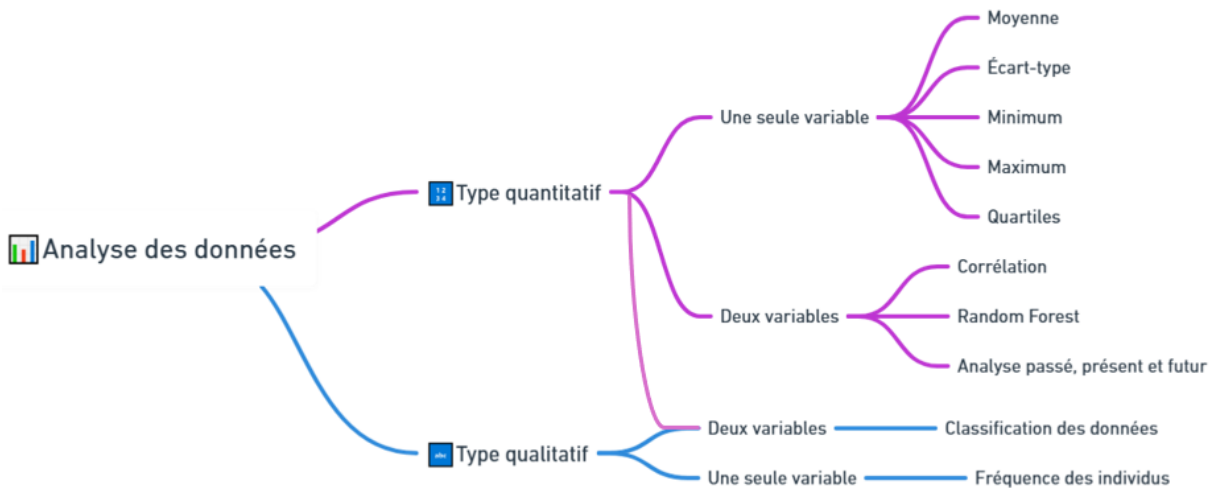


Figure 3.17 – Classification des types et méthodes d'analyse des données.

Tout d'abord, chaque type de données (qualitatives ou quantitatives) nécessite des analyses spécifiques, qui diffèrent selon qu'on examine une seule variable isolée ou qu'on cherche à établir des relations entre deux variables ou plus, qu'elles soient du même type ou de types différents. Dans le cas d'une seule variable, les mesures les plus courantes sont la médiane, les quantiles, la moyenne, la fréquence, la variance, et l'écart-type. Ces indicateurs sont souvent associés à des représentations graphiques telles que les diagrammes en bâtons, les histogrammes, les boîtes à moustaches, les graphiques cumulatifs.

Lorsqu'il s'agit de deux variables, l'analyse se concentre sur la corrélation et le coefficient de corrélation [110], ainsi que sur l'importance des variables en utilisant des méthodes telles que les forêts aléatoires (random forests). Ces concepts sont généralement accompagnés de graphiques tels que le nuage de points (scatterplot), les diagrammes de boîtes, ou encore les graphiques de corrélation.

L'étude simultanée de deux variables quantitatives X et Y a pour objectif principal de mettre en évidence une éventuelle variation conjointe des deux variables, que

l'on qualifie de liaison. Dans certains cas, cette liaison peut être interprétée a priori comme causale, où la variable X expliquerait la variable Y. Dans d'autres situations, cette relation n'est pas causale et les deux variables jouent des rôles symétriques. En pratique, il est important de bien distinguer ces deux cas, car une liaison entre deux variables ne signifie pas nécessairement qu'il existe une relation de cause à effet. C'est ainsi que sont introduites les notions de covariance, de coefficient de corrélation linéaire [110].

$$\text{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \tag{3.9}$$

$$Corr(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3.10}$$

Ou

- X_i : Valeur de la variable X pour l'observation i.
- Y_i : Valeur de la variable Y pour l'observation i.
- $\overline{X}$: Moyenne de la variable X.
- $\overline{Y}$: Moyenne de la variable Y.
- n : Nombre d'observations dans l'échantillon.
- σ_X : Écart-type de la variable X.
- σ_Y : Écart-type de la variable Y.

Une méthode pour représenter les données dans le cadre de l'étude simultanée d'une variable quantitative et d'une variable qualitative consiste à utiliser des boîtes parallèles. Sur un même graphique avec une échelle unique, on représente pour la variable Y un diagramme en boîte pour chacune des sous-populations définies par la variable X. La comparaison de ces boîtes permet d'obtenir une idée claire de l'influence de X sur les valeurs de Y, c'est-à-dire de la relation entre les deux variables. Et aussi il est nécessaire de classifier les données en sous-groupes pour une analyse plus détaillée, en les classant par exemple par type de jour, par période de l'année, ou par type des rames. Cela permet de détecter des anomalies, c'est-à-dire des points de données qui diffèrent significativement des autres et qui pourraient indiquer des erreurs ou des phénomènes inhabituels.

Étant donné que les données d'exploitation que nous étudions sont de type séries temporelles, chaque jour étant une nouvelle observation, une analyse des tendances intrinsèques est nécessaire pour relier le passé, le présent et le futur. Cela permet d'évaluer l'impact du comportement des rames dans le passé, ainsi que celui de leurs prédécesseurs et successeurs, sur les performances des rames actuelles et futures. Dans cette optique, nous intégrerons ces tendances passées comme variables d'entrée (features) dans nos modèles d'apprentissage. L'inclusion de ces données historiques enrichit le modèle en lui fournissant une base d'informations complète et diversifiée, ce qui est essentiel pour analyser et prédire les retards avec une plus grande précision. Ce processus permet d'exploiter les relations temporelles observées dans les données afin de renforcer la capacité prédictive du modèle.

La section suivante présente l'approche de conception des caractéristiques, qui repose sur une préparation rigoureuse des données d'entrée, basées sur des données soigneusement nettoyées. Cette démarche tient compte des principes "le passé impacte le présent" et "le présent impacte le futur" dans le cadre du trajet d'une rame. En structurant les données selon ces concepts, l'analyse permet d'identifier les facteurs influençant le trajet en fonction des événements passés tout en anticipant

les effets potentiels sur les étapes futures. Cette approche de conception garantit que les caractéristiques des données soient adaptées aux besoins des modèles prédictifs, contribuant ainsi à une meilleure précision des prévisions.

3.5 Ingénierie des caractéristiques temporelles

L'ingénierie des caractéristiques temporelles constitue une étape essentielle de notre méthodologie, comme présenté dans la figure 3.2 sous l'étape 3. Cette phase repose sur le principe selon lequel le passé influence le présent, qui à son tour impacte le futur. Le choix d'une approche basée sur un horizon à court terme s'est imposé en raison de nos données. En effet, les corrélations entre les variables explicatives et la variable cible, à savoir le retard, se sont révélées relativement faibles lors de notre analyse initiale. Ce faible niveau de corrélation constitue un défi majeur pour le développement d'un modèle performant. Des exemples concrets illustrant ces enjeux sont détaillés dans le chapitre 4, afin de mettre en évidence les limites et les solutions proposées pour surmonter ce problème.

La manière dont les données sont préparées joue un rôle crucial dans la performance finale du modèle. La phase de préparation des données, qui inclut le nettoyage, la transformation et l'enrichissement des informations, vise à optimiser la qualité des prédictions. Nous postulons que cette méthodologie, si elle est bien exécutée, pourrait être généralisée à d'autres types de données de transport ayant des caractéristiques similaires. Cependant, pour valider cette hypothèse, nous avons effectué une série de tests rigoureux, également présentés dans le chapitre 4, afin d'évaluer l'efficacité de notre approche et de vérifier si elle améliore réellement les performances du modèle. Ces tests permettent non seulement de confirmer ou infirmer notre hypothèse, mais aussi de mettre en lumière les ajustements potentiels qui pourraient encore améliorer la prédiction des retards dans des conditions opérationnelles complexes.

Cette section présente une préparation des données d'entrée basée sur des données soigneusement nettoyées, en tenant compte des concepts de "le passé impacte le présent" et "le présent impacte le futur" pour le trajet d'une rame, comme illustré à la figure 3.3. Ainsi, les données d'entrée du modèle d'apprentissage sont préparées en intégrant les trois notions temporelles comme caractéristiques (features) du modèle : le passé du trajet, le présent qui représente instantanément le passage d'une rame, et le futur proche de cette même rame.

3.5.1 Structuration temporelle : impact du passé et projection vers le futur

Les données d'exploitation, après avoir subi deux phases de prétraitement et d'analyse, nécessitent une sélection et une préparation spécifiques avant d'être intégrées dans un modèle d'apprentissage. L'analyse des données a démontré l'importance du comportement des rames dans le passé sur le présent, ainsi que sur le futur proche, grâce aux analyses de corrélation effectuées dans la section précédente. Approuvé par les OCC, il apparaît que le facteur déterminant influençant le retard d'une rame repose sur son comportement passé, ainsi que sur celui de ses prédécesseurs et successeurs. Cette observation permet de préparer les données d'entrée pour nos modèles d'apprentissage, en appliquant le principe selon lequel *"le passé impacte le*

présent, et le présent influence le futur".

Les opérateurs ont répondu à la première question en estimant qu'un horizon à court terme est suffisant pour permettre une réaction adéquate avec le système proposé, offrant ainsi un laps de temps raisonnable pour que les prédictions soient prises en compte avant que le retard réel ne survienne. En ce qui concerne la fenêtre du passé, cette question a été traitée à travers les tests détaillés dans le chapitre 4, où plusieurs analyses ont été menées afin d'identifier la fenêtre temporelle optimale pour chaque ensemble de données.

Pour intégrer les notions de passé et de futur avec l'ensemble des données d'entrée décrites dans la figure 3.16, nous avons développé des algorithmes permettant de capturer les variables de retard issues du passé et de les intégrer comme données d'entrée dans nos modèles d'apprentissage. De plus, ces algorithmes fournissent une estimation de la station où se trouvera la rame dans un futur proche, noté f.

La figure 3.3 illustre le concept de base de l'horizon de prédiction f pour une rame donnée, noté t_i, où i appartient à l'ensemble des rames circulant pendant la journée. L'idée est d'utiliser l'historique des rames précédant et suivant t_i, ainsi que les horaires de passage de cette rame aux stations précédentes (indiquées en gris). L'objectif est d'utiliser le comportement passé des rames et son potentiel sur le comportement futur proche f d'une rame t_i. En adoptant cette approche, le modèle vise à fournir au gestionnaire des opérations ferroviaires des informations prédictives avant l'arrivée d'une rame t_i à la prochaine station S_{j+e}, où e est une estimation de la prochaine station dans l'avenir de la rame t_i. En intégrant des données historiques et en s'appuyant sur des schémas temporels complexes. Cela donne aux opérateurs des informations critiques à l'avance. Prenons le moment présent, marqué par le départ de t_5 de la station S_5 à $16h07$. Nous examinons les heures de départ réelles des rames ayant précédé et suivi t_5 au cours de la dernière heure. En comparant l'heure actuelle $t = 16h07$ avec l'heure planifiée (TTP) de t_5, nous estimons que l'heure d'arrivée de t_5 à la station dans le futur proche estimé par exemple à $f = 15min\ t + 15 = 16h22$, ce qui signifie que notre t_5 devrait arriver à la station S_{12} dans les 15 minutes.

Pour intégrer les trois notions de passé, présent et futur dans nos modèles d'apprentissage, nous avons préparé trois ensembles d'entrées distincts. Le premier ensemble de données dans les deux sections suivantes, noté XOP, représente la notion de présent et regroupe toutes les caractéristiques opérationnelles, comme illustré dans la figure 3.16, et détaillé dans l'ensemble de données 3.2.4. Il reste à déterminer les éléments du passé et du futur. La seconde notion, correspondant au passé, est notée Y^{past}. Cet ensemble de données capture les retards passés d'une rame, ainsi que les informations concernant ses prédécesseurs et successeurs sur la ligne. Enfin, la notion de futur est représentée par l'ensemble XOP^{+f}, qui contient les informations liées à l'évolution future d'une rame. Cet ensemble inclut la station où la rame se trouvera dans un futur proche f, ainsi que l'heure de départ prévue pour cette rame.

3.5.2 Données d'entrée : Comportement historique des rames dans une fenêtre temporelle Y^{past}

Le principe des retards précédents repose sur la collecte et l'utilisation des données de retard historiques provenant de XOP, comme illustré dans la figure 3.3, afin d'anticiper les retards futurs d'une rame. L'idée centrale est de créer un sous ensemble de données spécifique XOP pour chaque rame t_i, appelé retards précédents de la rame

(Y^{past}), qui consolide les retards historiques de la rame t_i dans toutes les stations, en prenant en compte ses prédécesseurs et successeurs sur une période d'une heure dans le passé, noté $y^{\text{past}}_{t_i,s_j,rt_{t_i,s_j}}, t_i \in T, s_j \in S$ et $rt_{t_i,s_j} \in TTR$, capturant les retards historiques de la rame t_i dans toutes les stations sous forme de listes. Cet ensemble de données est formé par l'union des retards de toutes les rames précédentes de t_i sur l'ensemble du réseau de stations, comme présenté dans le tableau 3.9, où y^{past}_1 représente le premier retard d'une rame précédente dans une station et y^{past}_h représente le dernier retard d'une rame précédente avant l'heure actuelle. Dans le cadre de l'ensemble de données $XOP = (S, R, TTP, TTR, AI, A, D, PI, P, AT, Y)$, il est à noter qu'il existe un sous-ensemble Y^{past} représentant tous les retards précédents d'une rame t_i dans une station s_j avec un départ réel à rt_{t_i,s_j}. Ce sous-ensemble est tel que les retards précédents se situent dans une fenêtre temporelle d'une heure dans le passé, comme montré dans la fenêtre grisée de la figure 3.3. L'utilisation de l'ensemble complet des retards précédents dépend de la nature des données disponibles. Dans certains cas, on utilise l'intégralité des retards sur une heure dans le passé, tandis que dans d'autres cas, seules les 20 dernières minutes sont prises en compte. Le choix d'une période d'une heure a été proposé par les opérateurs Toutefois, l'utilisation de l'ensemble des données collectées durant cette heure dépend de l'algorithme d'optimisation que nous avons appliqué. Cet algorithme permet de sélectionner les retards les plus significatifs dans cette période d'une heure pour chaque modèle d'apprentissage. Cette approche est détaillée dans le chapitre 4, à travers l'analyse des deux ensembles de données : Tram et RER.

Nous définissons le sous-ensemble des retards précédents, noté $Y^{\text{past}}_{t_i,s_j,rt_{t_i,s_j}}$, pour chaque rame t_i à la station s_j, au moment du départ réel rt_{t_i,s_j}, comme suit :

$$Y^{\text{past}}_{t_i,s_j,rt_{t_i,s_j}} = \{y_{t_k,s_l,rt_{t_k,s_l}} \in Y | \quad rt_{(t_k,s_l)} < rt_{(t_i,s_j)} \text{ et } rt_{(t_k,s_l)} > rt_{(t_i,s_j)} - 3600\} \tag{3.11}$$

où :

— y_{t_k,s_l} est le retard de la rame t_k à la station s_l.
— rt_{t_k,s_l} est l'heure de départ réelle de la rame t_k à la station s_l.
— La condition $rt_{t_k,s_l} < rt_{t_i,s_j}$ garantit que nous ne considérons que les retards des rames qui ont déjà quitté les stations avant l'heure de départ réelle rt_{t_i,s_j}.
— La condition $rt_{t_k,s_l} > rt_{t_i,s_j} - 3600$ garantit que nous ne considérons que les retards des rames qui ont quitté les stations dans une fenêtre d'une heure avant l'heure de départ réelle rt_{t_i,s_j}.

3.5.3 Données d'entrée : Anticipation des rames sur un horizon temporel futur XOP^{+f}

Le sous-ensemble XOP^{+f} est formé en récupérant l'heure de départ réelle de la rame à sa station actuelle à partir de l'ensemble XOP, ainsi que les informations sur la rame associée. Ensuite, les données futures sont projetées sur un horizon f afin d'anticiper la position et l'état de la rame dans cet intervalle. Cette approche permet d'obtenir une estimation précise de l'évolution de la rame en fonction des conditions actuelles, tout en prévoyant son comportement dans le futur proche, ce qui est crucial pour anticiper d'éventuels retards ou ajustements opérationnels. Le processus détaillé est le suivant :

— Les données opérationnelles actuelles de la rame t_i à la station s_j sont prises en compte, y compris des informations telles que l'heure réelle de départ rt_{t_i,s_j}.
— Une estimation est faite pour l'arrivée de la rame t_i à la station $s_{(j+e)}$ dans une fenêtre temporelle de f minutes dans le futur.
— La position anticipée de la rame t_i dans le futur proche f est déterminée en fonction du TTP, comme illustré à la figure 3.3.
— Le sous-ensemble XOP^{+f} est formé, englobant à la fois les données actuelles et futures pour la rame t_i. Il intègre des informations relatives à l'heure réelle de départ, à l'arrivée future estimée, et à la position prévue dans le futur proche comme montré dans la figure 3.3 pour la rame 5 à la station 12.

Soit $x^{+f}_{(t_i,j_{j+e},rt_{t_i,s_j})} \in XOP^{+f}$ tel que $x^{+f}_{(t_i,j_{j+e},rt_{t_i,s_j})} = \{t_i, s_j, rt_{t_i,s_j}, s_{(j+e)}, pt_{(t_i,s_{j+e})}\}$, où :
— $s_{(j+e)}$ représente la station estimée pour une rame t_i dans les 15 prochaines minutes.
— $pt_{(t_i,s_{j+e})}$ désigne l'heure de départ prévue de la rame t_i à la station $s_{(j+e)}$.

En résumé, ce sous-ensemble XOP^{+f} est utilisé pour représenter les données futures d'une rame et est utilisé conjointement avec les ensembles XOP et Y^{past} qui représentent respectivement les variables opérationnelles (horaires) des rames et l'ensemble des retards passés d'une rame à une station. Notez que l'ensemble global de données pour notre modèle d'apprentissage est X 3.12, comme indiqué dans le Tableau 3.9. Le problème de la prédiction des retards Y est formulé sous la représentation mathématique suivante :

$$X = XOP \cap Y^{\text{past}} \cap XOP^{+f} \tag{3.12}$$

où X est l'ensemble de données global résultant de la fusion des caractéristiques actuelles XOP, des retards passés Y^{past}, et des éléments futurs XOP^{+f}. Cette formulation exprime la dépendance fonctionnelle entre les caractéristiques combinées et la variable cible.

$$\hat{y}_{(t_i,s_j)} = f(x_{(t_i,s_j,rt_{(t_i,s_j)})}), \quad x_{(t_i,s_j,rt_{(t_i,s_j)})} \in X, \quad \hat{y}_{(t_i,s_j)} \in \hat{Y}^{+f} \tag{3.13}$$

L'équation (3.13) décrit une fonction de prédiction utilisée dans le cadre d'un modèle d'apprentissage. La prédiction, notée $\hat{y}_{(t_i,s_j)}$, est obtenue en appliquant la fonction f aux variables d'entrée $x_{(t_i,s_j,rt_{(t_i,s_j)})}$. Ces variables d'entrée incluent des informations temporelles et contextuelles comme les retards passés et actuels à un instant t_i et à une station s_j. L'ensemble X représente l'espace des données d'entrée possibles, tandis que la sortie $\hat{y}_{(t_i,s_j)}$ appartient à un espace de prédiction noté $\hat{Y}^{+f}$, qui peut inclure des prévisions sur un horizon temporel futur f.

La section suivante présente les différents modèles d'apprentissage utilisés dans le cadre d'une étude comparative visant à sélectionner le modèle le plus adapté à notre cas d'utilisation. Cette analyse détaille le mécanisme d'entraînement propre à chaque modèle, ainsi que les optimisations des hyperparamètres effectuées pour améliorer leurs performances. L'objectif est de présenter les mécanismes pour comparer les résultats obtenus pour identifier le modèle offrant les meilleures prédictions tout en répondant aux exigences spécifiques du projet.

Table 3.9 – Exploration des caractéristiques pour la prédiction de retard y^{+f}

Entrées X															Sortie
XOP										XOP^{+f}		Y^{past}			Retards
t	s	rt	ai	a	d	pi	p	at	y	pt^{+f}	s^{+15}	y_1^{past}	...	y_h^{past}	y^{+f}
t_{14}	s_{10}	2017-09-01 07 :33 :00	135	34	5	300	120	22	50	07 :47 :00	s_{17}	-185	...	68	147
t_6	s_{18}	2017-09-01 07 :33 :49	96	23	5	360	96	22	110	07 :48 :30	s_3	209	...	-28	96
..	..	..	..	..	..	..	..	..	..	..	..	...	..	..	..
t_m	s_{15}	2017-12-01 23 :47 :20	130	31	7	120	120	9	-170	00 :10 :00	s_{21}	-591	...	141	-524

3.6 Mécanisme d'Entraînement et Optimisation des Modèles

Cette section présente le mécanisme d'entraînement des modèles sélectionnés a été mis en place afin de les tester dans la phase comparative, en vue de choisir le modèle le plus performant. Cette comparaison repose sur deux approches d'évaluation. La première utilise des critères standards tels que la RMSE et le coefficient de détermination R^2. La seconde approche est spécifiquement conçue pour répondre aux besoins opérationnels du PCC, afin de s'assurer que le modèle sélectionné soit en adéquation avec leurs exigences pratiques.

la figure 3.18 présente le processus d'entraînement, qui ne varie pas beaucoup entre les différents modèles d'apprentissage testés. Étant donné que l'ensemble des données à utiliser est déjà préparé selon la section 3.5, il suffit de préparer les jeux de données d'entraînement et de test. Nous utilisons la méthode classique de division des données avec une répartition de 80% pour l'entraînement et 20% pour l'évaluation. Cependant, en raison de la nature des données sous forme de séries temporelles, il est crucial de ne pas diviser les données de manière aléatoire. Au lieu de cela, les données sont segmentées par jour : 80% des jours sont utilisés pour entraîner les modèles, et les 20% restants sont réservés pour l'évaluation.

La notion de *passé*, telle que définie dans l'ensemble des données Y^{past} (voir l'équation 3.11), correspond aux retards accumulés sur une fenêtre temporelle d'une heure dans le passé. Chaque modèle d'apprentissage utilise un ensemble de retards précédents couvrant cette fenêtre. Une étude préalable a été réalisée pour déterminer le nombre optimal de retards à prendre en compte pour chaque modèle. Cette optimisation repose sur une fonction récursive, présentée dans l'algorithme 4, qui ajuste dynamiquement ce paramètre en fonction des performances du modèle. La fonction évalue les performances à chaque itération et retourne le nombre optimal de retards à inclure avant la phase d'entraînement.

Par ailleurs, une validation croisée est appliquée sur les données d'entraînement afin d'assurer une meilleure convergence et une généralisation plus robuste des modèles. Contrairement à une validation classique, celle-ci respecte l'ordre chronologique des séries temporelles pour éviter de mélanger aléatoirement les données. Cette méthode est détaillée dans l'algorithme 3.

Enfin, chaque modèle entraîné nécessite une phase d'optimisation des hyperparamètres, comme expliqué dans la section 2.3.4. Deux algorithmes d'optimisation sont adoptées selon le type de modèle. Pour les modèles basés sur les arbres de décision, l'algorithme d'optimisation bayésienne, décrit dans l'algorithme 1, est utilisé pour affiner les hyperparamètres. En revanche, pour les modèles de Deep Learning, une

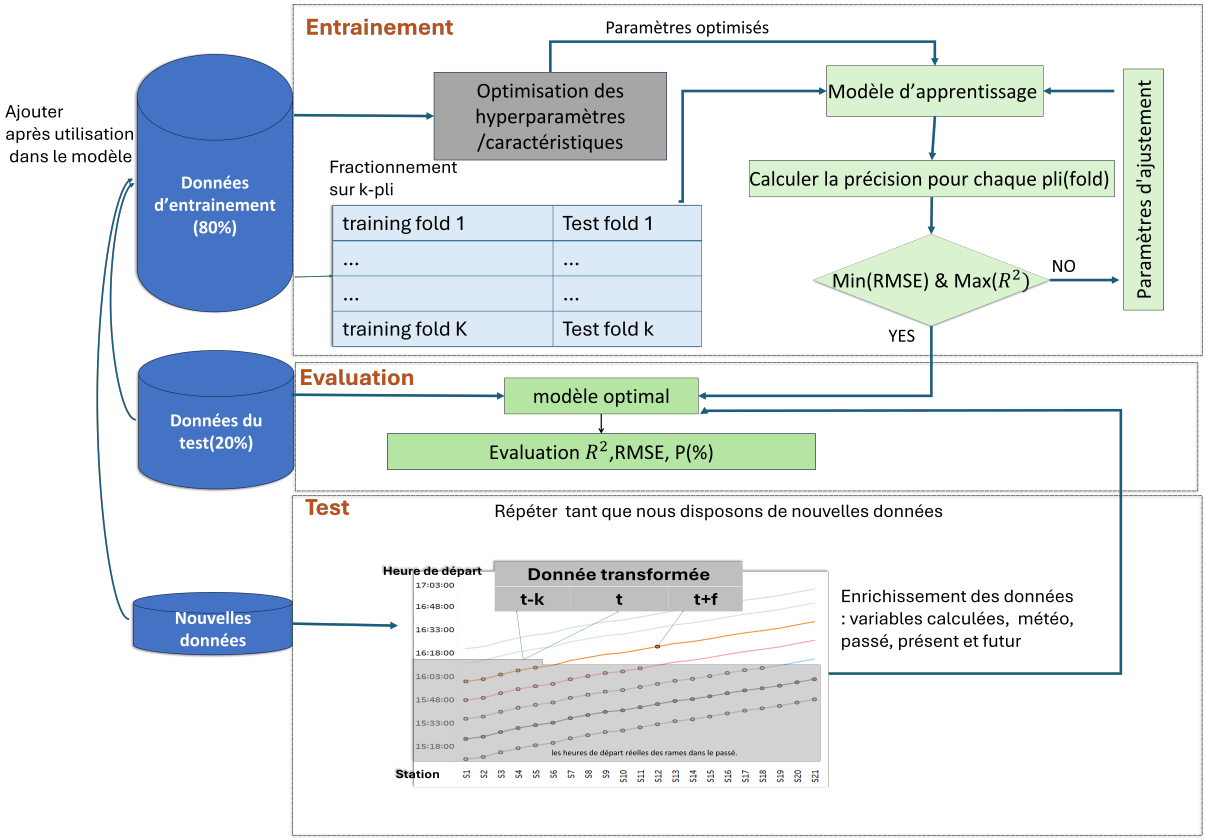


Figure 3.18 – Processus d'entraînement des modèles d'apprentissage pour une analyse comparative

recherche par grille (*Grid Search*), présentée dans l'algorithme 2, est mise en place pour déterminer les hyperparamètres optimaux. Cette méthodologie rigoureuse permet de garantir des performances optimales pour chaque modèle d'apprentissage, tout en améliorant leur précision et leur capacité de généralisation.

Nous avons ajouté la possibilité, en cas de réception de nouvelles données, de les tester facilement avec le modèle optimal. Cependant, avant d'utiliser ce modèle, il est nécessaire de préparer les données en suivant les étapes d'enrichissement présentées dans la figure 3.15, ainsi que de préparer les retards précédents avec l'estimation de la position future du train à court terme. Une fois ces étapes accomplies, il est possible de tester les nouvelles données sur notre modèle optimal en toute sérénité. Enfin, après chaque test de nouvelles données, nous intégrons celles-ci au jeu de données d'entraînement afin de ré-entraîner le modèle avec ces informations mises à jour.

La section suivante présente le travail réalisé sur l'évaluation des modèles sur les données de test, afin de garantir une performance optimale des modèles. L'accent est mis sur deux aspects de l'évaluation : d'une part, les métriques d'évaluation utilisées, et d'autre part, une méthode spécifique proposée par les opérateurs.

3.7 Évaluation des modèles

L'évaluation du modèle est une étape cruciale pour valider sa performance sur un jeu de données qu'il n'a jamais vu auparavant. Une fois le modèle minutieusement optimisé, il subit des tests rigoureux sur l'ensemble de données de test, représentant

Algorithme 3 : Détermination du nombre optimal de plis pour la validation croisée

Data : $\mathcal{X}$ (ensemble de données d'entraînement), $\mathcal{M}$ (modèle), K (liste des plis possibles)
Result : Nombre optimal de plis K_{opt}

$K_{\text{opt}} \leftarrow 0$;
$\text{perf}_{\text{best}} \leftarrow 0$;
for *chaque* $k \in K$ **do**
    Appliquer une validation croisée avec k plis sur $\mathcal{X}$;
    Calculer la performance moyenne perf_k du modèle $\mathcal{M}$;
    if $\textit{perf}_k > \textit{perf}_{\textit{best}}$ **then**
        $K_{\text{opt}} \leftarrow k$;
        $\text{perf}_{\text{best}} \leftarrow \text{perf}_k$;
return K_{opt};

Algorithme 4 : Sélection récursive de variables avec restauration en cas de diminution de la performance

Data : X (ensemble de données d'entraînement), $\mathcal{M}$ (modèle), y (cible)
Result : Variables sélectionnées optimales, meilleure performance du modèle

$V_{\text{restantes}} \leftarrow \{x_1, x_2, \ldots, x_n\}$; `// Initialiser l'ensemble complet des variables`
$\text{perf}_{\text{best}} \leftarrow 0$; `// Initialiser la meilleure performance à 0`
$V_{\text{opt}} \leftarrow V_{\text{restantes}}$; `// Ensemble optimal de variables`
while $|V_{\textit{restantes}}| > 1$ **do**
    for *chaque variable* $x_i \in V_{\textit{restantes}}$ **do**
        `// Étape 1 : Tester la suppression de la variable`
        Retirer temporairement x_i de $V_{\text{restantes}}$;
        Entraîner le modèle $\mathcal{M}$ sur l'ensemble réduit et calculer la nouvelle performance $\text{perf}_{\text{temp}}$;
        if $\textit{perf}_{\textit{temp}} < \textit{perf}_{\textit{best}}$ **then**
            `// Si la performance diminue, restaurer la variable`
            Réintégrer x_i dans $V_{\text{restantes}}$;
        else
            `// Si la performance s'améliore ou reste stable, mettre à jour la meilleure performance et supprimer la variable`
            $\text{perf}_{\text{best}} \leftarrow \text{perf}_{\text{temp}}$;
    $V_{\text{opt}} \leftarrow V_{\text{restantes}}$;
return V_{opt}, $\text{perf}_{\text{best}}$; `// Retourner l'ensemble optimal de variables et la meilleure performance`

20% de l'ensemble total des données. Cette phase d'évaluation est essentielle pour mesurer les performances du modèle dans des conditions réelles, sur des données inédites. Notre évaluation repose sur deux approches complémentaires. La première est basée sur des critères classiques d'évaluation des modèles d'entraînement, tels que la minimisation de la RMSE et l'optimisation du (R^2).

La seconde évaluation nommée P(%) adopte une approche spécifique pour analyser la capacité du modèle à prédire les retards. Nous utilisons ici une marge d'erreur définie par les PCC egis-rail pour chaque départ d'une rame. Les experts ont fixé une marge d'erreur de 5 minutes pour les arrivées, avec un totale de 150 secondes pour les retards et de -150 secondes pour les avances. Cette approche permet d'établir des critères précis pour évaluer la précision du modèle dans la prédiction des retards comme des avances. Ainsi, cette double évaluation offre une vue d'ensemble complète des performances du modèle, à la fois sur la précision générale de la prédiction et sur sa capacité à anticiper les retards avec une marge d'erreur contrôlée.

$$E = \{e \in E \mid e = y - \hat{y}\} \tag{3.14}$$

où, l'ensemble des erreurs admissibles E est divisé en deux sous-ensembles : TI (Prédictions Correctes) et FI (Prédictions Incorrectes), définis comme suit :

$$E = TI \cup FI \tag{3.15}$$

$$TI = \{ti \in TI \mid -150 \leq ti \leq 150\} \tag{3.16}$$

$$FI = \{FI \in FI \mid FI < -150 \text{ ou } FI > 150\} \tag{3.17}$$

De plus, on peut exprimer la précision P en termes de cardinalité d'ensemble, capturant la proportion de prédictions correctes parmi le nombre total d'éléments dans l'ensemble E tel que :

$$P(\%) = \frac{|TI|}{|TI| + |FI|} \tag{3.18}$$

Cette approche permet une évaluation exhaustive de la précision du modèle, en lien avec les procédures définies dans le PCC, en distinguant clairement les prédictions correctes (TI) des prédictions incorrectes (FI). Elle permet ainsi d'étudier si la marge d'erreur entre les prédictions et les données réelles se situe bien dans l'intervalle d'erreur défini par l'opérateurs.

La section suivante présente le principe de l'analyse de sensibilité appliqué à notre méthodologie, visant à garantir la fiabilité. Cette analyse permet d'évaluer dans quelle mesure les variations des paramètres et des données d'entrée influencent les performances des modèles, afin d'identifier les points faibles et les facteurs critiques.

3.8 Analyse de Sensibilité

Une étude de sensibilité dans le cadre de notre méthodologie proposée consiste à analyser l'impact des variations des paramètres du modèle ou des caractéristiques des données sur les performances des modèles à tester. Cette analyse permet de comprendre quels paramètres ou variables influencent le plus la sortie du modèle et comment des modifications mineures peuvent affecter les prédictions ou les résultats. Nous proposons cette analyse afin de vérifier la robustesse de notre méthodologie sur

différents ensembles de données et autres infrastructures, tout en gardant à l'esprit que nous travaillons toujours sur des données d'exploitation.

Les premiers tests sont effectués sur les données du tramway, tandis que l'étude de sensibilité est réalisée sur les données du RER afin de valider la méthodologie décrite dans ce chapitre. Cette approche permet d'évaluer les performances du modèle sur une autre infrastructure avec des paramètres différents, ce qui nous aide à étudier la fiabilité de notre méthodologie. L'objectif est de comprendre comment celle-ci se comporte lorsque l'infrastructure change, par exemple en passant à une autre infrastructure ferroviaire (tramway à RER) . De plus, l'étude se concentre sur les niveaux de corrélation entre les différentes périodes (mois, années, jours), ainsi que sur la relation entre le passé, le présent et le futur. La sélection de la fenêtre temporelle (passé, présent) est également testée afin de déterminer si notre méthodologie est applicable à divers types de données d'exploitation ou si elle est limitée aux premiers jeux de données qui ont servi à sa construction. Plusieurs tests sont effectués à ce niveau pour répondre à cette question.

3.9 Conclusion

En s'appuyant sur les exigences de notre partenaire et sur l'état de l'art 2, qui positionne scientifiquement les solutions dédiées à ce type de problématique dans les PCCs, une approche innovante a été développée. Celle-ci repose sur une architecture hybride, détaillée dans ce chapitre, et décrit les étapes suivies pour concevoir un modèle performant, capable de fonctionner efficacement avec l'ensemble des données d'exploitation.

Notre démarche s'appuie sur des tests successifs effectués sur les deux ensembles de données fournis par Egis Rail, enrichis par de nombreuses interactions avec des experts du domaine ferroviaire. Ces échanges ont permis de mieux cerner les défis liés à la gestion des retards, un phénomène particulièrement complexe à prédire et à contrôler en raison de la multiplicité des facteurs en jeu.

Notre méthodologie commence par le nettoyage et la préparation des données. Elle intègre ensuite des conceptions temporelles (passé, présent et futur) comme variables d'entrée dans les modèles que nous avons sélectionnés pour les tests. Chaque modèle est ensuite intégré dans une analyse comparative afin d'identifier celui qui s'adapte le mieux à chaque ensemble de données. Enfin, nous avons conclu notre recherche par une étude de la sensibilité de notre méthodologie, afin d'identifier les conditions sous lesquelles notre approche fonctionne efficacement et celles où un modèle moins performant pourrait émerger, même si les mêmes étapes sont suivies. Cette étude représente seulement le début d'une analyse plus approfondie visant à affiner notre approche et à améliorer la robustesse des modèles prédictifs.

Le chapitre suivant présente l'application de notre méthodologie sur deux ensembles de données réelles provenant de deux infrastructures différentes : le Tramway et le RER. L'objectif est de valider chaque étape de la méthodologie proposée et de sélectionner le modèle le plus optimal.

Dans un premier temps, l'ensemble de données du Tramway est dédié à une étude comparative des modèles d'apprentissage, en utilisant des algorithmes d'optimisation afin de déterminer le modèle le plus performant. Ensuite, le second ensemble de données, issu du RER, permet d'enrichir l'analyse grâce à une étude de sensibilité. Cette

étape vise à vérifier la validité et la robustesse de notre méthodologie sur une autre infrastructure, assurant ainsi sa capacité à s'adapter à différents contextes.

4

Analyse et résultats

Sommaire

3.1 Démarche proposée 58

3.2 Approche relationnelle pour la modélisation des données ferroviaires 61

3.2.1 Méthode de travail 63

3.2.2 Création du modèle relationnel 64

3.2.3 Cas d'usage 67

3.2.4 Représentation mathématique des données d'exploitation XOP 69

3.3 Prétraitement des données 70

3.3.1 Nettoyage des données 71

3.3.2 Correction des incohérences temporelles 73

3.3.3 Gestion des valeurs aberrantes 77

3.3.4 Enrichissement des données 80

3.3.5 Mise en forme et encodage des variables pour l'entraînement . 80

3.4 Analyse des données 82

3.5 Ingénierie des caractéristiques temporelles 86

3.5.1 Structuration temporelle : impact du passé et projection vers le futur 86

3.5.2 Données d'entrée : Comportement historique des rames dans une fenêtre temporelle Y^{past} 87

3.5.3 Données d'entrée : Anticipation des rames sur un horizon temporel futur XOP^{+f} 88

3.6 Mécanisme d'Entraînement et Optimisation des Modèles . . . 90

3.7 Évaluation des modèles 91

3.8 Analyse de Sensibilité 93

3.9 Conclusion 94

Ce chapitre présente la mise en œuvre de la méthodologie décrite dans le chapitre 3, en l'appliquant à deux ensembles de données réelles décrivant les infrastructures du RER et du tramway. L'objectif est de valider notre approche

en utilisant ces données et en tenant compte des retours des opérateurs des Postes de Commandement Centralisés (PCC). Cependant, nous n'avons pas pu obtenir de données supplémentaires pour poursuivre le travail de consolidation de la robustesse de notre méthodologie.

Après la récupération des données, il est essentiel d'analyser et de traiter ces données de manière adéquate, en tenant compte des types de données et des variables disponibles dans chaque ensemble. Notre méthodologie repose d'abord sur le prétraitement des données et l'analyse, comme décrit dans la section 3.3. Plusieurs étapes spécifiques sont suivies selon la nature des données traitées. Ce prétraitement permet de corriger les erreurs détectées dans chaque ensemble, d'analyser les données et d'identifier des méthodes susceptibles d'améliorer les prédictions. De plus, il nous aide à différencier les deux ensembles de données, un élément crucial pour notre étude de sensibilité, qui évalue l'impact de ces différences sur la méthodologie.

Suite à ce prétraitement, nous intégrons les données dans le modèle d'apprentissage présenté dans la section 2.3.4, afin de réaliser une étude comparative. Cette étude porte d'abord sur les données du tramway. Ensuite, une analyse de sensibilité est menée sur le second ensemble, les données du RER, afin de mettre en lumière les forces et les faiblesses de notre méthodologie. L'évaluation des résultats s'appuie sur des mesures standard, ainsi que sur celles proposées par les opérateurs des PCC. Cette structure nous permet d'explorer en profondeur l'efficacité de notre méthodologie sur deux types de réseaux de transport, tout en prenant en compte les contraintes opérationnelles des infrastructures étudiées.

Le plan du chapitre est présenté comme suit : La section 4.1.1, environnement de travail, présente les outils, packages, et bibliothèques utilisés pour l'implémentation des modèles d'apprentissage automatique, ainsi que les versions des logiciels et environnements nécessaires. La section 4.1.3, présentation des données collectées, organisation et format, décrit les données utilisées, leur origine, ainsi que leur structure et format, offrant une vue d'ensemble des variables analysées. La section 4.2, mise en forme et préparation des données, couvre les étapes de prétraitement appliquées aux données, incluant le nettoyage, la transformation, et la sélection des variables nécessaires à l'analyse. Dans la section 4.3, traitement analytique, deux sous-sections détaillent l'analyse des retards en fonction des variables XOP et Y_{past}, permettant de mieux comprendre les facteurs influençant les retards dans les réseaux de transport. La section 4.4, la prévision à court terme du retard : une première analyse exploratoire (Tramway), examine les méthodes de prévision appliquées au réseau de tramway, incluant le partitionnement des données, l'entraînement des modèles, l'ajustement des hyperparamètres, et l'évaluation des performances des différents modèles d'arbres de décision et d'apprentissage profond. Enfin, la section 4.5, analyse de sensibilité, explore la robustesse des modèles prédictifs en étudiant leur sensibilité aux modifications des infrastructures et aux différences entre les réseaux de tramway et RER. Cette section inclut également une seconde analyse exploratoire sur le réseau RER.

4.1 Cadre de travail

4.1.1 Environnement de travail

La figure 4.1 présente les différents packages et bibliothèques utilisés dans le cadre de ce projet de thèse pour la mise en place et l'entraînement des modèles d'apprentis-

sage automatique. L'ensemble de ces outils fonctionne sous la version `Python 3.11.3`, qui fournit la compatibilité nécessaire pour les bibliothèques utilisées.

Tout d'abord, les méthodes statistiques incluent la régression linéaire et la régression bayésienne, dont l'objectif est de modéliser les relations entre les variables de manière linéaire. Pour la régression linéaire, la bibliothèque `Scikit-learn` (version 1.2.2) a été utilisée, et le modèle est enregistré au format `.joblib`. Quant à la régression bayésienne, elle est implémentée avec `TensorFlow Probability` intégrée dans `TensorFlow` (version 2.13.0), et le modèle est sauvegardé en format `.h5` pour une gestion efficace des paramètres probabilistes.

Ensuite, les modèles basés sur les arbres de décision regroupent les arbres de décision, les forêts aléatoires et XGBOOST . Ces modèles, bien adaptés aux données tabulaires, permettent des applications en classification et en régression. Les arbres de décision et les forêts aléatoires sont implémentés avec `Scikit-learn` (version 1.2.2) et sont enregistrés chacun au format `.joblib`. Le modèle XGBOOST quant à lui utilise la version 1.7.6 de la bibliothèque `XGBOOST` et est enregistré au format `.pkl`.

Enfin, les modèles d'apprentissage profond incluent les réseaux de neurones artificiels (ANN), les réseaux de mémoire à long terme (LSTM), et les LSTM bidirectionnels (BILSTM). Ces modèles, particulièrement adaptés pour les données séquentielles, sont implémentés avec `TensorFlow` (version 2.13.0) et l'API `Keras`, ce qui facilite leur création et leur entraînement. Chaque type de modèle d'apprentissage profond est enregistré en format `.h5`, un format optimisé pour stocker les architectures complexes de réseaux neuronaux.

Pour la visualisation des données et des résultats, les bibliothèques `Matplotlib` (version 3.7.1), `Seaborn` (version 0.12.2), et `Plotly` (version 5.9.0) sont utilisées, offrant des outils flexibles et interactifs pour analyser graphiquement les performances des modèles et illustrer les résultats obtenus. L'environnement de travail ainsi décrit constitue une infrastructure technique solide et diversifiée, répondant aux exigences de cette thèse en termes de développement, d'entraînement et d'analyse des différents types de modèles d'apprentissage automatique.

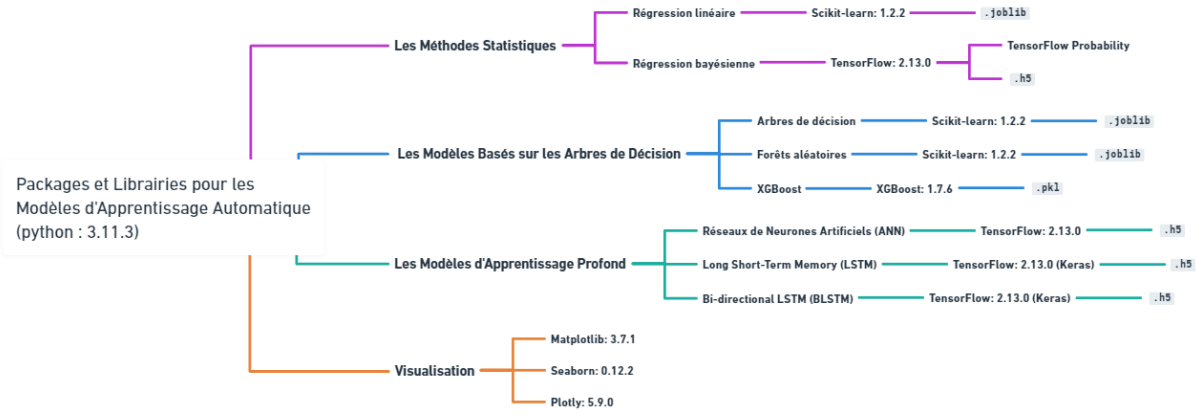


Figure 4.1 – Packages et extensions utilisés pour les modèles d'apprentissage automatique

La section suivante présente les deux infrastructures de données utilisées pour entraîner les modèles (Tramway, RER), en détaillant la description de chaque ensemble de données. Cette présentation inclut les caractéristiques principales des données, telles que leur structure, leur format.

4.1.2 Sources de données (Tram, RER)

Il nous est paru essentiel de rappeler les contraintes et difficultés rencontrées pour l'obtention des données nécessaires à notre travail. En effet, le secteur ferroviaire est souvent associé à une gestion collaborative liée à différents acteurs, ce qui rend difficile l'accès à ces informations. C'est d'autant plus fort avec les aspects confidentiel et concurrentiel.

Dans ce contexte, nous avons pu établir un état des lieux pour mieux appréhender cette situation en s'appuyant sur des discussions avec différents opérateurs ainsi que de nombreuses visites dans des postes de contrôle ferroviaires en France et à l'international. Nous avons ainsi identifié plusieurs problématiques/limitations à savoir : certaines données sont confidentielles, d'autres sont manquantes, et dans certains cas, les historiques de données sont supprimés, conservant seulement les trois derniers mois de données. De plus, certains opérateurs ont exprimé des réticences à intégrer le système d'intelligence artificielle dans leurs pratiques.

Par ailleurs, notre partenaire EGIS a pu se procurer deux jeux de données issus de sources de collaboratives : une ligne tram et une ligne RER. Ces données d'exploitation contiennent principalement des informations relatives aux départs des rames pour chaque station, incluant les temps de passage et les temps de retard. À la suite de notre analyse, nous avons enrichi ces données pour les rendre exploitables dans le cadre de notre projet. Aussi, un travail d'uniformisation des informations entre les deux ensembles de données a été mené pour adapter leurs formats de tel sorte à garantir un usage identique.

Figure 4.2 illustre une chronologie de l'évolution de la collecte des jeux de données utilisés dans notre recherche. Le premier ensemble de données concerne une ligne de tramway de 22 stations, que nous avons récupéré (ensemble des données) à la fin de la première année de notre travail, afin de lancer les tests des modèles d'apprentissage. La collecte des données restantes de cette ligne de tramway a été finalisée à la fin de la deuxième année. Le second ensemble de données, qui décrit une ligne ferroviaire RER comprenant 11 stations situées au milieu de la ligne, a été acquis en mai 2024, au cours de la troisième année. Cet ensemble a servi à valider le modèle développé avec le premier jeu de données et à commencer l'analyse de sensibilité de notre méthodologie. En plus des données fournies par Egis, nous avons également utilisé la méthode de web scraping pour récupérer des données météorologiques en ligne, afin de compléter et d'enrichir notre base de données. Ces données météorologiques, collectées pour chaque jour correspondant aux événements enregistrés dans notre jeu de données, nous ont permis d'analyser l'impact potentiel des conditions climatiques sur les retards des trains et des tramways.



Figure 4.2 – Évolution de la collecte des jeux de données pour le tramway et le RER au cours du projet de recherche

En raison des exigences strictes en matière de confidentialité liées aux données reçues, nous sommes tenus de protéger et de masquer toute information pouvant révéler l'infrastructure ou spécifier la ligne concernée. La figure 4.3 présente les étapes nécessaires pour assurer une protection adéquate des données, notamment l'anonymisation des noms des stations, la simplification de la représentation de l'infrastructure, ainsi que l'abstraction des terminus et des directions. Les noms des stations sont remplacés par des identifiants génériques comme "Station 1", "Station 2", etc, ou par des symboles, tandis que les terminus sont indiqués par des icônes génériques. Afin de garantir cette confidentialité, les données dans les fichiers CSV et les graphiques sont représentées de manière simplifiée : les stations sont symbolisées et les trains illustrés par des formes géométriques telles que des rectangles. L'infrastructure est également visualisée de façon abstraite, avec les stations disposées de manière linéaire, sans révéler d'informations sensibles.

Nous présentons ensuite le format et les caractéristiques de chaque ensemble de données étudié, à savoir celles du RER et du tramway.

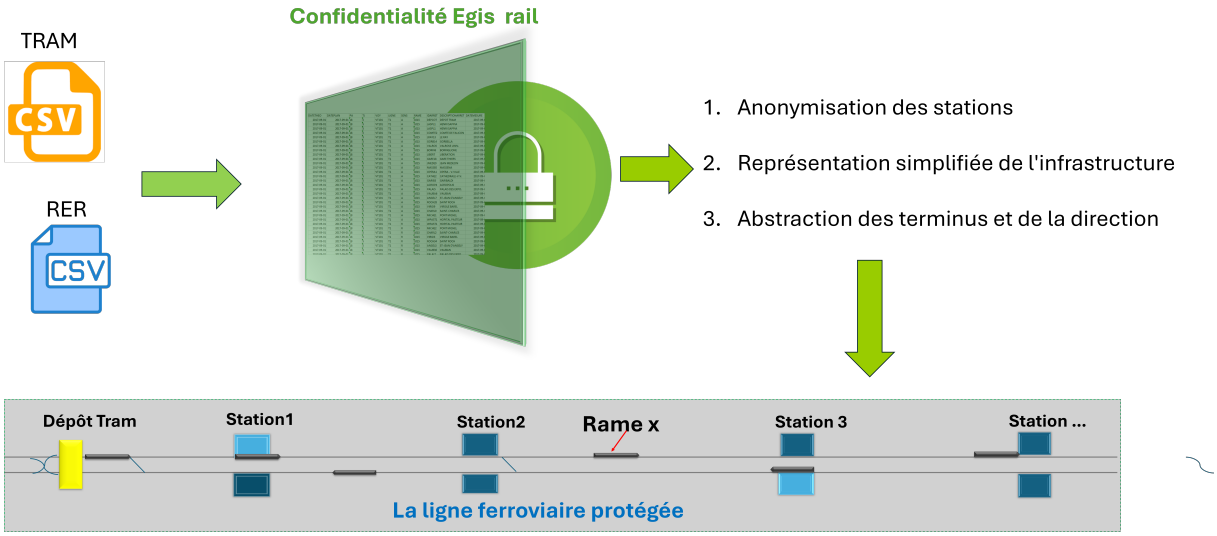


Figure 4.3 – Schéma abstrait d'une ligne ferroviaire avec confidentialité des données

4.1.3 Présentation des données collectées : organisation et format

Une ligne ferroviaire est un itinéraire fixe reliant plusieurs stations ou gares, utilisé pour le transport de passagers ou de marchandises. Durant une journée typique, plusieurs rames circulent sur cette ligne, se déplaçant d'une station à l'autre selon un TTP. Le nombre de stations desservies est fixe, déterminant les points de départ et d'arrivée de chaque rame. Selon le plan d'exploitation, certains trains peuvent être directs, semi-directs ou omnibus, s'arrêtant à toutes les stations ou seulement à certaines d'entre elles. Divers facteurs, tels que la météo, les incidents techniques et la coordination avec d'autres lignes du réseau, influencent la fluidité du service au quotidien. Ces éléments combinés garantissent le bon fonctionnement et l'efficacité de la ligne tout au long de la journée.

Les deux jeux de données contiennent un ensemble de lignes données, chaque ligne correspondant à des informations organisées en colonnes sous les format CSV. Ces colonnes décrivent, entre autres, la date et l'heure de départ, réelles et prévues, de

chaque rame, ainsi que l'identifiant de la rame et celui de la station correspondante. De plus, chaque ligne est divisée en deux voies, aller et retour. Pour les données RER, une nouvelle variable apparaît, qui n'existe pas dans celles du tramway : le nom du terminus. En effet, la ligne RER possède cinq terminus, et cette variable permet d'identifier la direction de chaque train vers l'un de ces terminus. À l'inverse, les données du tramway incluent des informations supplémentaires, telles que le temps de stationnement prévu et le temps interstation pour chaque rame, des informations qui ne sont pas présentes dans les données RER, comme illustré à la figure 4.4. Ces variables ont été fournies dans leur état brut par Egis Rail.

La ligne de tramway est au format de 2396287 lignes de données, elle est complète et comprend 22 stations, avec deux voies permettant les trajets aller et retour, ainsi que deux terminus. Les données couvrent la période du 1er septembre 2017 au 30 juin 2018. En revanche, la ligne RER est partielle, au format de 926615 lignes de données : elle inclut 11 stations sur un total de 46, situées au milieu de la ligne. L'infrastructure de la ligne comporte cinq terminus, Cependant, les données que nous utilisons se concentrent uniquement sur le premier trimestre des années 2017, 2018 et 2019 et ne comprennent pas les samedis, dimanches, ni les jours fériés.

La représentation des données d'exploitation reçues présente des similarités dans leur structure pour les deux ensembles, bien que des différences existent dans les noms des colonnes et le format des variables entre les différents jeux de données. Toutefois, le contenu reste fondamentalement le même pour tous les ensembles de données d'exploitation. Par exemple, dans un jeu de données, la colonne indiquant la station concernée peut être nommée "Description arrêt", tandis que dans un autre, elle sera désignée par "gare". Cependant, dans les deux cas, le contenu de ces colonnes correspond à des chaînes de caractères représentant les noms des stations de la ligne en question. De même, un autre exemple concerne les heures de départ des rames : dans un jeu de données, l'heure peut être présentée au format "12-09-2017 13 :14 :04", alors que dans un autre, elle pourrait être affichée sous le format "12/09/2017 13 :14 :04". Bien que les formats diffèrent, le contenu est identique. Notre objectif est d'harmoniser ces données en les convertissant dans un format uniforme à l'aide de fonctions spécifiques, afin de les rendre lisibles et exploitables par notre modèle d'apprentissage automatique.

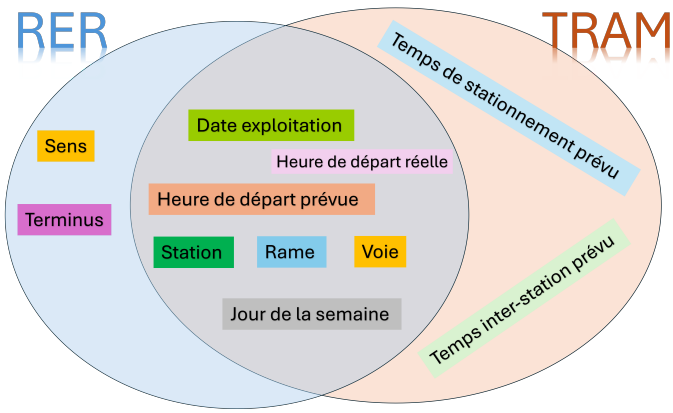


Figure 4.4 – Structure des données brutes : RER vs Tramway

La section suivante présente le nettoyage effectué sur les deux ensembles de données, en détaillant les étapes de correction et de suppression appliquées aux données.

Les valeurs erronées ou manquantes ont été corrigées lorsque cela était possible, et les données non réparables ont été supprimées pour assurer la qualité et la cohérence des informations utilisées dans l'entraînement des modèles.

4.2 Mise en forme et préparation des données

La phase de prétraitement des données, illustrée par la figure 3.7, décrit les divers problèmes rencontrés dans un ensemble de données d'exploitation. Cette phase vise à identifier ces problèmes et à proposer des solutions afin de maintenir l'intégrité de la base de données et de la rendre prête à être utilisée dans la prochaine étape, qui est l'analyse des données. Le tableau 4.1 présente le nombre de lignes impactées pour chaque ensemble de données (Tramway et RER), avec un code couleur. Les cases rouges indiquent les lignes qui ont été supprimées, tandis que les cases vertes montrent que ces lignes ont été corrigées à l'aide des méthodes décrites dans la section 3.3.

Phase de prétraitement	**Type d'erreur**	**TRAMWAY**	**RER**
Nettoyage des données	Format date	1 882	7 009
	Identifiant station	160	75
	Identifiant rame	0	40
	Lignes dupliquées	10 607	821
Gestion des incohérences temporelles	Valeur manquante	833	3 108
	Incohérence entre les horaires	2 089	14 400
Gestion des valeurs aberrantes		119 771	22 517

Table 4.1 – Nombre de lignes impactées par la phase de prétraitement

Ces méthodes incluent la correction des formats de date, des identifiants, ainsi que le traitement des valeurs manquantes (voir figure 3.8). Toutefois, le problème le plus critique concernait les incohérences entre les horaires de départ réels et prévus, qui représentaient le pourcentage le plus élevé des lignes impactées. Pour remédier à ces incohérences, des méthodes spécifiques ont été mises en place (voir figures 3.11 et équation 3.6) afin de garantir la cohérence des horaires. En ce qui concerne les valeurs aberrantes, deux méthodes ont été mises en place pour répondre à cette problématique, en accord avec la proposition des opérateurs, avec un intervalle suggéré de [-360, 900] secondes. Ces méthodes visaient à valider cet intervalle, la première étant basée sur l'évolution des quartiles des retards pour les deux ensembles de données. Les deux graphiques comparent les retards quotidiens pour le RER et le Tramway à travers leurs quartiles. Nous avons tracé le 1er quartile (Q1), la médiane (Q2), et le 3e quartile (Q3) des retards quotidiens. les graphique montrent ainsi une dispersion des retards, où l'on peut observer des pics dans le 3e quartile (ligne verte), indiquant des journées avec des retards plus importants. La médiane (ligne bleue) donne une vue d'ensemble des retards "typiques", tandis que le 1er quartile (ligne rouge) reflète les journées avec les plus faibles retards ou des avances. Pour le RER (voir figure

4.5), les retards sont plus extrêmes, avec des pics dépassant parfois 1750 secondes, notamment pour le 3e quartile (Q3), ce qui indique des événements perturbateurs majeurs. Toutefois, la médiane reste relativement stable autour de zéro, ce qui montre que la majorité des rames subissent peu de retards. En revanche, pour le Tramway (voir figure 4.6), les retards sont globalement plus modérés, avec des valeurs maximales atteignant rarement 1200 secondes et une médiane plus proche de zéro, suggérant une meilleure régularité des services. Ces deux graphiques mettent en lumière des différences significatives dans la gestion des retards entre ces deux infrastructures.

L'implémentation de la méthode Isolation Forest a joué un rôle crucial dans la validation de l'intervalle proposé par les opérateurs. Les figures 4.7 et 4.8 illustrent clairement que les valeurs aberrantes (en rouge) sont beaucoup plus dispersées et élevées, reflétant des événements inhabituels dans le système, tels que des pannes, des grèves ou d'autres incidents majeurs. À l'inverse, les valeurs normales (en bleu) correspondent au fonctionnement régulier avec des retards modérés, situés dans l'intervalle [-360, 1000] secondes. Une fonction a été implémentée pour remplacer ces valeurs aberrantes par l'intervalle acceptable de [-360, 1000], ce qui a conduit à la détection de 119771 valeurs aberrantes pour les données du Tramway et 22517 pour le RER. Cela permet de maintenir une circulation fiable des données, tant pour les analyses que lors de la phase d'apprentissage des modèles. Enfin, en ce qui concerne les lignes dupliquées, la solution adoptée a consisté à les supprimer pour garantir l'intégrité des données.

Après la phase de préparation des données pour les deux ensembles de données, Tram et RER, nous sommes passés d'un total de 2406894 à 2396287 lignes pour le tramway, et de 926615 à 925754 lignes pour le RER. Suite à la projection de notre diagramme de classe, présenté dans la figure 3.6, nous avons ajouté plusieurs variables calculées à partir de la phase d'enrichissement, illustrée dans la figure 3.15. Ainsi, pour chaque ensemble de données, nous sommes passés de 11 à 17 colonnes pour les données du tramway, et de 10 à 16 colonnes pour les données du RER plus la liste des retards précédents pour chaque ensemble de données. La phase suivante est dédiée à l'analyse des tendances des deux ensembles de données, Tram et RER. Cette étape vise à identifier des motifs récurrents, des variations saisonnières ou des comportements spécifiques qui pourraient influencer les performances du réseau ferroviaire. En examinant ces tendances, nous espérons mieux comprendre les facteurs sous-jacents qui contribuent aux retards et aux perturbations, ainsi que les différences éventuelles entre les deux infrastructures. Cette analyse servira de base pour affiner nos modèles prédictifs et améliorer les recommandations fournies aux opérateurs en temps réel.

La section suivante présente le traitement analytique réalisé sur les deux ensembles de données pour en comprendre la nature et les relations entre chaque entité. Une étude approfondie a été menée pour analyser les retards en fonction des données d'exploitation actuelles, ainsi que des retards en fonction des variables passées. Cette analyse permet d'identifier les facteurs influents et d'établir des corrélations utiles pour la modélisation et la prévision des retards dans les réseaux étudiés.

4.3 Traitement analytique

Cette section vise à analyser les deux ensembles de données afin de mieux comprendre les corrélations entre les variables et leur importance pour les modèles d'apprentissage automatique. L'objectif est d'identifier les tendances des retards pour

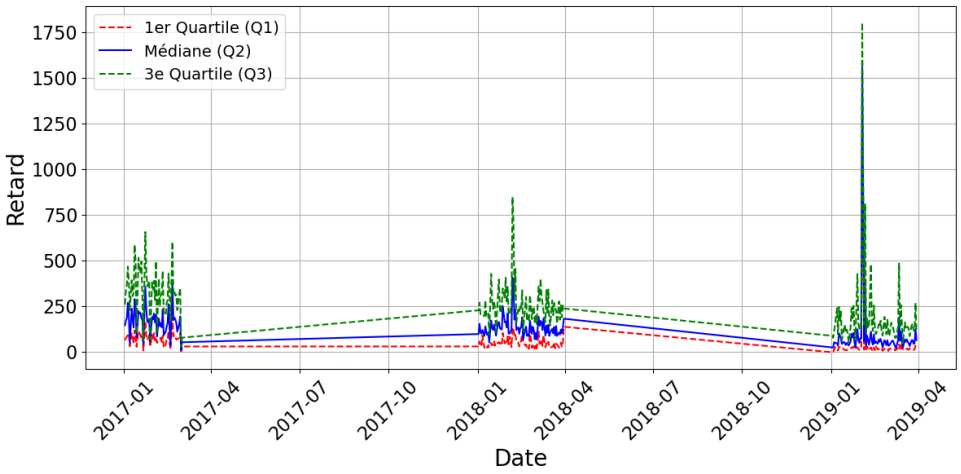


Figure 4.5 – Évolution des retards selon les quartiles (Q1, Q2, Q3) au cours du temps pour infrastructure RER

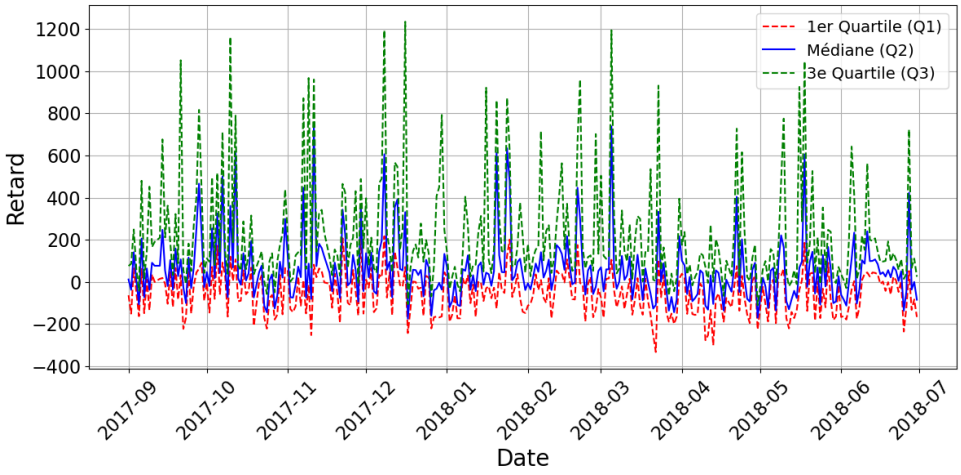


Figure 4.6 – Évolution des retards selon les quartiles (Q1, Q2, Q3) au cours du temps pour infrastructure Tramway

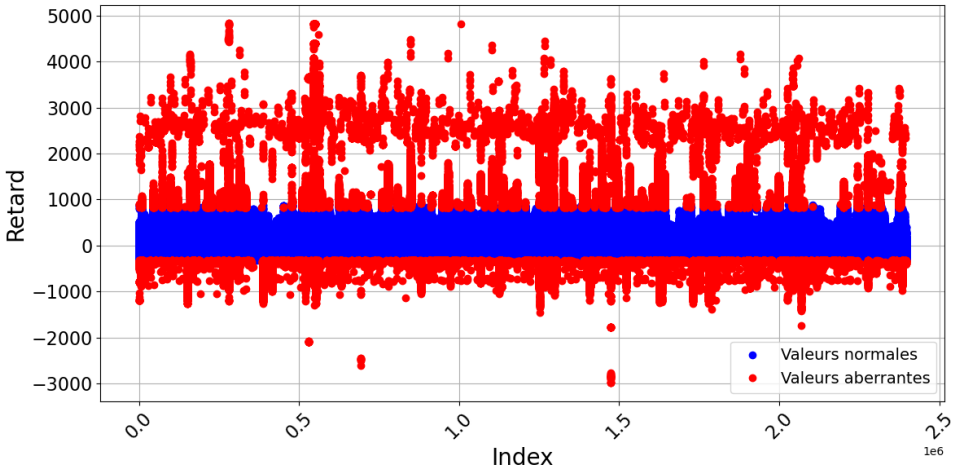


Figure 4.7 – Visualisation des retards avec détection des valeurs aberrantes par isolation forest pour infrastructure tramway

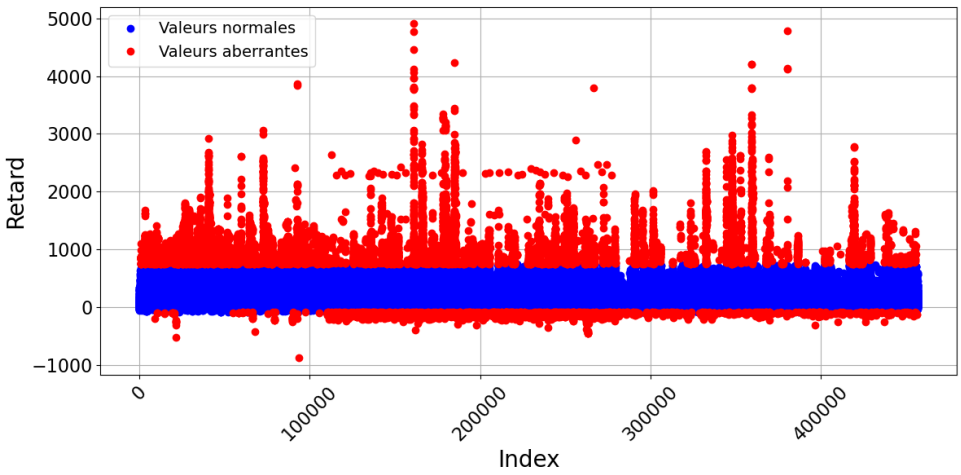


Figure 4.8 – Visualisation des retards avec détection des valeurs aberrantes par isolation forest pour infrastructure RER

chaque ensemble de données, notamment les jours de la semaine les plus perturbés et les heures les plus affectées par des retards importants. Nous allons également comparer les jours de la semaine pour déterminer s'il existe des similarités ou des différences notables en termes de retards. Par exemple, est-ce que les jours de la semaine se ressemblent dans leur comportement face aux retards, ou bien chaque jour présente-t-il un schéma unique, même s'il appartient à la même catégorie de jour (jour ouvrable, week-end, etc) ? En outre,Nous analyserons l'impact de divers facteurs externes, tels que la température et la période de l'année, ainsi que le jour de la semaine sur les retards. Est-ce que des températures particulièrement basses ou élevées ont un effet mesurable sur les retards des trains ? L'analyse de ces variables nous permettra de vérifier si la température influence la ponctualité des trajets. Enfin, cette étape d'analyse sera aussi l'occasion de continuer le nettoyage des données. En identifiant les anomalies, telles que des échantillons non représentatifs ou des jours particuliers présentant des comportements atypiques, nous pourrons affiner les ensembles de données pour les rendre plus cohérents et adaptés à l'entraînement des modèles.

4.3.1 Relation entre les retards et les variables d'exploitation (XOP)

Le tableau 4.2 présente les statistiques des retards pour les ensembles de données du tramway et du RER. Pour une journée typique, chaque train ou tramway peut connaître un retard pour les départs à chaque station. La moyenne des retards observés est de 101,05 secondes pour les données du tramway, tandis que pour le RER, la moyenne est significativement plus élevée, à 228,46 secondes. En termes de variabilité des retards, l'écart-type, qui mesure la dispersion des retards autour de la moyenne, est de 334,92 secondes pour le tramway et de 291,42 secondes pour le RER. Cela indique que les retards du tramway sont plus dispersés que ceux du RER. Les valeurs aberrantes sont également mises en évidence dans ce tableau avec des retards minimum et maximum. Le tramway peut avoir des avances allant jusqu'à -1993 secondes (environ 33 minutes d'avance) et des retards allant jusqu'à $+5000$ secondes (environ $1h23$ de retard). Pour le RER, les avances atteignent -878 secondes (environ 15 minutes d'avance) et un retard maximum de $+4024$ secondes (environ $1h07$ de retard). Ces statistiques permettent de mieux comprendre la distribution des retards dans les deux réseaux, en identifiant non seulement les retards typiques mais aussi les valeurs extrêmes. L'objectif de ces statistiques est de fournir une vision claire des écarts de performance pour chaque réseau, ce qui permet de mieux évaluer la fiabilité du service et de comprendre les anomalies possibles dans les retards.

Les deux figures, 4.9 et 4.10, présentent des boîtes à moustaches illustrant la distribution des retards pour chaque jour de la semaine. La première différence notable entre les deux graphiques réside dans le fait que la figure 4.10 (concernant le RER) n'inclut pas les retards des week-ends (samedi et dimanche), tandis que la figure 4.9 (concernant le Tramway) couvre toute la semaine. Malgré cette différence, une similarité ressort entre les jours de la semaine pour les deux lignes (Tramway et RER). En effet, il n'y a pas de grande différence visible entre les distributions des retards du lundi au vendredi, les médianes sont relativement proches et les boîtes sont d'une taille comparable, ce qui indique que les retards ont une répartition similaire pendant ces jours. Cependant, la figure 4.9 présente une particularité concernant le samedi, qui affiche une boîte plus petite, suggérant une plus faible variabilité des retards par rapport

aux autres jours de la semaine, bien que la médiane reste comparable. Le dimanche quant à lui constitue l'exception notable : la médiane des retards est plus élevée, ce qui pourrait indiquer des perturbations supplémentaires ou des changements dans les opérations, tels qu'un nombre réduit de services ou des perturbations spécifiques à ce jour. Ces observations peuvent être attribuées à la présence de valeurs aberrantes (outliers), qui influencent les distributions et créent ces variations d'un jour à l'autre.

Indicateur	**Tramway**	**RER**
Moyenne des retards	101.0461	228.4583
Écart-type des retards	334.9182	291.4153
Min	-1993.0	-878.0
Max	5000.0	4024.0

Table 4.2 – Statistiques des retards et valeurs aberrantes pour Tramway et RER

La distribution des moyennes des retards en fonction des jours est présentée dans les figures 4.11 et 4.12. Ces figures montrent les moyennes observées ainsi que l'identification des pics de retards, signalés par des points rouges. On observe que les pics de retards ne se limitent pas uniquement aux weekends, comme montré dans la figure 4.9, où la médiane la plus élevée est enregistrée le dimanche. Cependant, la figure 4.11 révèle que certains jours de la semaine présentent également des anomalies, comme le vendredi, qui montre un nombre important d'occurrences de retards élevés. Bien que la médiane du vendredi, dans la boîte à moustache, soit comparable à celle des autres jours, cette boîte présente des valeurs aberrantes considérables par rapport aux autres jours. De plus, il est intéressant de noter que des jours comme le jeudi montrent également des moyennes élevées, même si cela n'apparaît pas comme une anomalie dans la boîte à moustaches du jeudi. Cela s'explique par un phénomène de compensation des retards, par exemple, le jeudi 09-11-2017 présente une moyenne de retard de 864 secondes, tandis qu'un autre jeudi, le 28-06-2018, montre une avance de 13 secondes. Ce phénomène de compensation des retards pour certains jours de la semaine permet de comprendre pourquoi certaines anomalies ne sont pas visibles dans la boîte à moustache.

Un phénomène similaire est observé pour les jours de la semaine dans la figure 4.10. Les boîtes à moustaches du lundi et du mardi montrent des médianes normales, comparables à celles des autres jours. Cependant, dans la figure 4.12, on trouve des occurrences significatives pour ces deux jours sur les trois années (2017, 2018 et 2019), avec des pics de moyennes qui reflètent des valeurs aberrantes dans leurs boîtes à moustaches respectives. Enfin, la boîte à moustaches du vendredi montre des valeurs aberrantes plus importantes que les autres jours, ce qui s'explique également par le phénomène de compensation des retards mais dans la journnée. Par exemple pour le même jour, un retard important peut atteindre 1000 secondes, mais en parallèle, une avance pourrait compenser en atteignant 500 secondes, ce qui expliquerait cette perturbation dans les données. Toutes ces observations permettent d'analyser l'impact des types de jours sur les retards et d'identifier les jours non représentatifs dans les données (voir le tableau 4.3) afin de les éliminer dans les analyses futures.

La notion des heures de pointe et des heures creuses est bien présente dans les données du RER, mais elle est absente dans les données du tramway. La figure 4.13

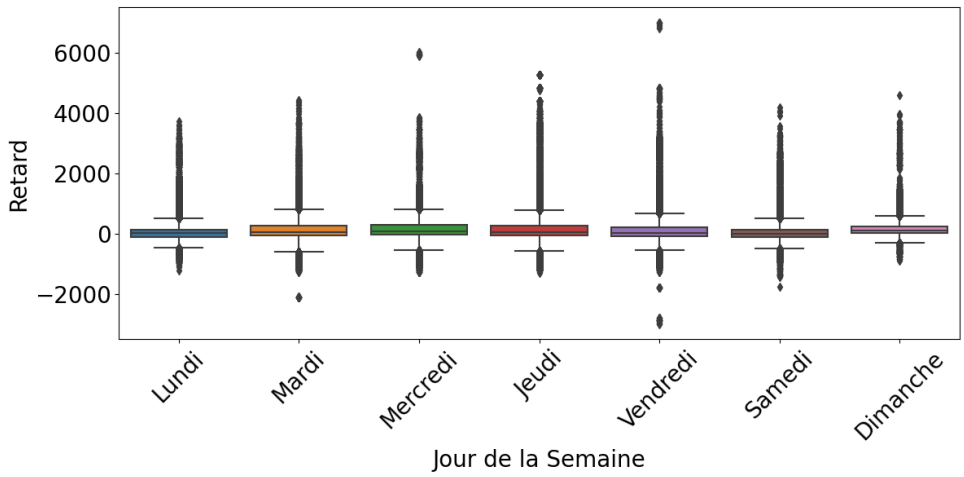


Figure 4.9 – Distribution des retards par jour de la semaine sur la ligne Tramway avec boîte à moustaches .

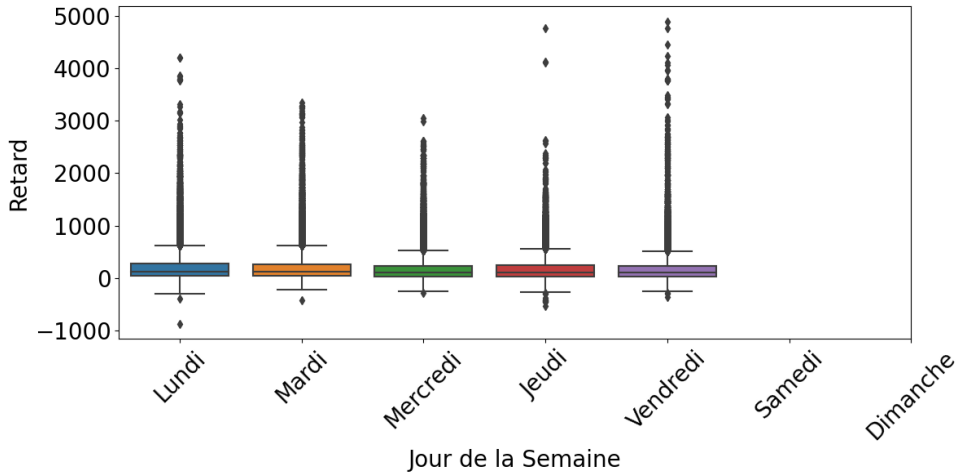


Figure 4.10 – Distribution des retards par jour de la semaine sur la ligne RER avec boîte à moustaches .

présente l'évolution des moyennes des retards par heure. La courbe verte, qui illustre les retards en fonction des heures pour les données RER, met en évidence des pics de retard aux heures de pointe : entre 8h et 10h le matin, ainsi qu'entre 17h et 21h le soir. En revanche, la courbe bleue représentant les données du tramway ne montre pas de pics de retard marqués. Les moyennes des retards par heure sont relativement fluides. Cependant, on observe une légère augmentation des retards en fin de journée. Cela pourrait s'expliquer par le comportement des voyageurs qui sortent davantage en soirée, dans le contexte d'une ville touristique. Cette spécificité fait que la notion d'heures de pointe et creuses n'est pas aussi marquée pour le réseau tramway que pour le RER.

La distribution des retards moyens pour les stations, présentée pour les deux ensembles de données du tramway et du RER, est illustrée dans les figures 4.14 et 4.15 par une courbe de moyenne en bleu. On observe qu'il n'y a pas de retards majeurs ni de pics dépassant la moyenne des retards, en tenant compte de l'intervalle de marge d'erreur présenté par l'équation 3.18. Le fait que les pics ne dépassent pas la zone colorée dans les deux graphiques indique que les retards sont normaux et bien répartis

pour l'ensemble des stations. Les figures 4.16 et 4.17 présentent la corrélation entre la variable expliquée, qui est le retard, et les variables explicatives pour les données du tramway et du RER respectivement. On observe une corrélation très faible pour les deux ensembles de données d'exploitation 3.2.4. Ce constat nous a conduit à l'idée d'intégrer l'historique des retards comme variable explicative supplémentaire afin d'améliorer la qualité des prévisions. Le fait que les données actuelles montrent une liaison faible avec le phénomène de retard étudié justifie cette approche par la suite 4.3.2.

4.3.2 Impact des variables passées sur l'analyse des retards (Y^{past})

La notion de "passé" est représentée par l'ensemble des données décrites dans l'équation 3.11, qui regroupe les retards des rames prédécesseuses et successeuses d'une rame dans une fenêtre temporelle d'une heure dans le passé. Cet ensemble de retards varie selon l'infrastructure étudiée, comme c'est le cas dans notre étude. Par exemple, dans la fenêtre d'une heure, on trouve environ 80 rames avec leurs retards pour le tramway, tandis que pour le RER, ce nombre s'élève à presque 100 rames. Ces retards sont intégrés dans l'ensemble des variables explicatives, représentant les retards précédents, comme illustré dans les figures 4.18 et 4.19 respectivement. L'analyse des corrélations observées dans ces dernières démontre l'importance de cette approche. En effet, la corrélation atteint $0,8$ pour les données du tramway et environ $0,6$ en moyenne pour le RER. Cette différence s'explique par le fait que la ligne ferroviaire du RER n'est pas complète dans l'ensemble de données, ce qui signifie que nous ne disposons pas de tout l'historique, impactant ainsi la corrélation observée. Cela reste néanmoins une amélioration qui permet d'obtenir des résultats significativement meilleurs par rapport à l'utilisation exclusive des données d'exploitation, lesquelles, comme montré dans les figures 4.16 et 4.17, présentent une corrélation plus faible. L'utilisation de l'ensemble des retards précédents dans les modèles d'apprentissage dépend de chaque modèle présenté dans la section 2.3.4. En effet, certains modèles d'apprentissage automatique sont mieux adaptés à l'intégration de données temporelles, comme les modèles séquentiels ou les réseaux neuronaux récurrents, tandis que d'autres, comme les arbres de décision, peuvent nécessiter des transformations supplémentaires des données pour exploiter pleinement les retards passés. Cette flexibilité permet d'adapter l'approche en fonction des caractéristiques spécifiques du modèle et des performances souhaitées.

Nous poursuivons le nettoyage des données en analysant la corrélation et l'importance des retards précédents pour les jours de chaque ensemble de données, afin de sélectionner les journées exceptionnelles ou non représentatives. Cette étape est cruciale pour permettre à nos modèles de converger correctement. Pour illustrer ce processus, les figures 4.20 et 4.21 présentent la corrélation entre les retards et la moyenne des retards précédents pour le tramway et le RER respectivement. Deux courbes sont tracées : la courbe bleue représente les jours importants, c'est-à-dire les jours représentatifs dans les données, avec $70,13\%$ pour le tramway et $40,59\%$ pour le RER, comme indiqué dans le tableau 4.3. La courbe rouge, quant à elle, représente les jours non représentatifs, avec $29,87\%$ pour le tramway et $59,41\%$ pour le RER. Ces proportions montrent que l'ensemble de données du RER contient un pourcentage plus élevé de jours non représentatifs. Cela s'explique par le fait que les données du RER couvrent le premier trimestre de chaque année, une période souvent marquée par

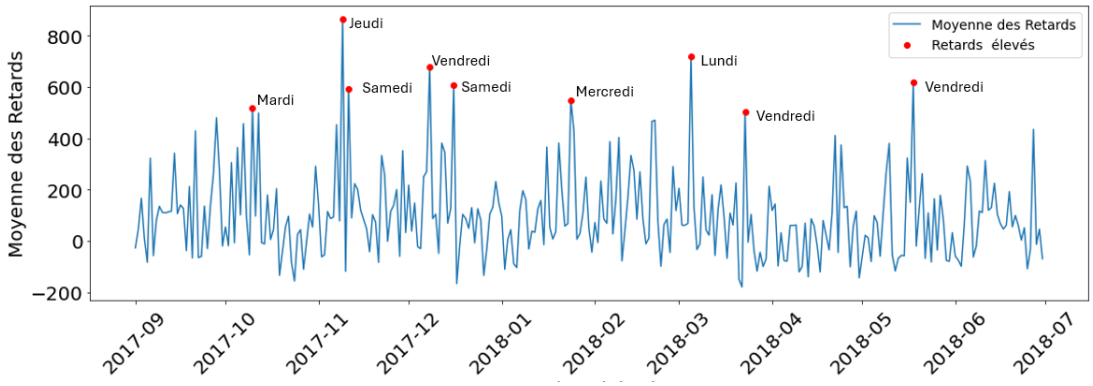


Figure 4.11 – Moyenne des retards quotidiens et identification des retards importants dans les données du tramway.

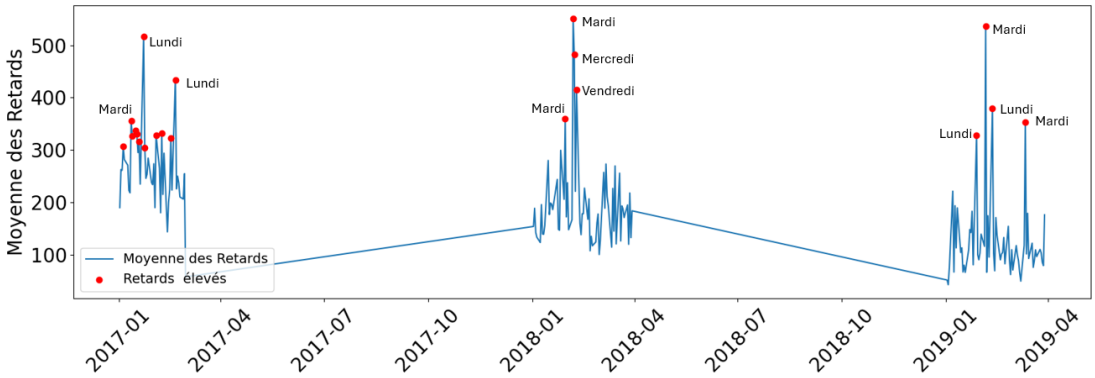


Figure 4.12 – Moyenne des retards quotidiens et identification des retards importants dans les données du RER.

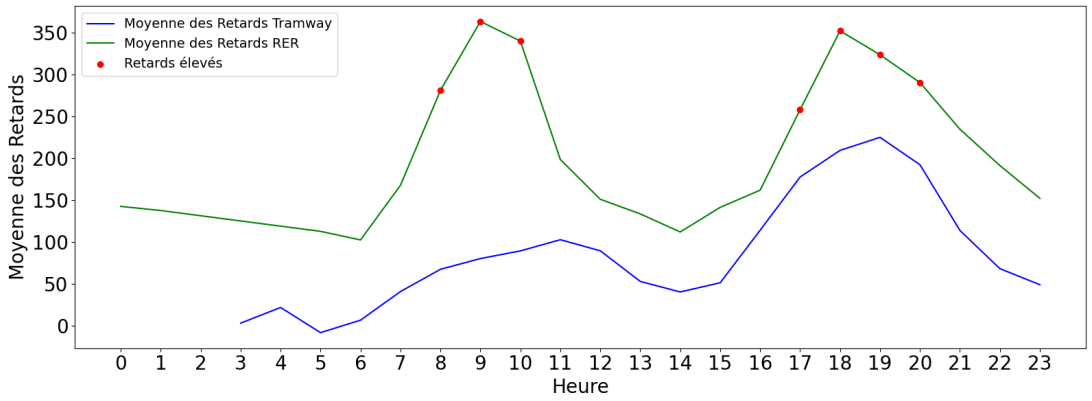


Figure 4.13 – Évolution des retards moyens par heure et identification des retards importants.

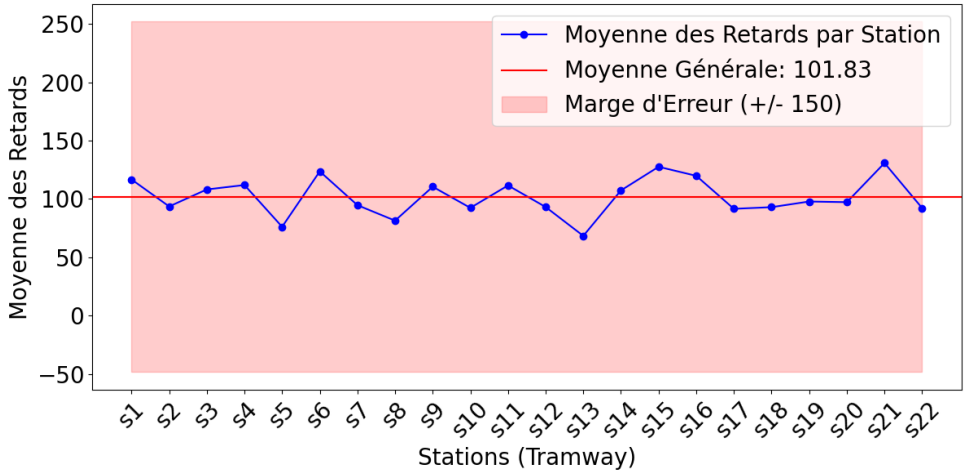


Figure 4.14 – Moyenne des retards par station de Tramway avec marge d'erreur de ±150

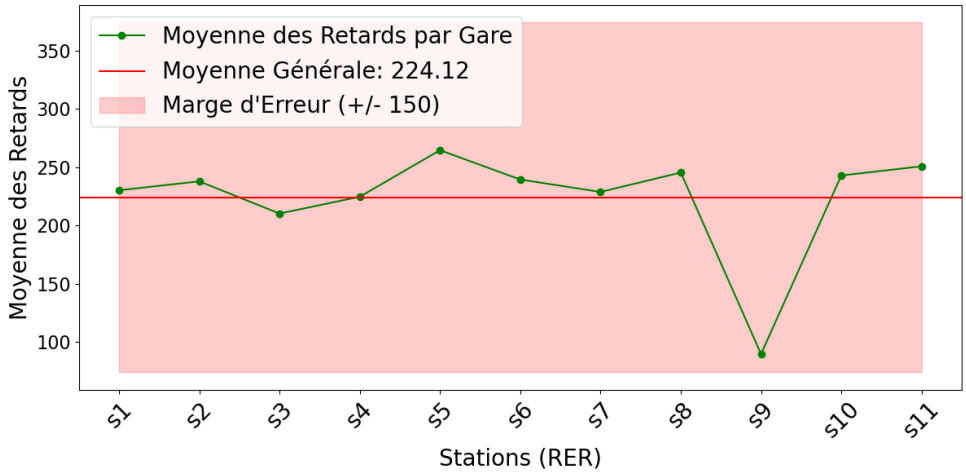


Figure 4.15 – Moyenne des retards par station de RER avec marge d'erreur de ±150

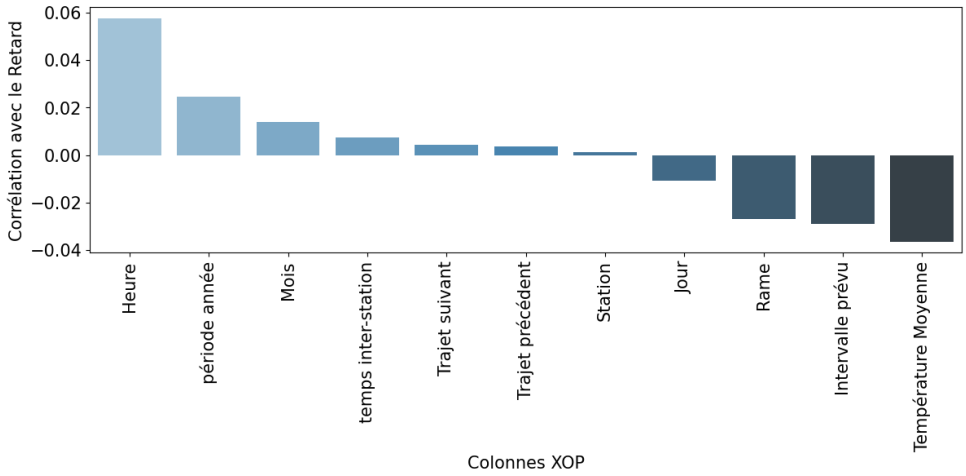


Figure 4.16 – Corrélation entre la variable expliquée "retard" et les variables explicatives représentant les données d'exploitation "XOP" pour le tramway

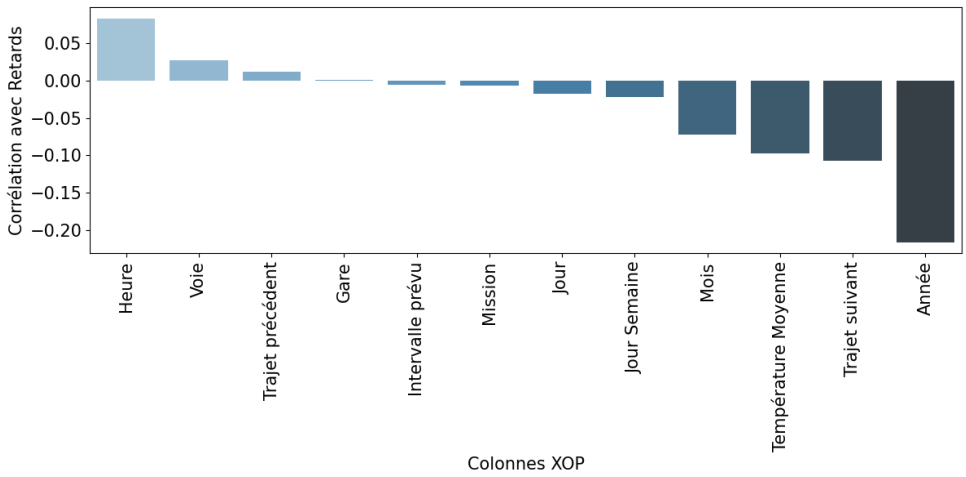


Figure 4.17 – Corrélation entre la variable expliquée "retard" et les variables explicatives représentant les données d'exploitation "XOP" pour le RER

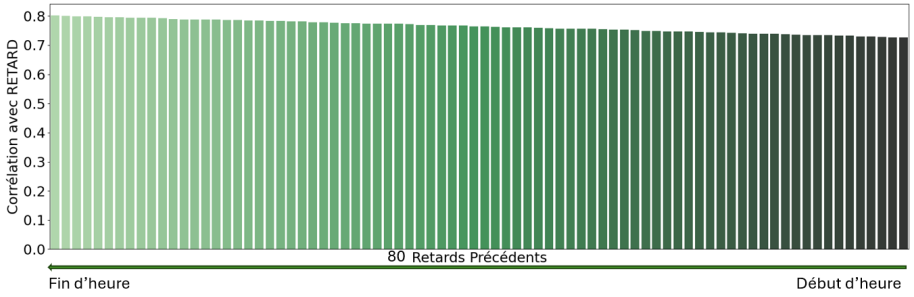


Figure 4.18 – Corrélation entre la variable expliquée "retard" et les 80 retards précedents pour le tramway

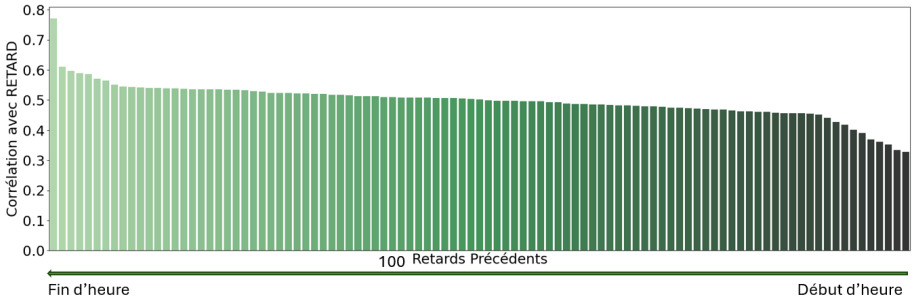


Figure 4.19 – Corrélation entre la variable expliquée "retard" et les 100 retards précedents pour le pour le RER

des problèmes spécifiques, des journées exceptionnelles, ainsi que des conditions climatiques défavorables, comme le froid ou les intempéries. De plus, la ligne ferroviaire du RER est incomplète dans l'ensemble des données, ce qui limite la disponibilité de l'historique et impacte ainsi les résultats.

Taux Statistique(%)	**Tramway**	**RER**
Trajet complet	91.91	68,08
Trajet non complet	8.79	31,92
Jours représentatifs	70.13	40.59
Jours non représentatifs	29.87	59.41
Heures corrélées	66.67	71.43
Heures non corrélées	23.81	28.57
Stations corrélées	100	81.82
Stations non corrélées	0	18.18

Table 4.3 – Les taux statistiques pour Tramway et RER.

Parmi les facteurs pouvant influencer la corrélation entre le retard actuel et les retards précédents, figure la présence de trajets incomplets dans l'ensemble des données. Ce phénomène est plus fréquent dans les données du RER que dans celles du tramway, avec $8,79\%$ de trajets incomplets pour le tramway et $31,92\%$ pour le RER, comme indiqué dans le tableau 4.3. Cette situation peut également affecter la corrélation, que ce soit de manière quotidienne ou globale. La figure 4.22 illustre ce

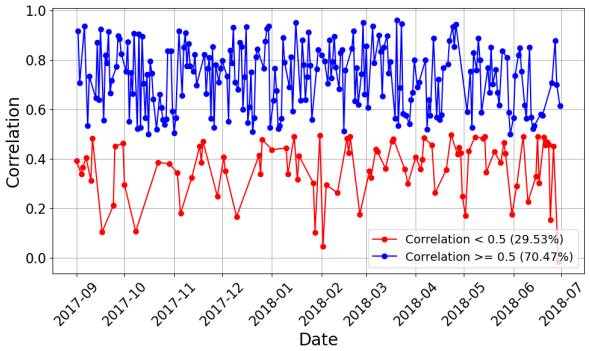


Figure 4.20 – Classification des jours en fonction de la corrélation entre retards historiques et anticipés pour le Tramway

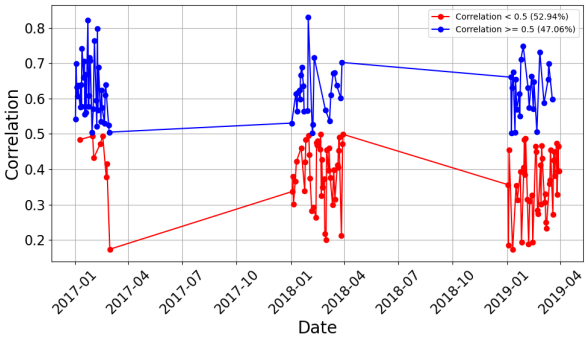


Figure 4.21 – Classification des jours en fonction de la corrélation entre retards historiques et anticipés pour le RER

phénomène, montrant des trajets incomplets effectués par une rame au cours d'une journée. Cela peut résulter soit d'anomalies lors de la collecte des données, soit des opérations de nettoyage décrites dans la section 4.2. Par exemple, la figure 4.22 présente le cas de la rame r_{28}, qui s'arrête à la station s_{15} avant la fin de son trajet, ce qui réduit la pertinence des retards précédents. Lorsqu'une journée contient de nombreux trajets incomplets, cela engendre des lacunes dans les données, nuisant à la qualité des corrélations. Ce problème est particulièrement présent dans les données du RER, expliquant la forte proportion de journées non représentatives, soit $59,41\%$, comme mentionné précédemment.

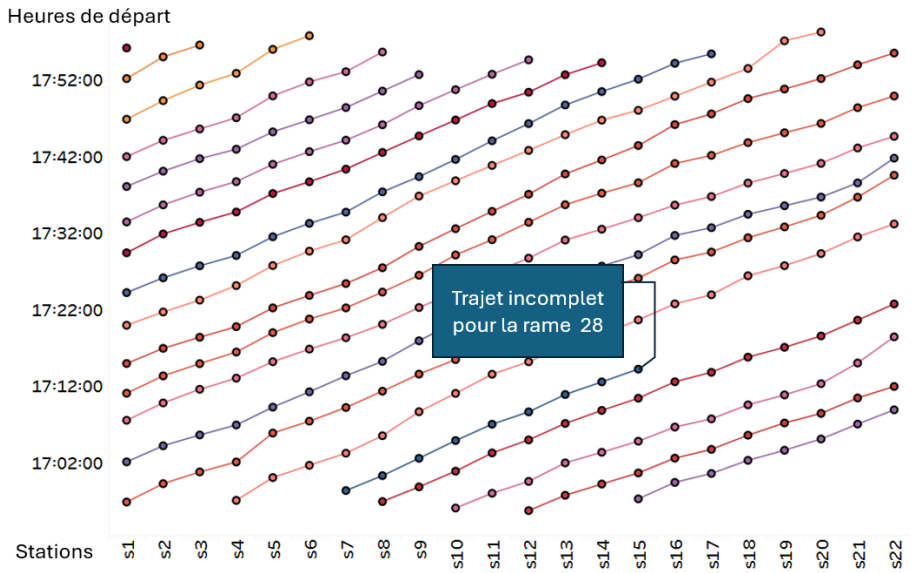


Figure 4.22 – Courbes de trajets pour différentes rames, avec un trajet incomplet visible pour la rame 28.

L'analyse réalisée sur l'importance des variables en lien avec notre phénomène d'étude, à savoir les retards dans les deux ensembles de données présentés jusqu'à présent, a pour objectif d'évaluer l'impact des variables proposées dans notre méthodologie sur les retards. Nous avons constaté que les retards ont une faible corrélation avec les données d'exploitation (3.2.4) et une corrélation plus forte avec les données historiques, notamment les retards précédents des trains prédécesseurs et succes-

seurs. Cela indique que les retards passés influencent significativement les retards futurs, ce qui contribue à une meilleure convergence des modèles. Une anomalie a été détectée concernant les jours non représentatifs dans les deux ensembles de données. Cela pourrait affecter les résultats, c'est pourquoi, dans chaque phase de test, nous proposons d'effectuer deux évaluations : une avec les jours non représentatifs et les trajets incomplets, et une autre sans ces anomalies, afin d'évaluer leur impact sur la précision des modèles d'apprentissage présentés dans la section 2.3.4. Cette étape permettra de comparer différents types de modèles testés dans la section suivante 4.4, et ainsi valider le modèle d'apprentissage le plus adapté aux deux ensembles de données. De plus, cela nous aidera à déterminer si notre méthodologie est applicable dans les deux cas, et si nous devons conserver ou supprimer les trajets incomplets et les jours non représentatifs.

La section suivante présente une étude comparative des données de l'infrastructure tramway afin de déterminer le modèle le plus pertinent. Cette comparaison se base sur divers critères de performance et de précision des prédictions, permettant d'évaluer l'efficacité des différents modèles appliqués à cet ensemble de données. L'objectif est de sélectionner la méthode la mieux adaptée pour fournir des prévisions fiables des retards, en tenant compte des spécificités propres à l'infrastructure tramway.

4.4 La prévision à court terme du retard : une première analyse exploratoire (tramway)

Cette section présente les résultats obtenus pour chaque modèle décrit dans notre méthodologie (section 2.3.4), ainsi que l'évaluation détaillée de leur performance. L'objectif est de déterminer si tous les modèles parviennent à converger avec les caractéristiques présentées dans le tableau 3.9, préparées dans la section précédente 4.3, en se concentrant dans un premier temps sur les données Tramway. Pour cela, nous avons utilisé les hyperparamètres spécifiques à chaque modèle d'apprentissage afin d'évaluer leur efficacité et leur capacité à traiter ces données, dans le but de sélectionner le ou les modèles les plus performants selon notre méthodologie. Les données RER, quant à elles, seront abordées dans la section 4.5 pour étudier la sensabilité de notre méthodologie.

4.4.1 Partitionnement des données pour l'entraînement et la validation : Tramway

La figure 4.23 illustre la répartition effectuée sur l'ensemble des données du tramway, mettant en évidence une division des données en deux parties distinctes : 80 % des données, soit 242 jours, sont dédiées à l'entraînement, tandis que les 20 % restants, soit 61 jours, sont réservés pour les tests. Ces données concernent le tramway et servent à évaluer les performances du modèle. Les colonnes vertes représentent les données qualitatives, tandis que la colonne orange, l'heure de départ réelle qui nécessite une ingénierie des caractéristiques afin d'extraire des informations telles que le jour, le mois, l'année, l'heure, et les minutes. Pour les autres colonnes contenant des données quantitatives, une méthode de normalisation(eq.3.8) est appliquée.

Pour chaque modèle d'apprentissage, nous avons besoin des données d'entraînement $(X_{\text{entraînement}})$ et des sorties correspondantes avec un horizon temporel de

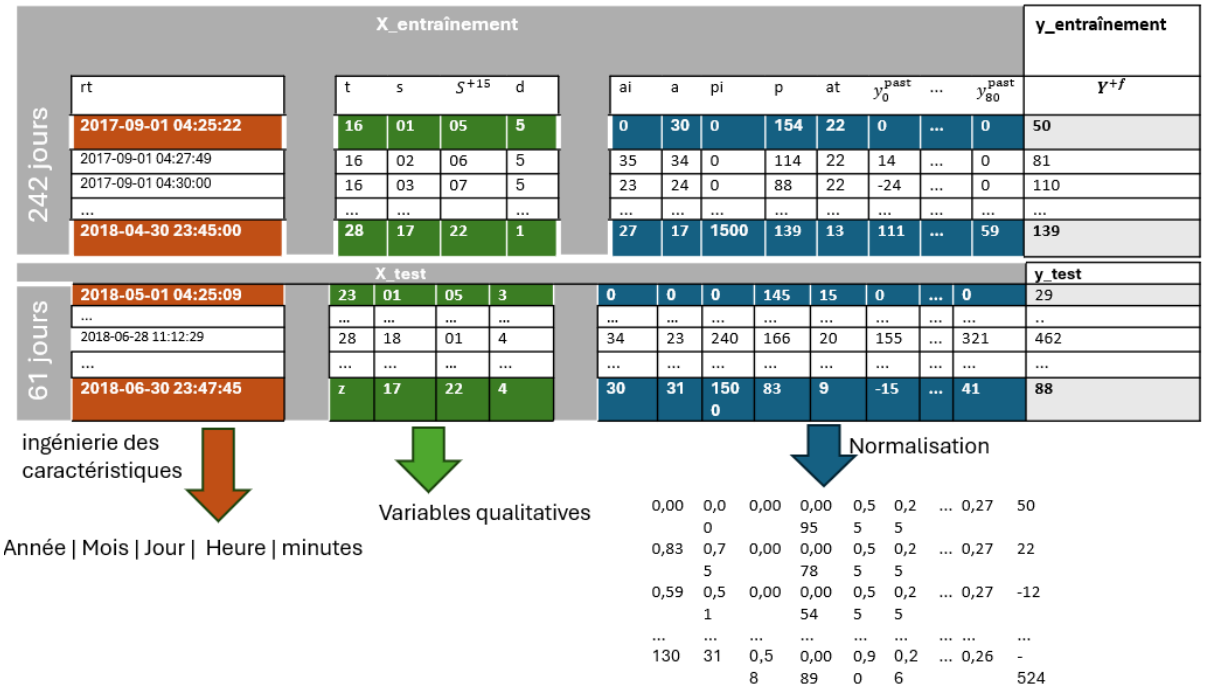


Figure 4.23 – Répartition des données en entraînement, test et catégorisation (quantitative et qualitative)

15 minutes dans le futur ($y_entraînement$). Pour l'évaluation, les données de test (X_test, y_test) sont utilisées afin de comparer les sorties du modèle avec les prédictions. L'objectif est de choisir un modèle qui converge bien pour les données du tramway en tenant compte des interactions entre le passé, le présent et le futur. Cette répartition est appliquée à chaque modèle mentionné dans la section 3.6. La seule différence réside dans l'utilisation de la validation croisée, appliquée de manière récursive (Algorithme 3), et dans le nombre de retards précédents pris en compte pour chaque modèle avec l'agorthime 4. De plus, chaque modèle possède un ensemble spécifique d'hyperparamètres à ajuster.

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{4.1}$$

Où :

- x : La valeur que vous souhaitez normaliser,
- x_{min} : La plus petite valeur possible dans l'ensemble de données,
- x_{max} : La plus grande valeur possible dans l'ensemble de données.

4.4.2 Méthodes d'entraînement

Chaque modèle d'apprentissage décrit dans la section 3.6 commence par une sélection des variables optimales, effectuée grâce à l'algorithme 4. Cet algorithme permet de déterminer les variables les plus importantes pour chaque modèle. , dont les résultats sont présentés dans le tableau 4.4, en indiquant les variables sélectionnées. Comme le montre cette table, tous les modèles utilisent des données d'exploitation futures, mais les modèles statiques prennent en compte un nombre limité de données d'exploitation et de données passées. Cela s'explique par leur nature, qui privilégie les

relations linéaires entre les variables, ce qui est inadapté dans notre contexte. Bien que ces modèles nécessitent peu de données pour capter les relations simples, leur performance diminue lorsque les relations entre variables sont plus complexes ou non linéaires. En revanche, les modèles basés sur des arbres de décision sont capables de modéliser des relations non linéaires et des interactions complexes entre variables, même en l'absence de corrélations fortes. Les méthodes d'apprentissage automatique profond offrent une capacité encore plus avancée à capturer des relations complexes entre les variables. Les modèles comme les réseaux de neurones artificiels ANN, ainsi les LSTM et BILSTM, sont particulièrement adaptés à la modélisation de séquences temporelles et d'interactions non linéaires à long terme. Ces modèles peuvent non seulement capturer les dépendances complexes dans les données, mais aussi traiter des volumes de données massifs avec des performances améliorées.

Les tests sur le nombre de plis à utiliser pour chaque modèle, effectués à l'aide de l'algorithme 3, sont présentés dans le tableau 4.4. Cet algorithme explore un nombre de plis compris entre 1 et 1000 afin de déterminer le pli optimal pour chaque modèle. Le choix du nombre de plis optimal dépend de la complexité de chaque modèle et de sa capacité à capturer les relations entre le passé, le présent et le futur. Les modèles statiques, tels que la régression linéaire et la régression Bayesian Ridge, ne nécessitent aucun pli (0 pli) car ils se basent sur des relations linéaires directes entre les variables du passé et les valeurs actuelles et futures. Cela signifie que ces modèles considèrent les données du passé comme suffisantes pour établir une relation linéaire avec le présent et le futur. En raison de cette simplicité, ces méthodes n'ont pas besoin de validation croisée étendue, car elles ne tentent pas de modéliser des relations complexes. les modèles à base d'arbres, tels que les arbres de décision, les forêts aléatoires et XGBOOST, requièrent un plus grand nombre de plis (généralement entre 80 et 120) pour modéliser des relations non linéaires complexes et des interactions multiples entre les variables. Ces modèles nécessitent une validation croisée plus fine pour bien évaluer leur performance et réduire le risque de surapprentissage. Enfin, les modèles d'apprentissage profond ANN, LSTM et BILSTM, demandent un nombre encore plus élevé de plis (jusqu'à 150) en raison de leur capacité à modéliser des dépendances temporelles à long terme et des relations non linéaires complexes. Grâce à leur architecture hiérarchique, ils prennent en compte les relations entre le passé et le présent sur des périodes plus longues, nécessitant ainsi une validation croisée approfondie pour assurer une généralisation robuste.

Table 4.4 – Exploration des modèles de machine learning et paramètres associés

Méthodes	**Modèle**	**Sortie de l'algorithme4** V_{opt}	**Liste des plis à tester**	**Sortie de l'algorithme 3** K_{opt}
Statique	Régression linéaire	$(XOP_{\{RT,PI,AT\}}, XOP^{+}, y_{past}^{30-80})$	1 à 1000	0
	Régression Bayésienne	$(XOP_{\{RT,PI,AT\}}, XOP^{+}, y_{past}^{35-80})$	1 à 1000	0
À base des arbres	Arbres de décision	$(XOP, XOP^{+}, y_{past}^{20-80})$	1 à 1000	80
	Forêts aléatoires	$(XOP, XOP^{+}, y_{past}^{15-80})$	1 à 1000	99
	XGBOOST	$(XOP, XOP^{+}, y_{past}^{0-80})$	1 à 1000	120
Apprentissage automatique profond	ANN	$(XOP_{\{RT,S,PI,AT\}}, XOP^{+}, y_{past}^{50-80})$	1 à 1000	99
	LSTM	$(XOP, XOP^{+}, y_{past}^{60-80})$	1 à 1000	150
	BILSTM	$(XOP, XOP^{+}, y_{past}^{60-80})$	1 à 1000	100

La section suivante présente l'application des algorithmes d'optimisation pour ajuster les hyperparamètres de chaque modèle d'apprentissage, qu'il s'agisse des modèles basés sur les arbres ou des modèles d'apprentissage automatique profond. L'objectif est de déterminer la combinaison d'hyperparamètres offrant les meilleurs résultats sur l'ensemble d'entraînement.

4.4.3 Ajustement des hyperparamètres

La sélection des hyperparamètres des modèles est une étape essentielle pour améliorer les performances et garantir la convergence du modèle. Parmi les modèles d'apprentissage, on distingue les modèles basés sur les arbres et les modèles d'apprentissage automatique profond, qui comportent de nombreux hyperparamètres à ajuster et à choisir.

L'objectif de l'optimisation des hyperparamètres est de trouver la meilleure combinaison qui offre les résultats les plus performants sur l'entraînement. Deux algorithmes, présentés respectivement dans les algorithmes 1 et 2, sont utilisés à cet effet. Le premier algorithme, dédié aux modèles basés sur les arbres, vise à identifier les valeurs optimales des hyperparamètres, comme présenté précédemment dans le tableau 2.2. Le second algorithme, quant à lui, est conçu pour optimiser les hyperparamètres des modèles d'apprentissage profond, dont les détails sont également décrits dans le tableau 2.3.

Les tableaux 4.5 et 4.6 présentent respectivement les espaces de recherche des hyperparamètres pour les modèles basés sur des arbres de décision et les modèles d'apprentissage profond. Pour les modèles basés sur des arbres de décision, l'algorithme 1 (Bayesian Optimization) a été utilisé afin de déterminer les hyperparamètres optimaux. Les résultats de cette optimisation montrent que le choix optimum de ces hyperparametres joue un rôle clé dans la performance. Pour les modèles d'apprentissage profond, l'algorithme 2 (Grid Search) a été employé pour explorer l'espace des hyperparamètres. Les résultats montrent que les paramètres tels que le taux d'apprentissage (*learning_rate*) et le nombre d'epochs ont un impact significatif sur la performance. Les réseaux neuronaux tels que LSTM et BILSTM nécessitent également une configuration optimale du nombre de couches et de la fonction d'activation.

À l'issue de chaque optimisation, un ensemble d'hyperparamètres est retourné, correspondant aux meilleures performances observées sur l'ensemble de données défini. Le temps de calcul pour l'optimisation des hyperparamètres varie selon le type de modèle. En moyenne, pour les modèles basés sur des arbres de décision, l'optimisation prend environ cinq jours, tandis que pour les modèles d'apprentissage profond, le calcul est plus long, nécessitant environ huit jours. Ce délai plus important est attribuable à la complexité des architectures des modèles d'apprentissage profond, qui requièrent une exploration plus exhaustive de l'espace des hyperparamètres.

La section suivante présente les résultats obtenus pour les modèles d'apprentissage testés avec les hyperparamètres optimaux sélectionnés.

4.4.4 Performance et convergence des modèles sur l'ensemble d'entraînement

Les trois optimisations effectuées au cours de ce travail comprennent la sélection des variables optimales, la détermination du nombre de plis pour la validation croisée, ainsi que l'optimisation des espaces des hyperparamètres. Ces étapes, présentées dans les tableaux 4.4, 4.5 et 4.6, respectivement, sont essentielles pour garantir un entraînement efficace des modèles et assurer leur convergence. La méthode utilisée pour analyser la performance des modèles est décrite en détail dans la section 3.7, où une comparaison approfondie entre les différents modèles est réalisée. Le tableau 4.7 présente les performances des différents modèles de prédiction des retards sur

Table 4.5 – Exploration des hyperparamètres - À base des arbres

Modèle	**Hyperparamètres**	**Espace des hyperparamètres**	**Sortie de l'algorithme θ^{meilleur}**
Arbres de décision	`max_depth`, `min_samples_split`, `min_samples_leaf`, `max_features`	(3,100), (2, 100), (1, 100), (0.1,1)	(8, 15, 20, 0.71)
Forêts aléatoires	`max_depth`, `min_samples_split`, `min_samples_leaf`, `max_features`, `n_estimators`	(3,100), (2,100), (1,100), (0.1, 1.0), (50, 1000)	(10, 45, 32, 0.88, 900)
XGBOOST	`max_depth`, `n_estimators`, `subsample`, `learning_rate`, `colsample_bytree`	(3,100), (50, 1000), (0.1, 1.0), (0.001,1.0), (0.1, 1.0)	(8, 500, 0.98, 0.016, 0.94)

l'ensemble d'entraînement. Les métriques utilisées pour l'évaluation incluent le RMSE (Root Mean Square Error), le MAE (Mean Absolute Error), le coefficient R^2, ainsi que le pourcentage des prédictions se situant dans l'intervalle de [-150, 150] secondes, permettant ainsi une évaluation complète de la précision des modèles.

Les résultats montrent que les modèles basés sur les arbres de décision, tels que les Forêts Aléatoires et XGBOOST, offrent des performances nettement supérieures. Le modèle XGBOOST atteint des scores très élevés avec un R^2 de 0.99 et un pourcentage de 99.08% des prédictions dans l'intervalle [-150, 150], ce qui signifie que presque toutes les prédictions sont très proches des valeurs réelles. De manière similaire, les Forêts Aléatoires montrent un RMSE très bas de 35.65 et un excellent R^2 de 0.98, confirmant leur efficacité pour capturer les relations complexes dans les données d'entraînement. En revanche, les modèles plus simples, tels que la Régression Linéaire et la Régression Bayésienne, obtiennent des performances relativement modestes avec un RMSE de 149.50 et un R^2 de 0.67. Bien qu'ils soient capables de fournir des prédictions correctes pour environ 80% des observations, ces modèles ne capturent pas les variations les plus complexes dans les données, ce qui limite leur précision globale. Les modèles d'apprentissage profond, notamment LSTM et BILSTM, affichent des performances intermédiaires, avec un R^2 autour de 0.75 et des pourcentages de prédictions précises de 86.26% et 85.87% respectivement. Ces résultats montrent que ces modèles sont bien adaptés à la gestion des séries temporelles, capturant efficacement les dépendances temporelles entre les retards, tout en ayant une certaine capacité de généralisation.Le modèle d'Arbres de Décision simple, quant à lui, se distingue par une performance nettement inférieure avec un RMSE de 207.56 et un R^2 de seulement 0.37, ce qui suggère une capacité limitée à capturer les complexités des retards, peut-être en raison de la tendance à surapprendre les données d'entraînement.

Certains modèles, tels que XGBOOST ou les Forêts Aléatoires, atteignent un R^2 proche de 0.99, ce qui indique une excellente performance sur les données d'entraînement. Ce résultat peut être attribué à plusieurs facteurs. Tout d'abord, ces modèles sont capables de capturer des relations non linéaires complexes entre les variables, grâce à leur structure qui combine plusieurs arbres de décision. Contrairement aux modèles plus simples comme les régressions linéaires, ces algorithmes exploitent efficacement

Table 4.6 – Exploration des hyperparamètres - Apprentissage profond

Modèle	**Hyperparamètres**	**Espace des hyperparamètres**	**Sortie de l'algorithme** $2\ \mathcal{H}^{\text{meilleur}}$
ANN	`batch_size`, `epochs`, `optimizer`, `learning_rate`, `dropout_rate`, `layers`, `init`, `activation`	(10, 500 avec un pas de 10), (50, 1000 avec un pas de 50), ({'adam', 'rmsprop'}), (0.001, 1 avec un pas de 0.01), (0.0, 0.5 avec un pas de 0.05), (2, 10 avec un pas de 1), (1, 500 avec un pas de 10 pour chaque couche), ({'relu', 'tanh'})	(31, 173, 'rmsprop', 0.19, 0.0748, 5, (320,380,160,160,60), 'relu')
LSTM	`batch_size`, `epochs`, `optimizer`, `learning_rate`, `dropout_rate`, `layers`, `init`, `activation`	(10, 500 avec un pas de 10), (50, 1000 avec un pas de 50), ({'adam', 'rmsprop'}), (0.001, 1 avec un pas de 0.01), (0.0, 0.5 avec un pas de 0.05), (2, 10 avec un pas de 1), (1, 500 avec un pas de 10 pour chaque couche), ({'relu', 'tanh'})	(64, 50, 'adam', 0.00108, 0.2, 3, (350, 160, 160), 'relu')
BILSTM	`batch_size`, `epochs`, `optimizer`, `learning_rate`, `dropout_rate`, `layers`, `init`, `activation`, `Lst_activation`	(10, 500 avec un pas de 10), (50, 500 avec un pas de 50), ({'adam', 'rmsprop'}), (0.001, 1 avec un pas de 0.01), (0.0, 0.5 avec un pas de 0.05), (2, 10 avec un pas de 1), (1, 500 avec un pas de 10 pour chaque couche), ({'relu', 'tanh'}), ({'linear'})	(260, 350, 'rmsprop', 0.00314, 0.16, 2, (260, 160), 'relu','linear')

les interactions entre les différentes variables d'entrée et les données historiques, réduisant ainsi l'erreur de prédiction. De plus, l'optimisation des hyperparamètres, en particulier dans des algorithmes comme XGBOOST, permet de trouver des paramètres tels que la profondeur maximale des arbres ou le taux d'apprentissage, qui maximisent la capacité du modèle à s'ajuster aux données d'entraînement. Toutefois, ce R^2 élevé doit être interprété avec prudence, car un score trop proche de 1 peut parfois indiquer un risque de surapprentissage (overfitting), ce qui peut limiter la capacité du modèle à bien généraliser sur des données de test non vues.

En revanche, les modèles d'apprentissage profond tels que LSTM, ANN et BILSTM n'atteignent pas un R^2 de 0.99 sur l'ensemble d'entraînement. Cela s'explique par plusieurs raisons. Premièrement, ces modèles sont conçus pour capturer les dépendances temporelles et les dynamiques séquentielles des données, ce qui les rend plus sensibles aux fluctuations des séries temporelles. Bien qu'ils soient très efficaces pour traiter des séries temporelles complexes, leur optimisation est plus difficile et requiert une exploration plus vaste des hyperparamètres, tels que le nombre de couches, la taille des neurones ou le taux d'apprentissage. De plus, ces modèles sont souvent régularisés

pour éviter le surapprentissage, ce qui réduit intentionnellement leurs performances sur l'ensemble d'entraînement afin d'améliorer leur capacité à généraliser sur des données non vues. Enfin, la convergence des modèles d'apprentissage profond peut être plus lente en raison de la complexité des architectures et de la nature stochastique de l'entraînement, empêchant ainsi d'atteindre un score parfait de R^2. Il est également possible que ces modèles nécessitent un volume de données plus important pour atteindre leur plein potentiel, car les modèles d'apprentissage profond sont généralement plus gourmands en données pour parvenir à une convergence stable et robuste.

Modèle	**RMSE**	**MAE**	**R²**	**Pourcentage [-150, 150]**
Régression Linéaire	149.5062	101.6920	0.6741	80.51%
Régression Bayésienne	149.5064	101.6947	0.6741	80.51%
Arbres de Décision	207.5667	94.8438	0.3719	79.98%
Forêts Aléatoires	35.6543	23.5249	0.9815	99.33%
XGBOOST	39.89	34.548	0.99	99.08%
ANN	129.7881	92.1497	0.7544	82.83%
LSTM	123.49	82.83	0.7593	87.26%
BILSTM	110.4639	75.5331	0.802	87.87%

Table 4.7 – Métriques de performance des modèles sur l'ensemble d'entraînement

La section suivante présente l'évaluation des modèles sur les données de test, accompagnée d'une comparaison des performances. Cette analyse vise à identifier le modèle le plus adapté à notre cas d'étude.

4.4.5 Evaluation et comparaison des modèles sur l'ensemble du test

Le tableau 4.8 présente les performances des différents modèles de prédiction des retards sur l'ensemble du test. Les résultats montrent une tendance similaire à ceux de l'entraînement, bien que les performances soient généralement légèrement inférieures. XGBOOST continue d'offrir des résultats remarquables avec un RMSE de 60.36 et un R^2 de 0.93, ce qui confirme sa capacité à généraliser efficacement. Le pourcentage de prédictions correctes dans l'intervalle est également très élevé à 97%, ce qui en fait le modèle le plus fiable pour la prédiction des retards sur les données de test. Le modèle Forêts Aléatoires reste également performant avec un RMSE de 98.94 et un R^2 de 0.83, ce qui indique une bonne généralisation. Il capture correctement environ 90.96% des prédictions dans l'intervalle, montrant qu'il est capable de gérer des données non vues avec un bon niveau de précision. Les modèles de Régression Linéaire et de Régression Bayésienne maintiennent des performances similaires à celles observées lors de l'entraînement, avec un RMSE d'environ 132 et un R^2 de 0.70. Ils atteignent également un pourcentage de prédictions précises d'environ 81.88%. Cela montre que bien qu'ils soient relativement simples, ces modèles peuvent fournir des prédictions raisonnablement fiables, mais sont moins capables de capturer les relations complexes entre les variables. En ce qui concerne les modèles d'apprentissage profond (ANN, LSTM, BILSTM), les résultats sont globalement satisfaisants. Le BILSTM montre un R^2 de

0.724 et un pourcentage de 85.87% de prédictions correctes, indiquant également une bonne généralisation. Le modèle LSTM quant à lui affiche également de bons résultats avec un R^2 de 0.73 et 87.14% de prédictions dans l'intervalle, confirmant sa capacité à capturer des dépendances temporelles tout en restant performant sur l'ensemble de test.

Modèle	RMSE	MAE	R²	Pourcentage [-150, 150]
Régression Linéaire	132.7926	95.2627	0.7019	81.88%
Régression Bayésienne	132.7942	95.2667	0.7018	81.88%
Arbres de Décision	237.9436	155.1433	0.0427	68.12%
Forêts Aléatoires	98.9414	67.1602	0.8345	90.96%
XGBOOST	60.36	41.0056	0.93	97%
ANN	127.8954	91.7623	0.7234	83.03%
LSTM	114.69	79.65	0.73	87.14%
BILSTM	115.6529	79.4241	0.7243	86.70%

Table 4.8 – Métriques de performance des modèles sur l'ensemble de test

La section suivante présente une analyse comparative des modèles d'apprentissage afin d'identifier celui qui est le plus adapté à notre cas d'étude.

4.4.6 Analyse comparative et discussion

Les résultats présentés dans les tableaux 4.7 et 4.8 montrent que les modèles Forêts Aléatoires et XGBOOST sont ceux qui convergent le mieux avec l'ensemble d'entraînement, atteignant un coefficient de détermination R^2 allant jusqu'à 0,93. La figure 4.24 illustre les temps d'entraînement de chaque modèle, sans optimisation des hyperparamètres. On observe que les modèles LSTM et BILSTM nécessitent le plus de temps, jusqu'à cinq jours, suivis par le modèle Forêts Aléatoires avec trois jours d'entraînement. Ensuite, les modèles ANN et XGBOOST requièrent chacun une journée d'entraînement. En revanche, les autres modèles ne dépassent pas trois heures d'entraînement. Selon les résultats obtenus, le modèle XGBOOST s'avère le plus adapté à notre situation, tant en termes de temps d'entraînement que de performances.

La figure 4.25 présente la moyenne des performances des différents modèles de prédiction des retards pour chaque heure de la journée sur l'ensemble des données de test, à travers quatre graphiques : RMSE, MAE, R^2 et le pourcentage $p(\%)$ dans l'intervalle [-150, 150]. Ces plots permettent de visualiser si les modèles conservent des performances cohérentes tout au long de la journée. L'exploitation commence à 5h00 du matin et il est visible que les modèles éprouvent des difficultés à maintenir des performances constantes. On observe une diminution notable des performances entre 5h00 et 8h00, ainsi qu'après 22h00. Le premier phénomène, en début de journée, peut s'expliquer par l'absence de données historiques (retards passés), car les premiers trains de la journée ne disposent pas d'informations passées pour alimenter les modèles. Cependant, XGBOOST parvient à maintenir une bonne performance, malgré cette absence de données historiques, probablement en raison de sa capacité à capturer des relations non linéaires dans les données d'exploitation. Quant à la détérioration

des performances en fin de journée, elle pourrait s'expliquer par une irrégularité accrue des rames, perturbant les modèles. Néanmoins, on remarque que XGBOOST reste le plus performant, minimisant le RMSE et le MAE tout en maximisant R^2 et $p(\%)$, aussi bien au début qu'à la fin de la journée. Par ailleurs, au milieu de la journée, on observe une augmentation du RMSE et du MAE pour tous les modèles par rapport au début et à la fin de la journée. Ce phénomène peut s'expliquer par un effet de cumul des retards, les retards étant en moyenne plus importants au milieu de la journée. Cela est dû à l'accumulation progressive des retards tout au long de la journée d'exploitation.

La figure 4.26 présente la moyenne des performances des différents modèles pour chaque jour des données de test. Tous les modèles affichent des perturbations à certains jours de cette période. Parmi eux, XGBOOST se distingue en offrant une meilleure stabilité au cours des journées, contrairement aux autres modèles qui peinent à maintenir cette stabilité. Cependant, même XGBOOST présente des difficultés à stabiliser ses performances lors de trois jours spécifiques des données de test. Les deux jours fériés, le 10 mai 2018 (Ascension) et le 21 mai 2018 (lundi de Pentecôte), sont particulièrement marqués par des irrégularités, avec une moyenne de RMSE à 170 secondes et un R^2 de 0,63. Ces jours fériés perturbent probablement les horaires habituels et les données historiques utilisées par les modèles, impactant leur capacité à prédire les retards avec précision. Le 28 juin 2018, bien qu'il ne s'agisse pas d'un jour férié, correspond à une période particulière à Nice avec l'événement "Innovative City", un forum international sur les solutions urbaines innovantes. Ce jour a également été marqué par des températures estivales avec un ciel voilé et des températures comprises entre 23 et 25 degrés. Aucun événement majeur tel que des grèves ou des perturbations n'a été signalé, bien que la ville ait connu des événements culturels et des animations régulières durant cette période. Malgré la robustesse de XGBOOST, on observe que le modèle peut rencontrer des difficultés à maintenir des performances stables lors de journées caractérisées par des irrégularités importantes dans les retards. Ce constat souligne la complexité de prédire avec précision les retards dans des conditions exceptionnelles ou peu régulières.

La figure 4.27 présente la moyenne des performances des différents modèles sur les 22 stations représentées dans les données de test. On observe que, quel que soit le modèle utilisé, les performances restent relativement stables d'une station à l'autre. Il n'y a pas de différences significatives entre les stations situées au début, au milieu, ou à la fin de la ligne, ce qui montre que les modèles parviennent à prédire de manière cohérente sur l'ensemble des stations. XGBOOST se distingue une fois de plus en offrant les meilleures performances globales, en minimisant les erreurs (RMSE et MAE) et en maximisant le R^2 et le pourcentage de prédictions correctes dans l'intervalle. En conclusion, XGBOOST se révèle être le modèle le plus fiable et le plus stable pour prédire les retards, que ce soit en fonction des stations, des jours ou des heures. Il parvient à maintenir des performances constantes. Cette robustesse lui permet de s'adapter efficacement à diverses conditions opérationnelles, le plaçant au-dessus des autres modèles testés en termes de précision et de cohérence dans les prédictions.

La section suivante présente l'analyse de sensibilité effectuée sur notre méthodologie en utilisant l'infrastructure RER. Cette analyse compare les performances des deux modèles, tramway et RER, afin d'identifier les points faibles de notre approche. En examinant les divergences de performance entre les deux infrastructures, cette étude permet de préciser les limites de la méthodologie et d'envisager des améliorations pour une meilleure généralisation des résultats.

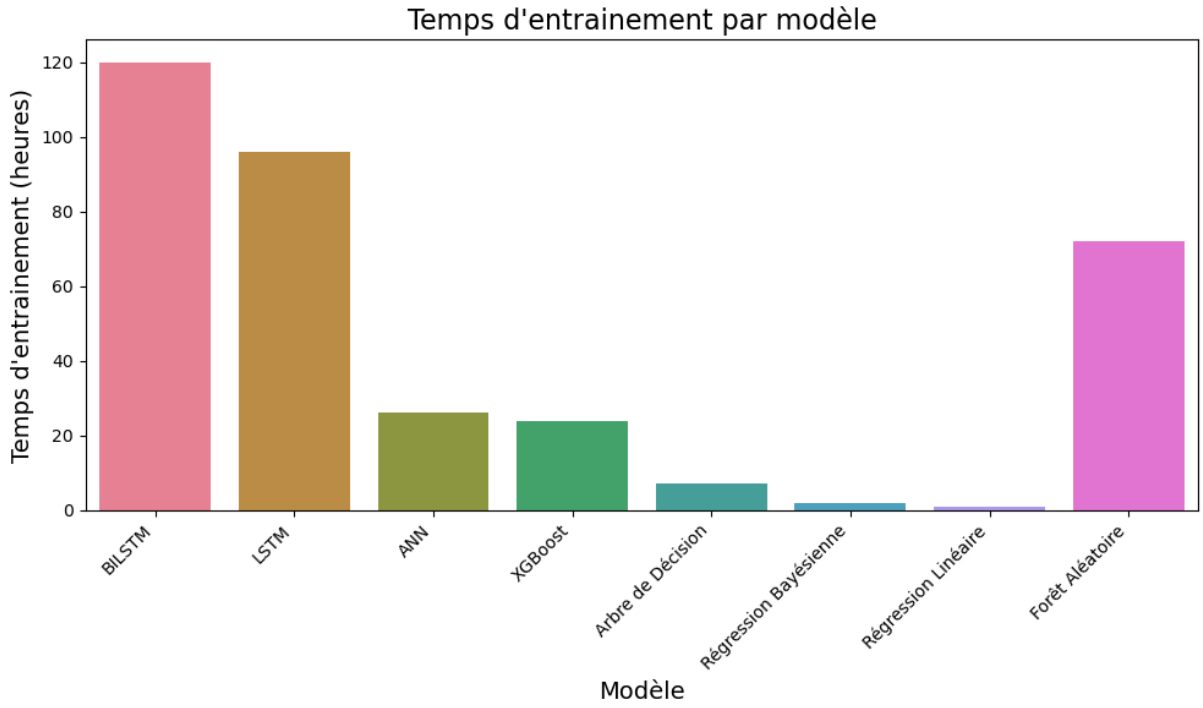


Figure 4.24 – Comparaison des temps d'entraînement des différents modèles

4.5 Analyse de sensibilité

Cette section présente les résultats obtenus pour l'entraînement des données du RER à l'aide du modèle XGBOOST, sélectionné précédemment dans la section 4.4.6, après une étude comparative des différents modèles d'apprentissage exposés dans la section 2.3.4 sur les données tramway. L'objectif ici est d'évaluer la sensibilité de notre méthodologie, déjà présentée, face à un nouvel ensemble de données issues de l'infrastructure RER. Nous cherchons à comprendre comment ces nouvelles données interagissent avec la méthodologie, qui commence par la préparation des données, l'intégration des données historiques (passé) influençant les valeurs actuelles (présent), ainsi que leur impact sur les prédictions futures. Ces données sont ensuite utilisées comme entrée pour le modèle XGBOOST sélectionné, afin de valider sa robustesse dans ce nouveau contexte.

4.5.1 La prévision à court terme du retard : Une 2eme analyse exploratoire (RER)

Nous avons effectué un test sur les données RER, en appliquant le même cadre d'entraînement décrit dans la section 3.6, impliquant une division des données en deux parties : 80 % des données, soit 192 jours, sont dédiées à l'entraînement, tandis que les 20 % restants, soit 38 jours, sont utilisés pour le test, en conservant le même principe appliqué sur les données du Tramway. Les données qualitatives sont directement intégrées, tandis que l'heure de départ réelle nécessite une ingénierie des caractéristiques. Pour les autres colonnes contenant des données quantitatives, une méthode de normalisation (eq.3.8) est appliquée. La figure 3.18 présente l'optimisation appliquée sur ce nouvel ensemble de données pour déterminer, en premier lieu, le nombre de plis

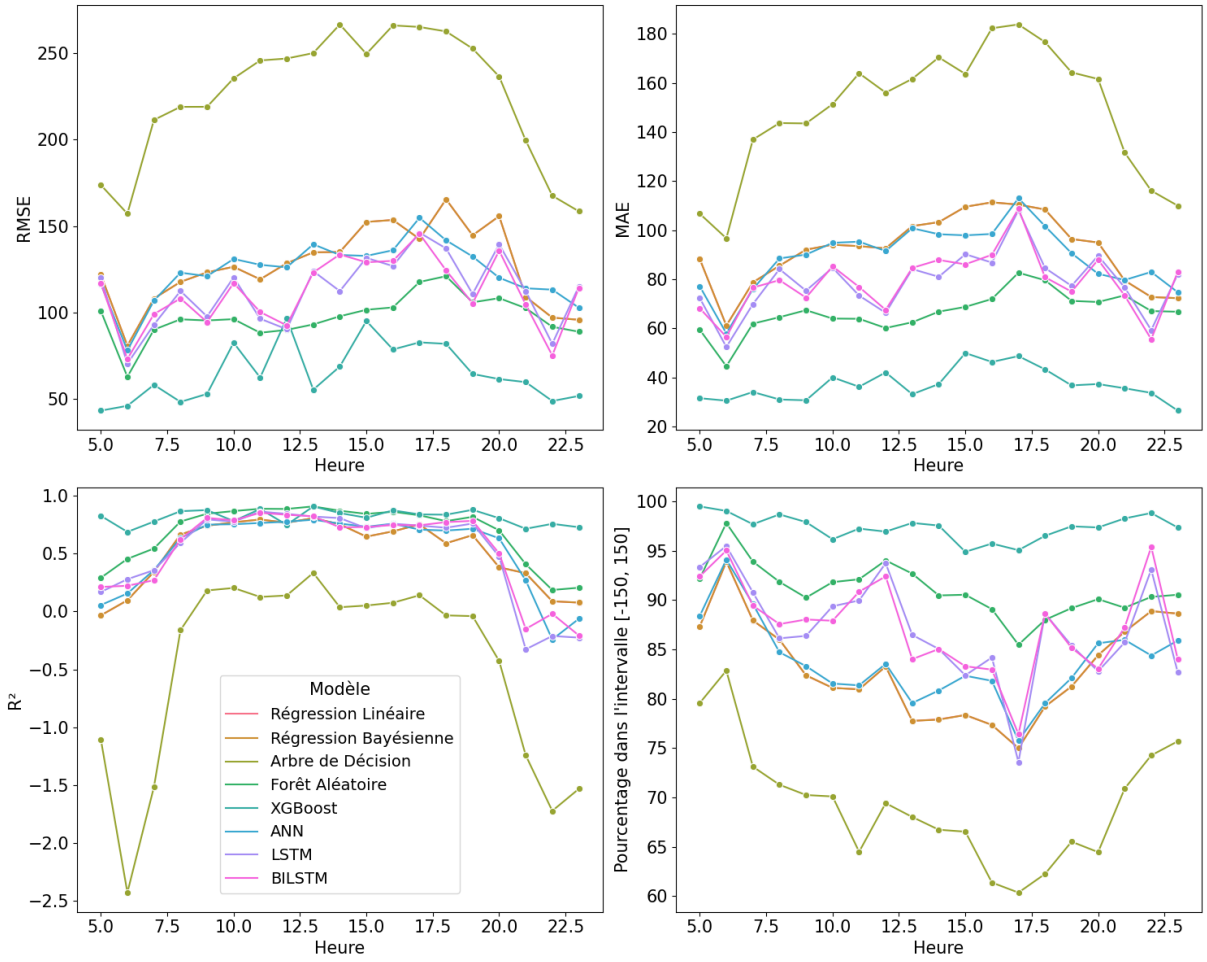


Figure 4.25 – Comparaison des performances de différents modèles de prédiction des retards en fonction de l'heure de la journée

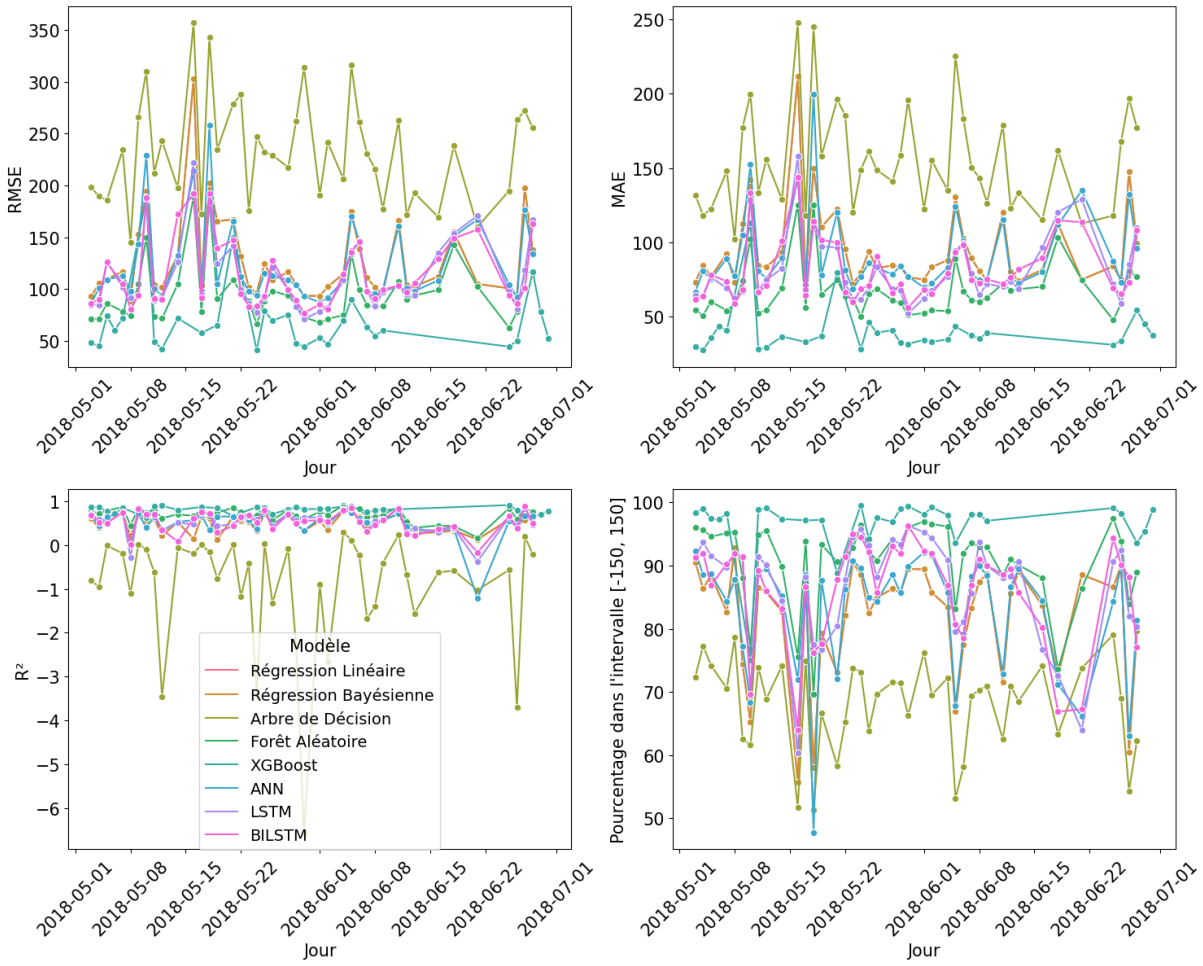


Figure 4.26 – Comparaison des performances de différents modèles de prédiction des retards en fonction du jour

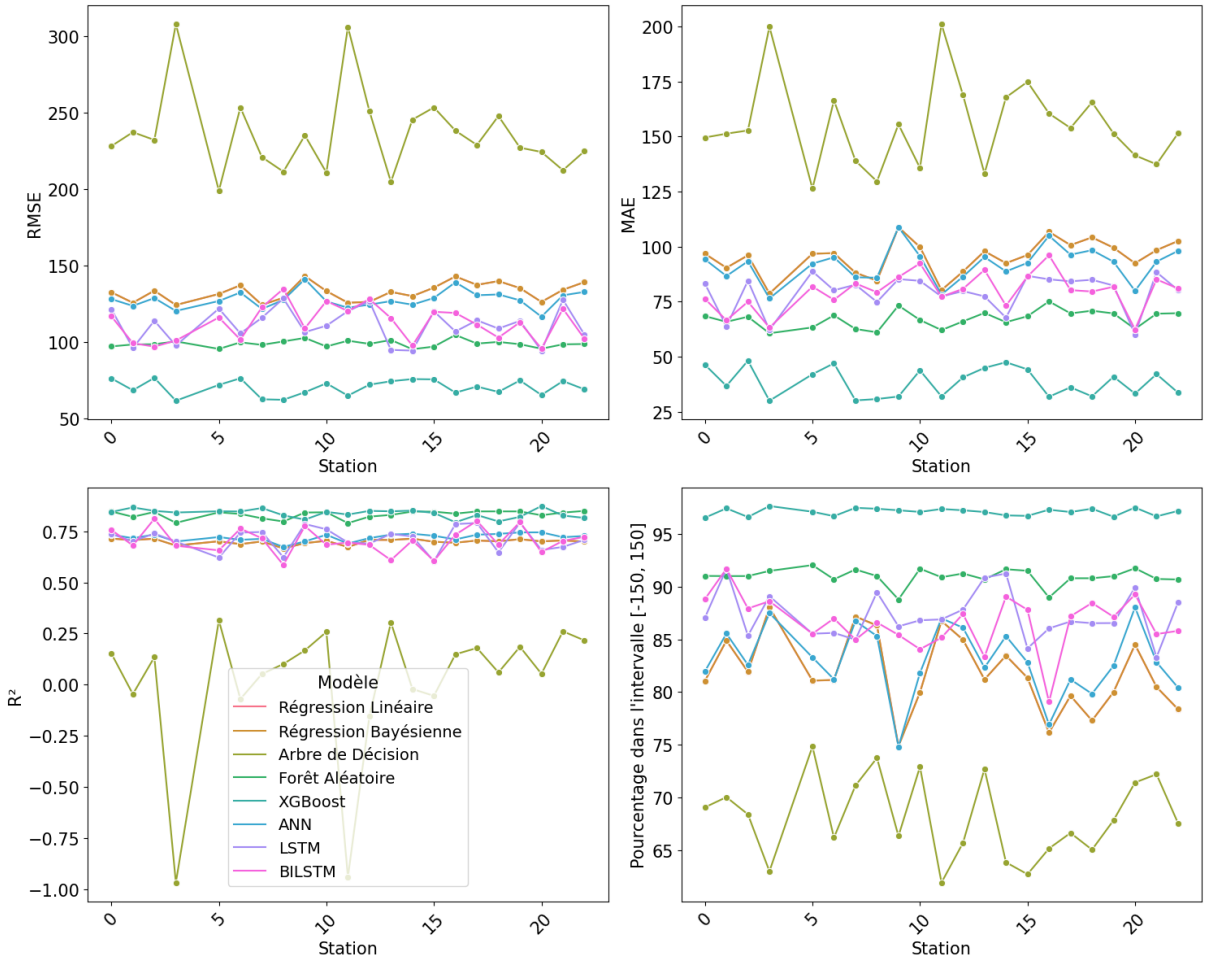


Figure 4.27 – Comparaison des performances de différents modèles de prédiction des retards en fonction de la station

à utiliser avec l'algorithme 3, qui a donné 120 plis optimaux pour le modèle XGBOOST. Toutes les variables $(XOP, XOP^{+f}, Y_{past}^{0-100})$ ont été intégrées comme variables d'entrée du modèle XGBOOST, en conservant la même fenêtre de prédiction $f = 15$ minutes. L'application de l'algorithme 1 a permis d'obtenir les hyperparamètres optimaux pour notre modèle présenté dans le tableau 4.9 avec une durée d'entraînement 2 jours.

Le modèle affiche des résultats remarquables, avec une précision de 96 % sur la moyenne des retards par jour et de 94 % sur la moyenne des retards par heure. Concernant les retards à l'heure de départ prévue, le modèle atteint une précision de 74 % pour un intervalle d'erreur de 150 secondes. De plus, il se révèle efficace pour prédire les avances avec un intervalle d'erreur de 50 secondes. Il est également capable de distinguer aisément les journées normales des journées perturbées. Cependant, quelques limitations persistent, le modèle n'atteint pas encore une précision parfaite pour l'heure exacte de départ prévue et présente des difficultés à prédire les retards dans l'intervalle de 150 secondes. De plus, le fait que les stations utilisées pour la modélisation se situent au milieu de la ligne peut limiter la corrélation avec les données passées. Les perturbations de la ligne, qui ont souvent pour origine les premières stations, échappent partiellement à la capacité du modèle à capturer l'intégralité des dynamiques de trafic. Enfin, les performances sur les données RER sont encore en deçà de celles obtenues sur les jeux de données TRAM, où le modèle montre une meilleure précision.

RER	Plis optimum	Hyperparamètre optimum	Temps d'entraînement	RMSE	MAE	R^2	P(%)
XGBOOST	120	'max_depth' : 6, 'n_estimators' : 800, 'subsample' : 0.986, 'learning_rate' : 0.016, 'colsample_bytree' : 0.977	2 jours	169	95	67	74

Table 4.9 – Résultats du modèle XGBOOST pour les données RER

4.5.2 Sensibilité des performances des modèles à la standardisation des infrastructures Tramway-RER

La comparaison entre le modèle du RER et celui du tramway met en lumière les défis liés aux différences structurelles entre les deux infrastructures de transport, telles qu'illustrées dans la section 4.1.3 de la présentation des données. La ligne de tramway étudiée est une ligne complète, tandis que les données du RER ne couvrent qu'une partie de la ligne, ce qui pourrait limiter la capacité du modèle à capturer la dynamique complète du réseau. La comparaison entre le modèle appliqué aux données du RER et le modèle XGBOOST optimisé pour le tramway, telle que présentée dans le tableau 4.8, vise donc à évaluer la sensibilité de notre approche en ajustant certains paramètres de l'infrastructure, la marge d'erreur (3.14), ainsi que les fenêtres de prédiction et du passé.

Le tableau 4.10 présente les résultats des tests effectués sur l'infrastructure ferroviaire. Actuellement, le modèle RER affiche une précision de 74% sur la marge d'erreur [-150,150], ce qui témoigne d'une certaine robustesse, bien que cette performance puisse être limitée par l'incomplétude des données disponibles. En revanche, le modèle optimisé pour le tramway, basé sur la ligne complète, présente des performances nettement supérieures par rapport au RER. Afin d'explorer l'impact de la couverture

incomplète, des tests ont été réalisés en divisant la ligne de tramway en sous-parties similaires à celles du RER. Les analyses ont d'abord été effectuées sur une sélection de 8 stations, puis réduites à des groupes de 5 stations, dans différentes configurations, afin d'observer comment les variations dans la couverture des stations influencent la convergence des résultats du modèle. En testant les cinq premières stations puis les cinq dernières, cette analyse vise à déterminer si une ligne incomplète impacte significativement la performance du modèle, en particulier en termes de capacité à capter les retards.

Les résultats, comme le montre le tableau 4.10, indiquent une diminution notable de l'indicateur R^2, passant de 0,93 à 0,70 pour le modèle du tramway lorsque celui-ci est limité à une partie de la ligne. Cette baisse suggère une sensibilité accrue du modèle aux changements de couverture des stations, révélant que la modélisation d'une ligne incomplète réduit sa capacité à capturer efficacement les retards et à maintenir un niveau élevé de précision. Ces observations soulignent la nécessité d'ajuster l'infrastructure du modèle pour en maximiser la robustesse face aux variations entre réseaux complets et incomplets. Par ailleurs, des tests ont été réalisés sur différents marges d'erreur fixés par les opérateurs. Lorsque cet intervalle est augmenté, la précision du modèle RER à 10 minutes passe à 86%, ce qui soulève des questions importantes en termes de tolérance de précision discutée avec les responsables des centres de contrôle.

Plusieurs tests ont été réalisés sur les fenêtres temporelles indiquées dans la section 4.2 afin d'analyser la sensibilité de notre méthodologie aux fenêtres temporelles capturant les retards précédents, présentées dans le tableau 4.11, et à l'horizon temporel de prédiction, présenté dans le tableau 4.12. Le premier tableau présente différentes tailles de fenêtres pour capturer les retards précédents : la fenêtre est étendue jusqu'à 1h30 et réduite à 15 minutes. On observe qu'une augmentation de la fenêtre n'améliore pas significativement la performance de notre modèle RER, bien qu'elle augmente le temps d'apprentissage. En revanche, une réduction de la fenêtre entraîne une diminution de la précision, passant de 74% à 61%, ce qui souligne l'importance de conserver une taille de fenêtre suffisante pour les retards dans le passé. De même, le second tableau illustre l'impact de l'horizon temporel de prédiction à court terme. Si l'horizon de prédiction est augmenté, la performance diminue ; à l'inverse, une réduction de l'horizon de prédiction permet d'augmenter la précision jusqu'à 80%. Cependant, un horizon de 5 minutes est considéré comme trop court pour être exploitable par les opérateurs en temps réel. Un horizon de prédiction de 10 minute reste également un point de discussion avec les opérateurs pour évaluer son utilité dans l'exploitation.

En somme, cette analyse de sensibilité révèle l'importance de bien calibrer la taille des fenêtres temporelles et l'horizon de prédiction pour maximiser l'efficacité de notre modèle dans différents contextes d'infrastructure de transport. Ces ajustements sont essentiels pour garantir que le modèle conserve une précision suffisante et demeure exploitable pour les opérateurs, que ce soit pour des lignes complètes ou incomplètes.

4.6 Conclusion

Ce chapitre présente l'ensemble du processus de préparation, d'entraînement et d'évaluation de notre méthodologie pour la prédiction des retards dans les transports publics, en tenant compte des spécificités des infrastructures du tramway et du RER.

La première étape a consisté en la préparation des deux ensembles de données, l'un pour le réseau tramway, couvrant une ligne complète de 22 stations, et l'autre pour le

	Modèle performance	R^2	RMSE	P% [-150,150]	P% [-200,200]	P% [-300,300]
RER A	Modèle	67	169	74	80	86
Tramway	Modèle de base (22 stations)	0.93	60	97	98	99
	Modèle (8 stations centrales)	79	149	83	91	95
	Modèle (6 stations centrales)	74	163	82	88	93
	Modèle (5 stations centrales)	70	182	75	85	91
	Modèle (premières 5 stations)	75	164	81	89	92
	Modèle (dernières 5 stations)	70	190	70	82	90

Table 4.10 – Comparaison des performances des modèles pour RER et Tramway

XGBOOST	R²	RMSE	P(%)
1h30	0.66	170	74
1 h	0.67	169	74
45 min	0.47	184	69
30 min	0.40	195	68
15 min	0.36	220	61

Table 4.11 – Performances du modèle XGBOOST en fonction de la fenêtre temporelle capturant les retards précédents pour le RER

réseau RER, qui contient uniquement 11 stations centrales sur une ligne de 46 stations. La différence entre ces infrastructures, avec une couverture complète pour le tramway et partielle pour le RER, est cruciale car elle influence la capacité du modèle à capturer les dynamiques de retard dans chaque réseau. La préparation des données a également impliqué le nettoyage et l'organisation des variables temporelles, en s'assurant qu'elles reflètent fidèlement les tendances annuelles, mensuelles et journalières qui influencent les retards.

Une série d'entraînements initiaux a permis de comparer différents modèles pour évaluer leurs performances sur la ligne tramway. À l'issue de ces tests, le modèle XGBOOST a été identifié comme le plus performant en raison de sa capacité à capturer les spécificités des retards dans ce réseau. Ensuite, une analyse de sensibilité approfondie a été réalisée pour évaluer l'impact de plusieurs paramètres critiques. Parmi ces paramètres, la couverture des stations dans la base de données d'entraînement s'est avérée essentielle, confirmant l'importance d'avoir une ligne complète pour capturer les spécificités de chaque segment de la ligne. La corrélation entre les retards passés et futurs a également été jugée cruciale pour une prédiction précise des retards à venir. La sélection judicieuse des fenêtres temporelles, combinée à l'utilisation de techniques de récursivité, a permis d'identifier les périodes pertinentes pour l'entraînement du modèle.

Enfin, il a été constaté que la présence de données non représentatives dans les ensembles de données d'entraînement et de test peut compromettre la capacité du modèle à se généraliser efficacement. Ces observations soulignent l'importance de calibrer correctement la taille des fenêtres temporelles et l'horizon de prédiction pour maximiser l'efficacité et la robustesse du modèle, qu'il soit appliqué à des lignes de transport complètes ou incomplètes. Ces facteurs influencent directement la capacité

XGBOOST	R^2	RMSE	P(%)
5 min	0.72	145	80
10 min	0.69	153	78
15 min	0.67	169	74
20 min	0.60	200	64
30 min	0.43	247	64

Table 4.12 – Performances du modèle XGBOOST en fonction de l'horizon temporel pour le RER

du modèle à identifier des schémas pertinents et à se généraliser de manière efficace à de nouvelles données, comme illustré dans la figure 4.28. Ainsi, des bases de données de qualité doivent répondre à plusieurs critères importants. Tout d'abord, une infrastructure de transport complète, qu'il s'agisse du tram, du RER ou du métro, est cruciale pour la mobilité urbaine ; avoir toutes les stations d'une ligne dans la base de données d'entraînement est essentiel pour capturer les spécificités de chaque segment de la ligne. Ensuite, une forte corrélation temporelle entre les années, les mois et les jours est nécessaire pour saisir les tendances temporelles influençant les retards, et la corrélation entre les retards passés et futurs est également indispensable pour prédire avec précision les retards à venir. De plus, la sélection judicieuse des fenêtres temporelles et l'utilisation de techniques de récursivité sont fondamentales pour identifier les périodes pertinentes pour l'entraînement du modèle. Enfin, la présence de données représentatives dans les ensembles d'entraînement et de test est cruciale pour garantir que le modèle puisse se généraliser efficacement.

Le chapitre suivant présente l'intégration du modèle le plus performant, XGBoost, appliqué aux données du tramway dans une interface homme-machine conviviale. Cette interface offre des prédictions avec un horizon de 15 minutes dans le futur et intègre une animation des rames, synchronisée avec les horaires de départ réels issus de la base de données tramway.

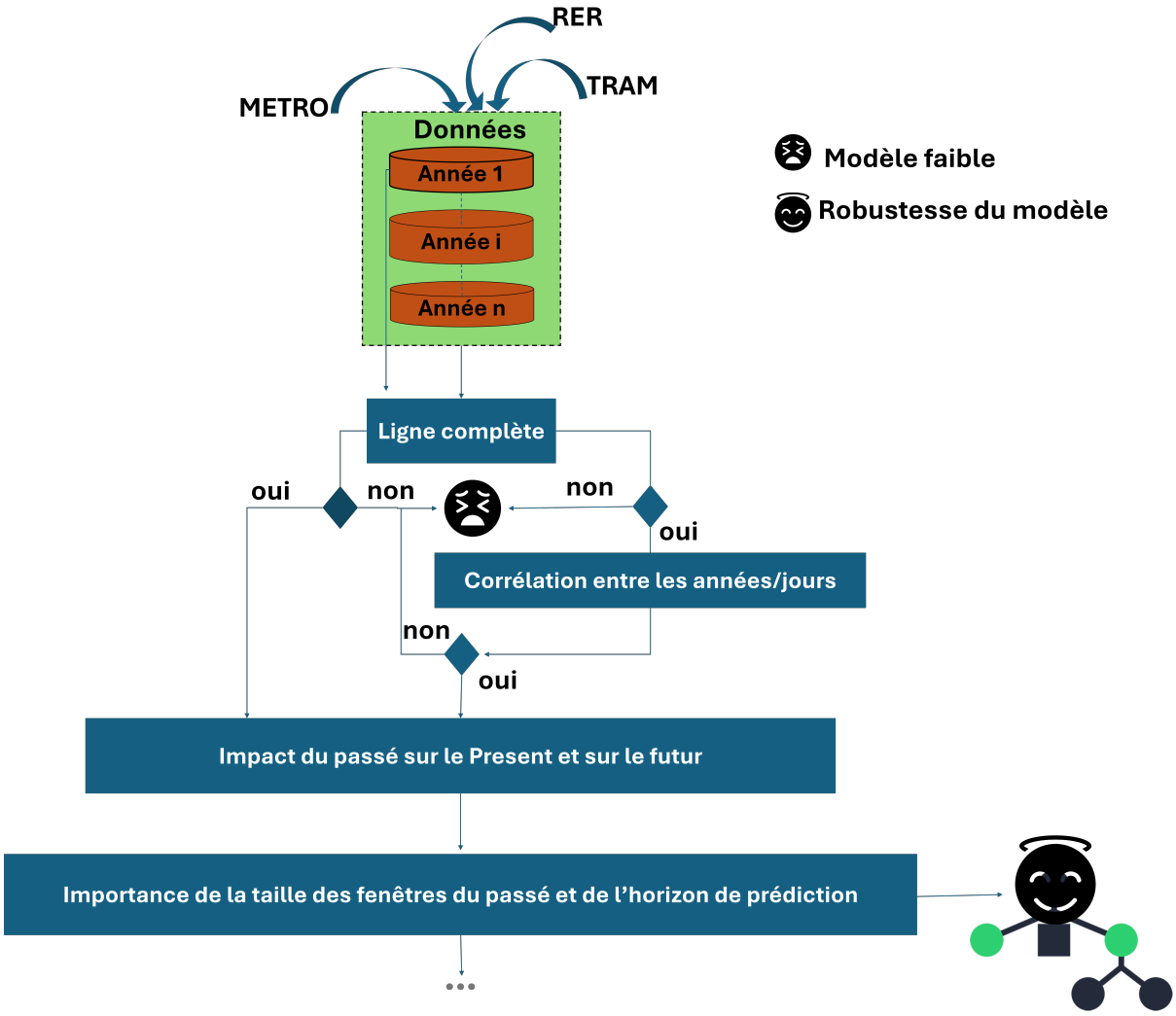


Figure 4.28 – Processus d'évaluation de la robustesse et de la sensibilité du modèle.

5

Implémentation du prototype (Système d'Aide à la Prise de Décision)

Sommaire

4.1 Cadre de travail 98

4.1.1 Environnement de travail 98

4.1.2 Sources de données (Tram, RER) 100

4.1.3 Présentation des données collectées : organisation et format . . 101

4.2 Mise en forme et préparation des données 103

4.3 Traitement analytique 104

4.3.1 Relation entre les retards et les variables d'exploitation (XOP) 107

4.3.2 Impact des variables passées sur l'analyse des retards (Y^{past}) . 110

4.4 La prévision à court terme du retard : une première analyse exploratoire (tramway) 116

4.4.1 Partitionnement des données pour l'entraînement et la validation : Tramway 116

4.4.2 Méthodes d'entraînement 117

4.4.3 Ajustement des hyperparamètres 119

4.4.4 Performance et convergence des modèles sur l'ensemble d'entraînement 119

4.4.5 Evaluation et comparaison des modèles sur l'ensemble du test . 122

4.4.6 Analyse comparative et discussion 123

4.5 Analyse de sensibilité 125

4.5.1 La prévision à court terme du retard : Une 2eme analyse exploratoire (RER) 125

4.5.2 Sensibilité des performances des modèles à la standardisation des infrastructures Tramway-RER 129

4.6 Conclusion 130

Ce chapitre détaille l'implémentation du modèle sélectionné et décrit dans le chapitre précédent 4. Ce modèle, basé sur l'algorithme XGBoost, a été optimisé à l'aide de méthodes bayésiennes pour ajuster ses hyperparamètres et obtenir

des performances optimales. L'objectif est de déployer ce modèle dans le cadre d'une interface utilisateur afin de créer un système d'aide à la prise de décision.

L'implémentation de notre interface utilisateur (IHM) repose sur le modèle XGBoost, préalablement entraîné avec les données de la ligne de tramway. Ce processus a comporté plusieurs étapes clés, telles que l'intégration du modèle de prédiction des retards dans l'architecture de l'IHM, le développement des différents composants de l'interface, et la validation de l'ensemble du système. L'IHM a été conçue pour présenter les prédictions de retards de façon claire et intuitive, garantissant une expérience utilisateur fluide et facile à appréhender. Ce premier prototype a pour objectif de démontrer la faisabilité de cette intégration et de préparer son déploiement futur dans un poste de commande centralisé (PCC), tout en soulignant l'importance de l'intelligence artificielle (IA) dans le domaine du transport public. Ce système, en rendant plus accessible la visualisation et l'interprétation des prédictions de retards, pourrait considérablement améliorer la prise de décision des opérateurs dans un environnement de contrôle ferroviaire. Chaque étape de l'implémentation est décrite de manière détaillée pour illustrer le processus suivi ainsi que les choix technologiques effectués.

Ce chapitre s'articule autour de quatre sections principales. La première, intitulée "Conception de l'interface homme-machine", détaille le processus de conception en présentant les "Maquettes" et les "Fonctionnalités clés avec diagramme des cas d'utilisation", fournissant ainsi une base visuelle et fonctionnelle pour le prototype.

La seconde section, "Architecture générale du prototype complet (IHM)", propose une vue d'ensemble de la structure technique et organisationnelle de l'interface utilisateur.

La troisième section, "Phase de Développement du Prototype", explore en détail les différentes étapes du développement à travers plusieurs diagrammes de séquence.

Enfin, la dernière section, "Mise en œuvre sur un cas test : Scénario simplifié", décrit l'utilisation concrète du prototype sur des scénarios simplifiés couvrant des périodes spécifiques.

En somme, ce chapitre fournit une vue complète et structurée du processus de conception, de développement et de validation du prototype, en mettant en avant les choix techniques et méthodologiques effectués à chaque étape.

5.1 Conception de l'interface homme machine

L'implémentation du prototype a pour objectif de clarifier comment le logiciel envisagé pourrait fonctionner tout en servant de preuve de concept. Ce prototype ne couvre pas l'ensemble des étapes de conception, mais constitue l'une des phases les plus cruciales : la conception des sous-systèmes horizontaux de l'interface d'aide à la prise de décision pour un poste de commande centralisé (PCC). Cette phase inclut l'intégration de l'IHM, qui vise à offrir aux opérateurs une vue d'ensemble précise et en temps réel des prédictions de retards, facilitant ainsi la prise de décisions stratégiques.

L'interface a été réalisée en prenant en compte les spécificités détaillées dans notre méthodologie, avec pour objectif de proposer un écran supplémentaire aux opérateurs. Cet écran affiche les résultats optimisés de notre modèle de prédiction des retards en temps réel, offrant ainsi un outil d'aide à la prise de décision pour la régulation du trafic. L'intention n'est pas de remplacer le système de régulation existant, mais de fournir des propositions complémentaires que les opérateurs peuvent considérer pour affiner leurs décisions.

Ce premier prototype a été conçu pour gérer des données en temps réel. Pour y parvenir, il a été nécessaire de manipuler les données historiques du tramway de manière à simuler un flux de données en temps réel. Cette manipulation permet de reproduire des conditions d'exploitation réalistes et de démontrer comment notre interface peut réagir et s'adapter de manière dynamique, comme le ferait un système utilisé en conditions réelles. L'interface est inspirée par les écrans déjà existants dans les PCCs. L'objectif est de garantir que notre prototype se rapproche le plus possible d'une interface fonctionnant en temps réel, permettant aux opérateurs d'expérimenter et de tester son potentiel dans un environnement proche de celui du PCC.

5.1.1 Maquette

Nous proposons un écran qui combine trois concepts : les données en temps réel des PCC, les modèles d'apprentissage et le système de régulation des PCC, que nous avons intégrés dans notre modèle statique du prototype proposé par les opérateurs d'egis-Rail, tel que représenté par la figure 5.1. Cette figure présente trois angles : 1. le temps de la simulation (date, heure de début et de fin de la simulation avec un facteur d'accélération), permettant de définir le cadre temporel de la simulation. Bien que notre objectif soit de développer une interface homme-machine (IHM) en temps réel, nous n'avons pas un accès direct aux données en temps réel, mais uniquement à des données statiques du tramway. Pour pallier cette limitation, nous gérons ces données dans la simulation afin de créer un flux simulé de données en temps réel; 2. la zone principale de la simulation, qui constitue l'espace central où la simulation s'exécute et s'affiche, contenant les résultats ou visualisations générés en fonction des paramètres définis par l'utilisateur; et 3. la section de téléchargement des données de simulation, permettant de télécharger les résultats de la simulation, y compris les prédictions générées durant la période de simulation, afin de réaliser des statistiques en fin de session et d'évaluer l'interface. Cette représentation vise à offrir une vision intégrée de la simulation, permettant d'articuler les différents aspects des données et des modèles pour mieux évaluer les performances et la pertinence de notre approche dans un contexte réaliste.

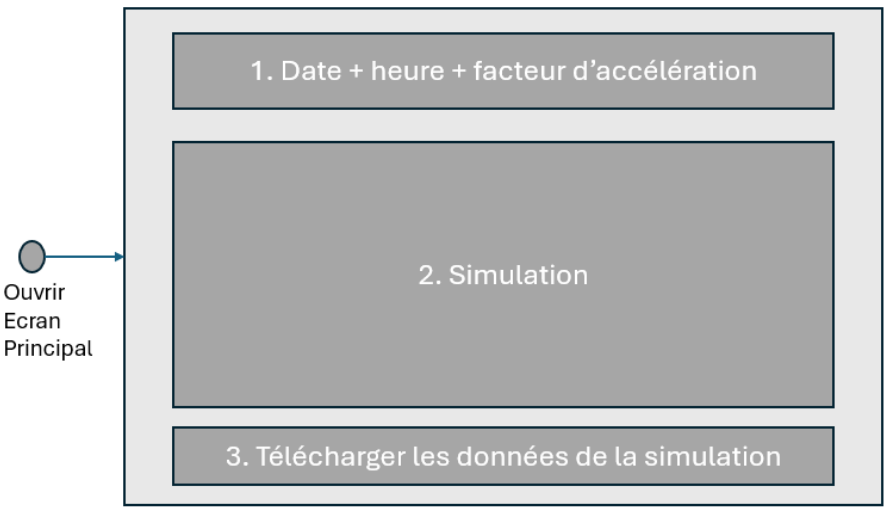


Figure 5.1 – Modèle statique du prototype IHM

La figure 5.2 présente la conception de notre interface, qui s'articule autour de

plusieurs éléments clés visant à offrir une représentation dynamique et interactive du système ferroviaire. L'interface s'appuie sur une base de données ferroviaire, dans notre cas les données du tramway, pour dessiner les stations et les rames. Une station occupée sera représentée par une couleur bleu clair pour indiquer le quai concerné, tandis qu'une station libre sera marquée en bleu foncé. Un grand rectangle jaune symbolisera le déport des rames, offrant une meilleure visualisation de la structure ferroviaire. Les aiguilles, qui indiquent les jonctions entre différentes voies, doivent être représentées visuellement pour fournir une meilleure compréhension des connexions et de la direction des rames. La circulation des trains entre les stations doit se baser sur l'heure de départ réelle, assurant une représentation précise de leur mouvement. L'interface animera les rames en déplacement d'une station à une autre en utilisant l'heure de départ réelle comme référence. Les rames peuvent être représentées par un rectangle ou par un symbole ou une image spécifique représentant une rame.

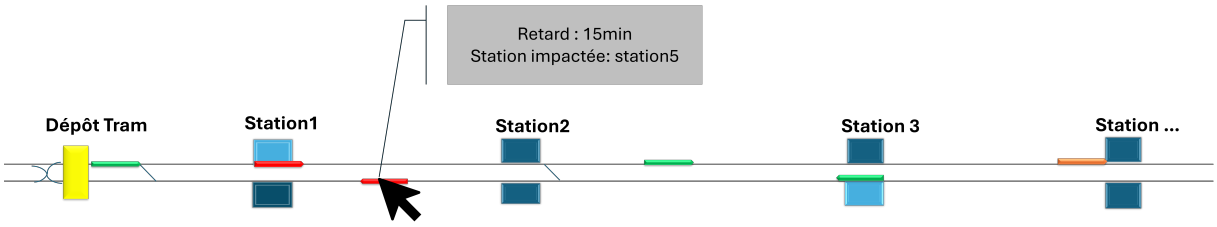


Figure 5.2 – Maquette interactive du système ferroviaire

La couleur des rames variera en fonction du retard ou de l'avance. Pour les retards : une rame avec un retard de cinq minutes ou moins sera représentée en vert avec un symbole "R". Si le retard se situe entre 5 et 10 minutes, la rame sera affichée en orange, et pour les retards critiques (supérieurs à 10 minutes), la rame sera représentée en rouge et clignotera pour attirer l'attention. Pour les avances : une rame avec une avance de moins de 5 minutes sera représentée en vert avec un symbole "A". Les avances comprises entre 5 et 10 minutes seront indiquées en orange, tandis que les avances supérieures à 10 minutes seront marquées en rouge clignotant.

Le modèle d'apprentissage utilisé est celui choisi dans le chapitre 4, à savoir le modèle XGBoost, avec un horizon de prédiction de 15 minutes. Lorsqu'une rame présente un retard critique (indiquée par la couleur rouge) prédit par le modèle intégré à l'IHM, l'opérateur peut interagir avec l'écran en cliquant sur la rame concernée pour afficher une étiquette contenant des informations telles que les stations impactées dans les 15 prochaines minutes et plus, ainsi que le retard prédit. Cette approche permet de reproduire les conditions d'exploitation réelles et de démontrer comment notre interface peut réagir de manière dynamique, tout en s'inspirant des écrans existants dans les postes de contrôle. L'objectif est de garantir une expérience proche du temps réel pour les opérateurs, leur permettant d'expérimenter et de tester le potentiel de cette interface dans un environnement simulé.

5.1.2 Fonctionnalités clés avec diagramme des cas d'utilisation

La figure 5.3 présente un diagramme de cas d'utilisation [53] illustrant les différentes interactions et fonctionnalités disponibles pour l'interface homme-machine (IHM) destinée à la simulation et à l'affichage des prédictions. Ce diagramme met en évidence

le rôle central de l'opérateur, qui interagit avec le système pour configurer, lancer et consulter les simulations.

Le processus commence par le cas d'utilisation "Saisir les configurations", où l'opérateur doit fournir divers paramètres tels que la date, l'heure de début, l'heure de fin et le facteur d'accélération. Ces paramètres permettent de définir le temps de la simulation afin d'obtenir des résultats. Une fois ces configurations saisies, l'opérateur peut déclencher le cas d'utilisation "Lancer la simulation", permettant ainsi l'exécution du modèle de simulation. Après le lancement, l'opérateur peut accéder au cas d'utilisation "Consulter la simulation", qui inclut l'affichage d'informations telles que le code couleur des rames basé sur les prédictions du modèle XGBoost. Ce code couleur permet d'identifier rapidement l'état des rames en fonction des prédictions. La consultation inclut également la position des rames, fournissant une visualisation en temps réel de la localisation des rames dans la simulation, et la circulation inter-station, qui montre la dynamique des rames entre les différentes stations de manière interactive.

Enfin, le cas d'utilisation "Afficher les prédictions" permet de visualiser les informations prédites telles que la station impactée, l'heure de départ prévue et le retard prédit, fournissant des indications essentielles pour l'optimisation et la gestion des opérations ferroviaires. Si la simulation est terminée, il est possible de télécharger les données à l'aide du cas d'utilisation étendu "Télécharger les données de la simulation", offrant ainsi à l'opérateur une possibilité de stocker et d'analyser les résultats obtenus.

La section suivante présente l'architecture générale de notre conception du prototype de l'IHM à réaliser, en détaillant la logique ainsi que les technologies utilisées dans l'interface.

5.2 Architecture générale du prototype complet (IHM)

La figure 5.4 présente l'architecture globale de IHM [111, 112] mise en place pour la conception détaillée, décrite dans la section 5.1, repose sur quatre entités essentielles : la base de données, le front-end, le back-end et le modèle d'apprentissage. La base de données a été conçue pour intégrer et traiter les données des deux derniers mois du tramway, comme expliqué dans la sous-section 4.4.1. Cette structure permet de tester notre modèle optimal, XGBoost, sur les données de test afin d'évaluer ses performances en temps réel pour la prédiction des retards.

Le back-end contient tous les fichiers et scripts relatifs au traitement de l'application, incluant l'API Flask pour la gestion des requêtes HTTP, la gestion de la base de données, et l'intégration du modèle de machine learning pour les prédictions. Il est dédié à la préparation des données en temps réel et à tous les calculs nécessaires pour assurer un flux dynamique de données, évitant ainsi une approche statique. Flask [113] a été choisi pour sa flexibilité, sa syntaxe simple et intuitive, ainsi que pour son support d'extensions qui facilitent l'ajout de fonctionnalités supplémentaires. Les fichiers CSV des tramways ont été transformés en une base de données SQLite [114], permettant des requêtes SQL rapides pour récupérer des données spécifiques. Comparativement aux fichiers CSV, cette approche optimise la recherche et la manipulation des données, les rendant plus rapides et plus efficaces.

Le front-end, construit avec PySide [115], regroupe tous les éléments liés à l'interface utilisateur de l'application. Il utilise le langage QML[116] pour définir et gérer les composants visuels, tandis que des scripts en Python et JavaScript contrôlent la logique applicative. L'interface est conçue pour présenter et afficher les résultats de manière

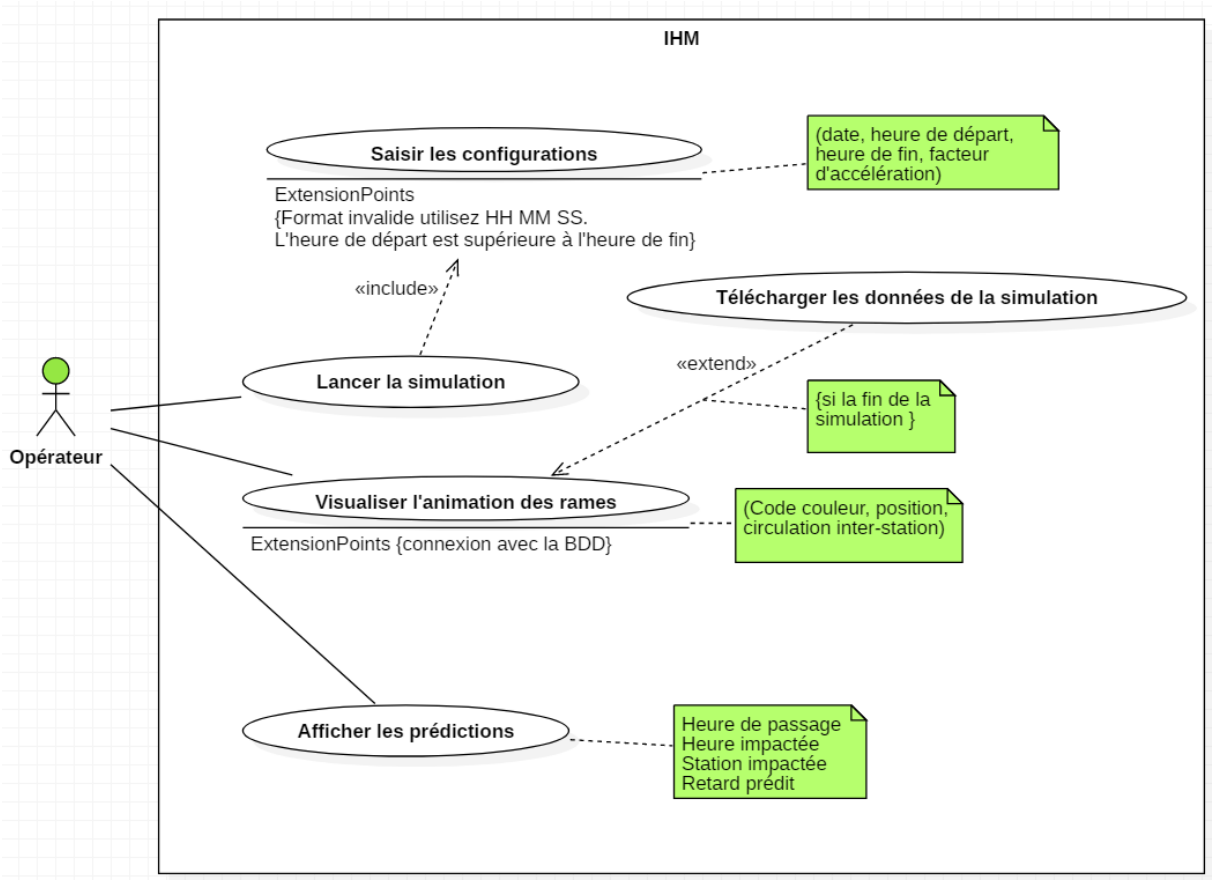


Figure 5.3 – Diagramme de cas d'utilisation pour la simulation et l'affichage des prédictions dans l'IHM

claire et intuitive aux opérateurs, facilitant leur interaction avec le système. Le front-end envoie des requêtes HTTP pour récupérer les données nécessaires à l'affichage, et l'API agit comme un intermédiaire, traitant les demandes et fournissant les réponses pertinentes. Cette architecture garantit un fonctionnement fluide et réactif, renforçant l'interactivité et l'efficacité du système d'aide à la prise de décision utilisé dans le contexte d'un PCC (Poste de Commande Centralisé). En plus de son rôle d'affichage, le front-end constitue le point d'entrée de l'interface utilisateur (IHM) de l'application. Grâce à PySide, il charge et affiche les vues QML, tandis que la logique métier est gérée via Python. Il se connecte à une API Flask pour récupérer des données au format JSON, qui sont ensuite exposées aux composants QML de l'IHM. Le front-end offre également des fonctionnalités avancées pour transformer, valider et exporter les données, tout en gérant les erreurs de manière efficace afin d'assurer une expérience utilisateur optimale.

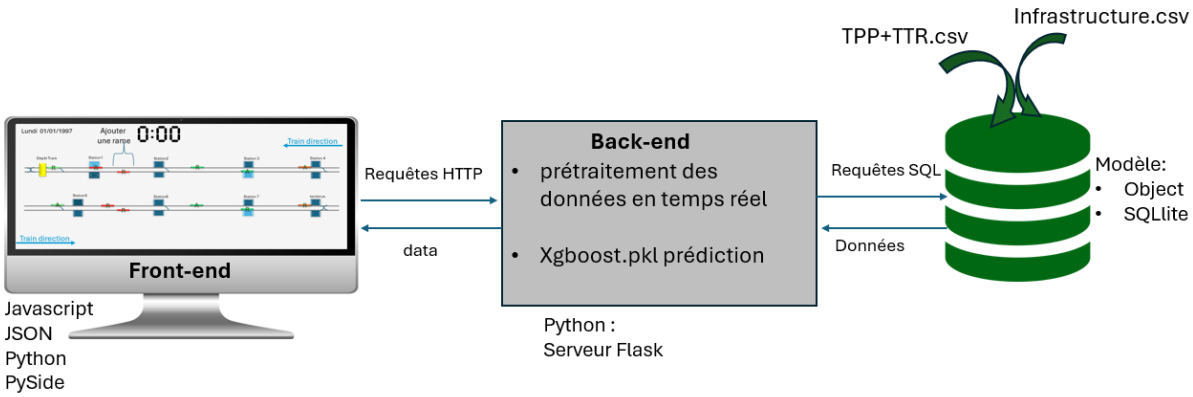


Figure 5.4 – Architecture globale de IHM : intégration du front-end, back-end, base de données et Xgboost.pkl pour la prédiction des retards

La section suivante présente le développement des fonctionnalités nécessaires pour les composantes back-end et front-end afin de construire notre IHM (Interface Homme-Machine). Les interactions entre chaque élément sont représentées à l'aide de diagrammes de séquence, permettant de visualiser chaque itération. Cette approche facilite la compréhension des échanges de données, de la logique applicative et des flux d'information, en mettant en lumière les relations et les dépendances entre les différents composants de l'architecture.

5.3 Phase de Développement du Prototype

Après avoir établi la conception générale de l'interface homme-machine (IHM), cette section se concentre sur le développement des fonctions spécifiques de chaque composant, incluant le front-end et le back-end. Le développement de ces fonctionnalités vise à garantir une intégration fluide et une interaction cohérente entre les différents éléments du système. Nous présentons, à travers des diagrammes de séquence, les interactions entre les composants pour mettre en évidence les échanges de données, la logique applicative et les processus d'intégration, assurant ainsi un flux d'informations optimal entre le front-end et le back-end.

Le choix d'utiliser des diagrammes de séquence dans ce contexte repose sur leur capacité à illustrer de manière claire et précise les interactions dynamiques entre

les différents composants de l'application. En visualisant le flux d'information et les échanges entre le front-end et le back-end, ces diagrammes mettent en évidence les étapes de communication, les déclenchements de processus, et les relations temporelles qui régissent le comportement de notre système. Ils permettent ainsi de mieux comprendre la logique sous-jacente, de clarifier les dépendances fonctionnelles et de garantir une intégration cohérente entre les différentes fonctionnalités développées. Ce choix s'inscrit dans une volonté de rendre le développement plus transparent, tout en facilitant l'analyse, la validation et la documentation des interactions complexes au sein de l'architecture logicielle.

Dans cette optique, le diagramme de séquence 5.5 présente les cinq lignes de vie essentielles pour les interactions et la logique de notre application dans le contexte d'une simulation ferroviaire, depuis son lancement jusqu'à la fermeture de l'interface. Le processus commence par le lancement de l'application, déclenchant automatiquement deux références principales : la "création de la ligne à simuler" 5.6 et la définition des "caractéristiques de la simulation" 5.7. La création de la ligne à simuler permet de visualiser la ligne ferroviaire, incluant les stations et les aiguilles, tandis que la définition des caractéristiques de la simulation consiste à obtenir une liste de données spécifiques.

Si les données de la simulation sont disponibles (`données=True`), le processus entre dans une boucle pour simuler le flux de données en temps réel. Ce flux est structuré pour récupérer les données à des intervalles de deux minutes avec des mises à jour toutes les deux secondes. À chaque itération de la boucle, les caractéristiques et les prédictions des retards sont calculées à l'aide du modèle de prédiction (XGBoost) 5.8, générant une liste des retards prédits. Une fois les prédictions effectuées, l'animation des rames avec un code couleur démarre, permettant une visualisation dynamique et intuitive de la situation sur la ligne simulée 5.9.

Lorsque la simulation atteint la fin des données disponibles, une étape de téléchargement est initiée, permettant à l'utilisateur d'exporter les données de la simulation pour une analyse plus approfondie. Si les données retournées sont incorrectes ou absentes (`données=False`), une gestion des erreurs intervient, menant potentiellement à la fermeture ou à la destruction de l'interface de simulation. Les références (`ref`) indiquées dans le diagramme représentent d'autres diagrammes de séquence 5.6, 5.7, 5.8 et 5.9 qui seront expliqués plus en détail par la suite pour clarifier leur rôle dans les interactions et les processus représentés.

5.3.1 Diagramme de séquence : Création de la ligne à simuler

Le diagramme de séquence 5.6 illustre le processus de création de la ligne à simuler, en commençant par l'obtention d'une liste de stations (`liste des stations = True`). Le processus démarre par une interaction entre le front-end et le back-end, où ce dernier envoie une requête à la base de données via la méthode `créationStation` pour générer et récupérer la liste des stations nécessaires à la simulation. Les informations sur les stations sont ensuite renvoyées au back-end et transmises au front-end.

À partir de là, le front-end procède à plusieurs étapes de traitement : d'abord, il réorganise les stations reçues grâce à la méthode `réorganiser stations`, permettant de structurer les stations selon l'ordre requis pour la simulation. Ensuite, le front-end crée les inter-stations (`créationInterStation`), représentant les segments de liaison entre les stations. Une fois ces éléments préparés, la visualisation des stations et des

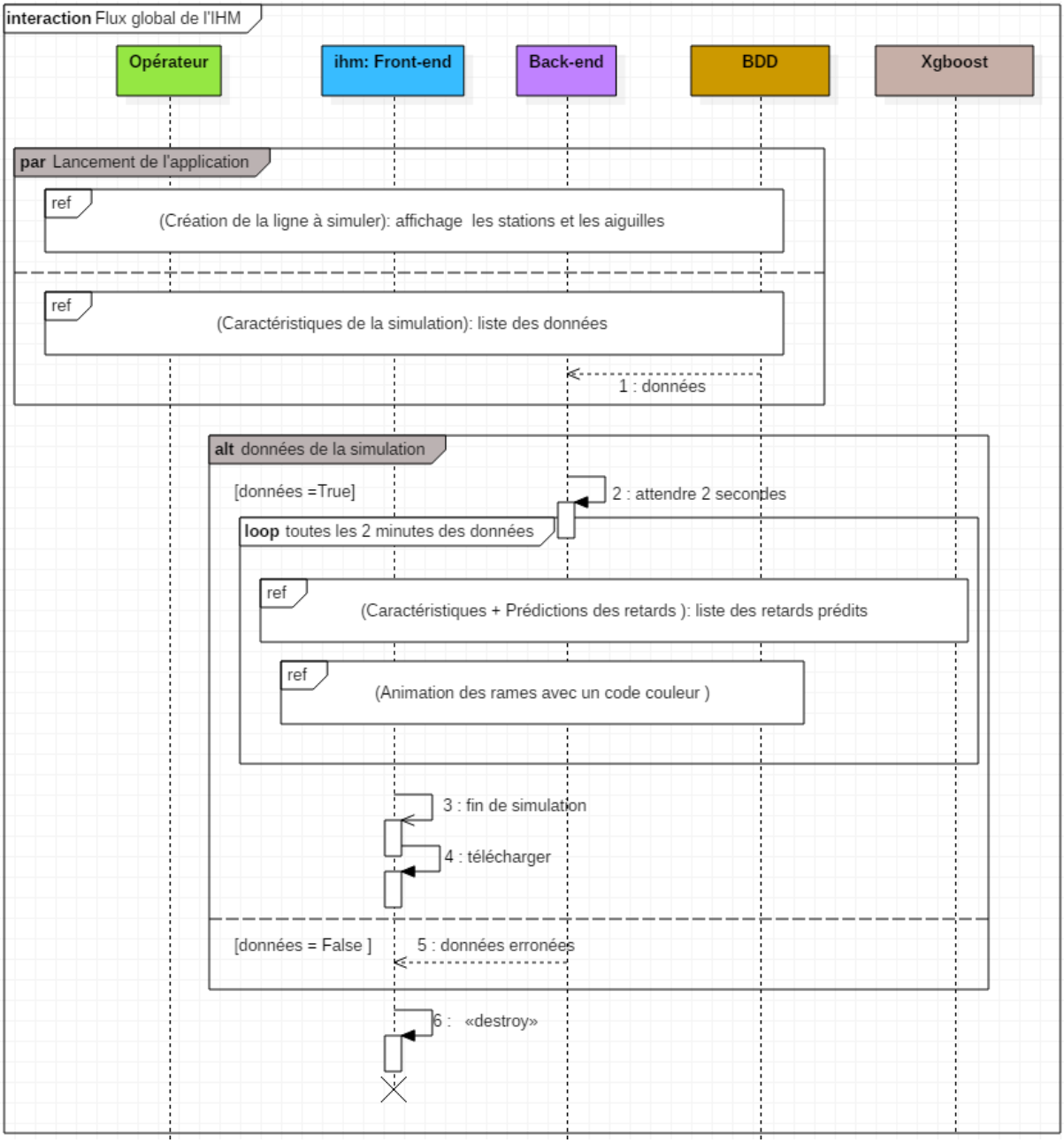


Figure 5.5 – Diagramme de séquence illustrant le flux global de l'application, depuis le lancement de la simulation jusqu'à sa fermeture : la création de la ligne, la récupération et le traitement des données, la prédiction des retards, l'animation des rames.

inter-stations (`visualisation des station et interStation`) est effectuée, rendant l'ensemble visible sur l'interface utilisateur.

Le processus se poursuit par la création des aiguilles. Cela commence par la procédure `Création aiguille`, qui comporte la création de deux types d'aiguilles : `Aiguille A` et `Aiguille B`. Le front-end appelle successivement les méthodes `créationA` et `créationB`, en fournissant les stations et le paramètre `isFirst` pour indiquer l'ordre ou la priorité. Enfin, le processus aboutit à la visualisation des aiguilles (`visualisation des aiguilles`), assurant une cohérence graphique et fonctionnelle dans l'interface utilisateur.

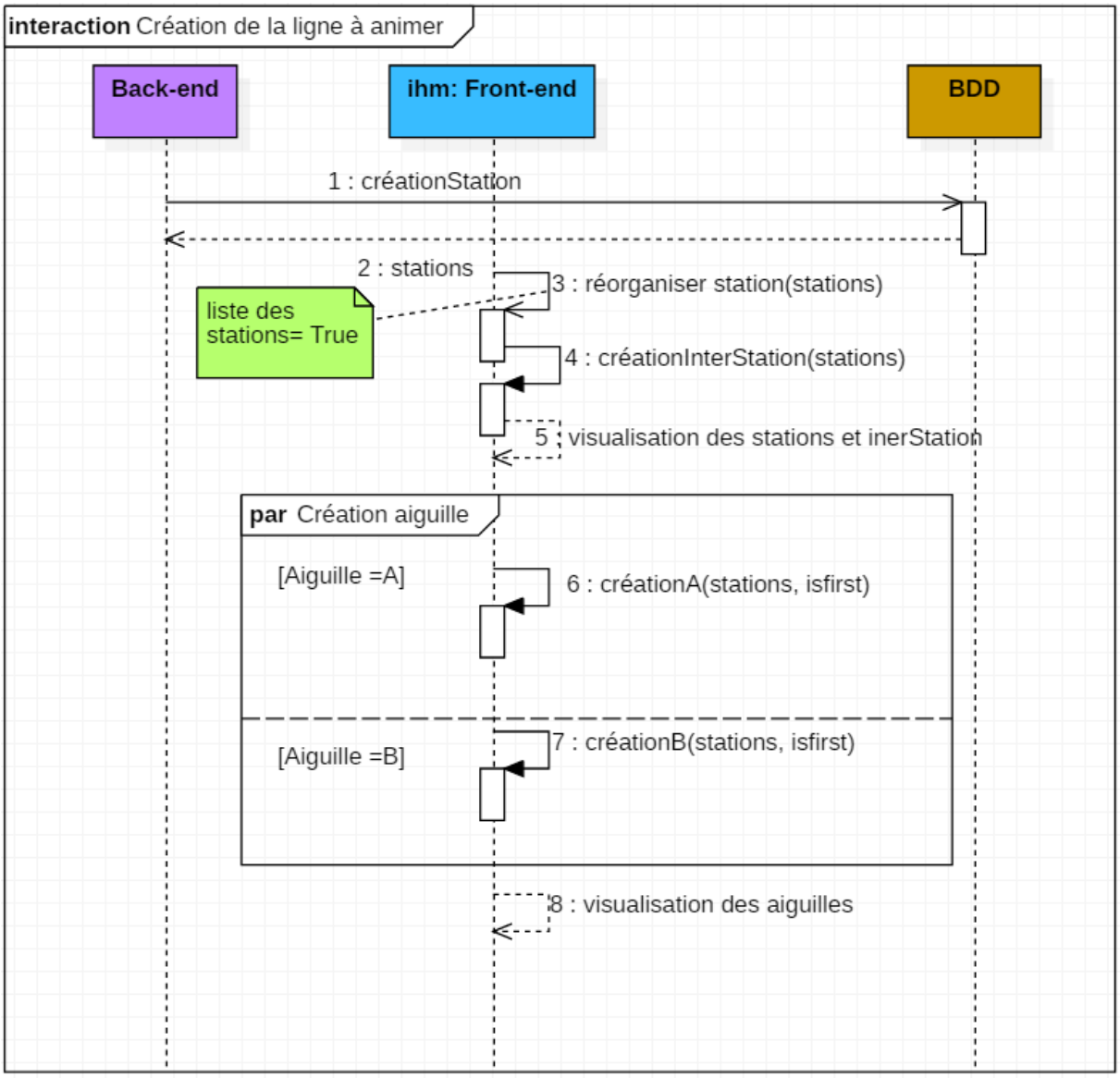


Figure 5.6 – Diagramme de séquence pour la création de la ligne à simuler : la réorganisation des stations, la génération des inter-stations et la visualisation des aiguilles.

5.3.2 Diagramme de séquence : Caractéristiques de la simulation

Le diagramme de séquence 5.7 décrit le processus d'initialisation des caractéristiques de la simulation, impliquant une interaction entre l'opérateur, le front-end, le back-end et la base de données (BDD). Le processus commence par une boucle `Initialiser la Simulation`, qui s'exécute tant que les entrées fournies par l'opérateur sont correctes. L'opérateur doit saisir plusieurs paramètres essentiels : la date (`entrer la date`), qui correspond au jour à sélectionner dans la base de données pour la simulation, l'heure de départ (`entrer heure de départ`) et l'heure de fin (`entrer heure de fin`) de la simulation, ainsi que le facteur d'accélération de la simulation (`choisir le facteur d'accélération`).

Les facteurs d'accélération de notre interface comprennent cinq valeurs : $F_a \in \{1, 20, 30, 40, 60\}$. Le facteur 1 représente une simulation en temps réel. À mesure que le facteur augmente, la simulation est accélérée proportionnellement. Le principe des

facteurs d'accélération est le suivant : pour chaque intervalle de 2 minutes en temps réel, la durée correspondante dans la simulation est ajustée en fonction du facteur choisi. Par exemple, si un facteur d'accélération de 20 est utilisé, alors 2 minutes en temps réel seront équivalentes à 20 secondes dans la simulation. En revanche, avec un facteur de 1, 2 minutes en temps réel seront équivalentes à 2 minutes dans l'application, sans accélération. Ainsi, la durée de la simulation (en minutes) peut être calculée à l'aide de la fonction suivante :

$$T_s \text{ (secondes)} = \frac{T_r \text{ (s)} \times F_a}{120} \tag{5.1}$$

— T_s : Temps de la simulation, c'est-à-dire la durée totale pendant laquelle la simulation s'exécute.
— T_r : Temps de la plage sélectionnée, représentant la période de temps réelle que l'on souhaite simuler (par exemple, 08 :08 :08 à 08 :50 :50).
— F_a : Facteur d'accélération, utilisé pour accélérer la simulation et réduire le temps d'exécution par rapport à la durée réelle.
— La division par 2 indique que les données sont récupérées ou traitées à intervalles réguliers de deux secondes, optimisant ainsi le temps de simulation.

Une fois ces informations renseignées, le front-end transmet une requête au back-end via la méthode `get_passage`, en utilisant les paramètres de date, d'heure de départ et d'heure de fin. Le back-end interroge alors la base de données pour récupérer les lignes de données correspondant à l'intervalle de temps spécifié. Si les données récupérées sont valides (`données=True`), elles sont renvoyées et utilisées dans la simulation. Ce processus garantit que les données de passage des rames entre l'heure de départ et l'heure de fin sont correctement prises en compte, permettant ainsi une exécution précise et conforme à la configuration définie par l'opérateur.

5.3.3 Diagramme de séquence : Caractéristiques + Prédiction des retards

Le diagramme de séquence 5.8 présenté illustre le processus de prédiction des retards pour chaque passage de rame, basé sur un flux de données en temps réel. Ce processus se déroule dans une boucle qui traite les données toutes les deux minutes pour chaque passage devant une station.

Le processus commence par la recherche de la station future estimée 15 minutes dans le futur (`trouver_futur_station`). Si une station future est trouvée, le back-end récupère les caractéristiques de ce passage estimé, incluant la station, l'heure prévue et le retard potentiel. Ces informations sont combinées avec les données du passage actuel (`combiner(passage, passage_estimé)`). Si aucune station future n'est trouvée ou si la fin de la journée est atteinte, aucune donnée supplémentaire n'est ajoutée.

Ensuite, le processus se concentre sur les retards précédents. Pour le premier passage, le système récupère les 80 derniers retards enregistrés (`récupérer_retards_précédents`), qui sont stockés dans une variable `retards_précédents`. Pour les passages suivants, le dernier retard est récupéré et utilisé pour mettre à jour les retards précédents (`mise_à_jour(retards_précédents, dernier_retard)`). Cela garantit que les retards historiques sont maintenus à jour pour chaque passage.

Après la combinaison des données du passage actuel avec les retards précédents, le système sélectionne les caractéristiques nécessaires 3.9 pour le modèle de pré-

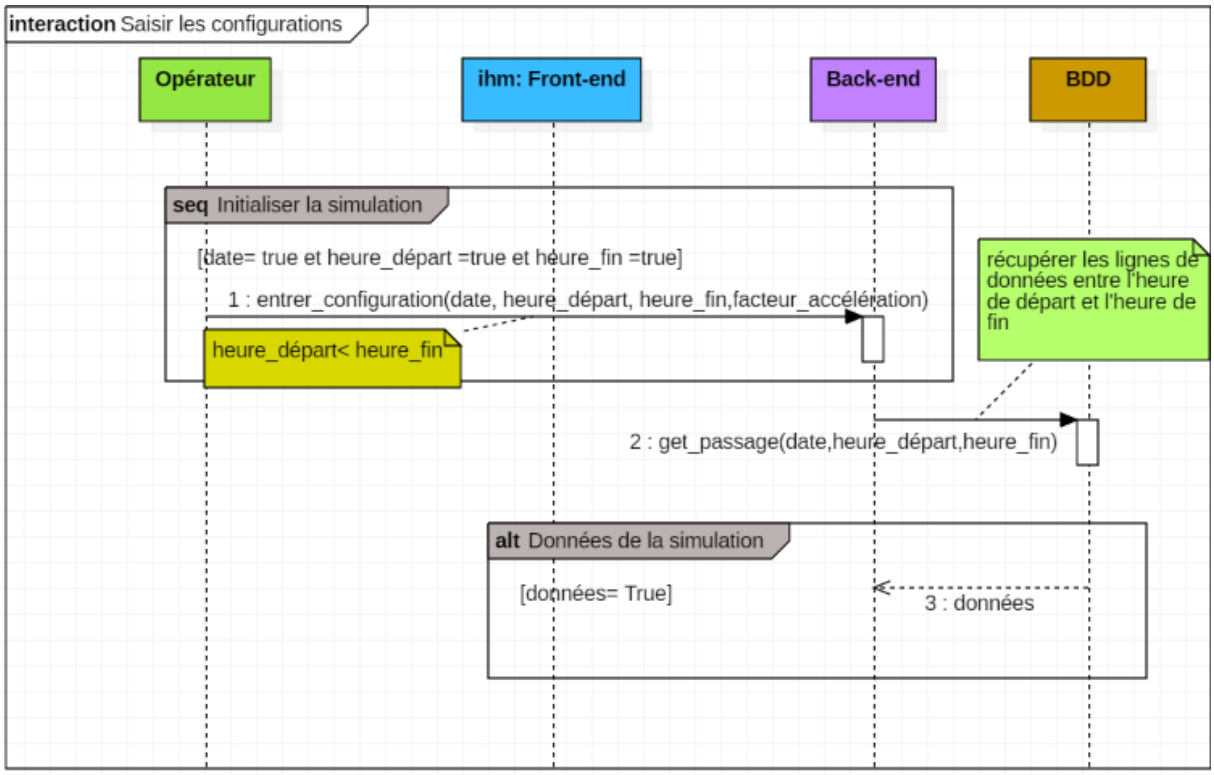


Figure 5.7 – Diagramme de séquence pour l'initialisation des caractéristiques de la simulation.

diction (`sélectionner_caractéristiques_modèle`). Le modèle de prédiction (XGBoost) est ensuite appliqué pour générer une prédiction de retard (`prédire(passage)`), qui est associée au passage. Cette prédiction est renvoyée pour une utilisation dans la simulation, offrant des informations précises et actualisées sur les retards.

5.3.4 Diagramme de séquence : Animation des rames avec un code couleur

Après le processus de prédiction des retards, le diagramme de séquence 5.9 illustre les étapes d'animation des rames en utilisant un code couleur pour indiquer les prédictions. Ce processus commence par la récupération de la prédiction associée au passage actuel (`get_prediction(passage)`). Le front-end reçoit la prédiction et utilise cette information pour déterminer la couleur associée à l'état de la rame (`trouver_couleur(prédiction)`) suivant le code couleur expliqué dans section 5.1.1.

Ensuite, le système détermine le symbole à associer à la rame pour indiquer si la prédiction correspond à une avance (`A`) ou un retard (`R`). Cela est basé sur la valeur de la prédiction, où une valeur négative représente une avance et une valeur positive un retard.

Une fois les couleurs et les symboles déterminés, le front-end appelle la méthode `get_station_suivante(rame)` pour identifier la station suivante. La création de la rame, incluant la station, la couleur et le symbole, est ensuite effectuée (`création_rame`). L'animation de la rame est lancée, représentant visuellement son déplacement le long de la ligne simulée. L'utilisateur a également la possibilité de cliquer sur une rame pour afficher des informations détaillées de prédiction, telles

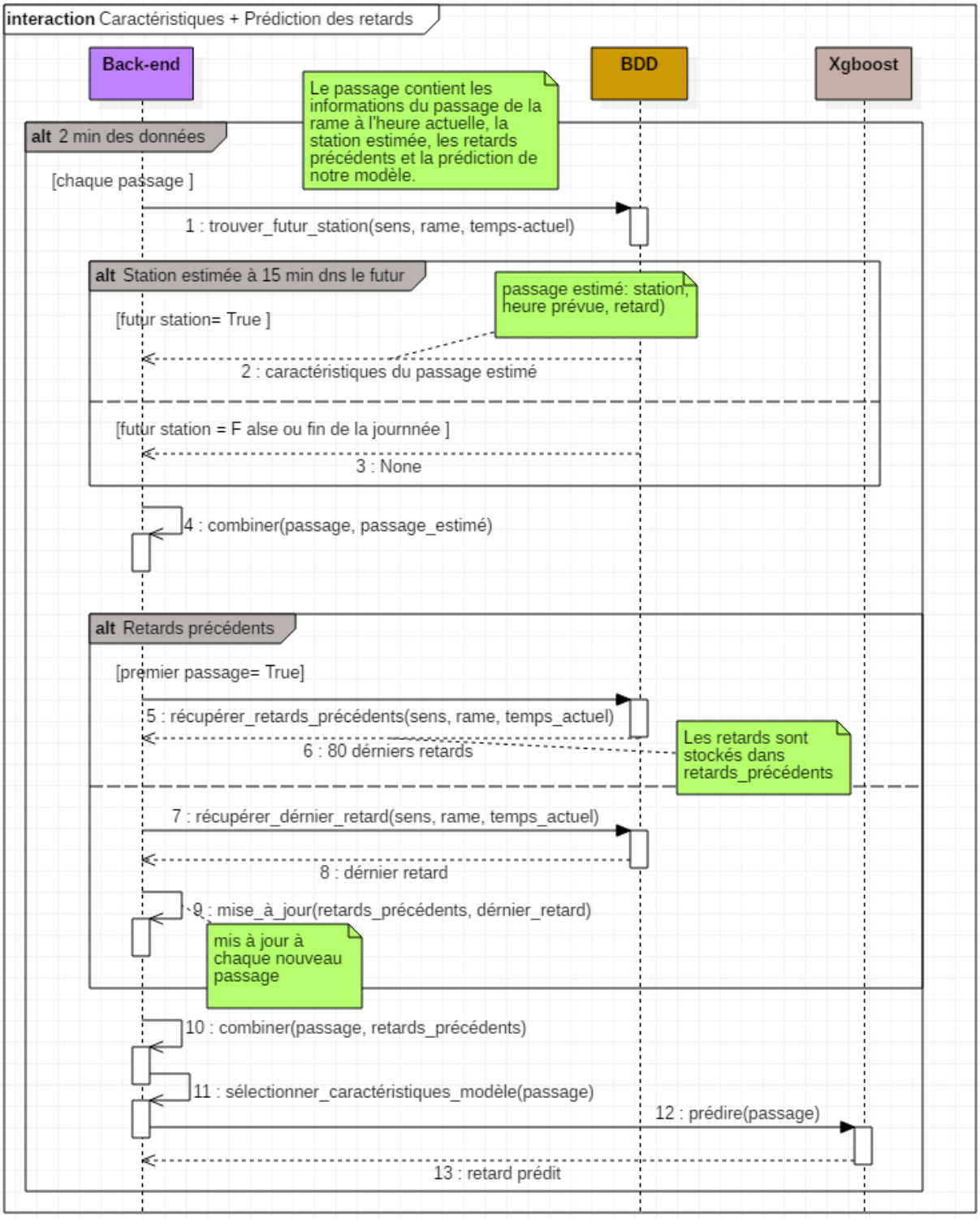


Figure 5.8 – Diagramme de séquence du processus de prédiction des retards

que la station impactée et le retard prédit, comme illustré dans le diagramme 5.10.

La section suivante présente les cas d'utilisation de notre interface homme-machine (IHM) appliqués à différents scénarios ferroviaires (différents jours et heures). Cette démonstration met en évidence les fonctionnalités clés de l'interface à travers des captures d'écran illustratives, montrant l'interaction entre les utilisateurs et le système pour faciliter la gestion des opérations ferroviaires.

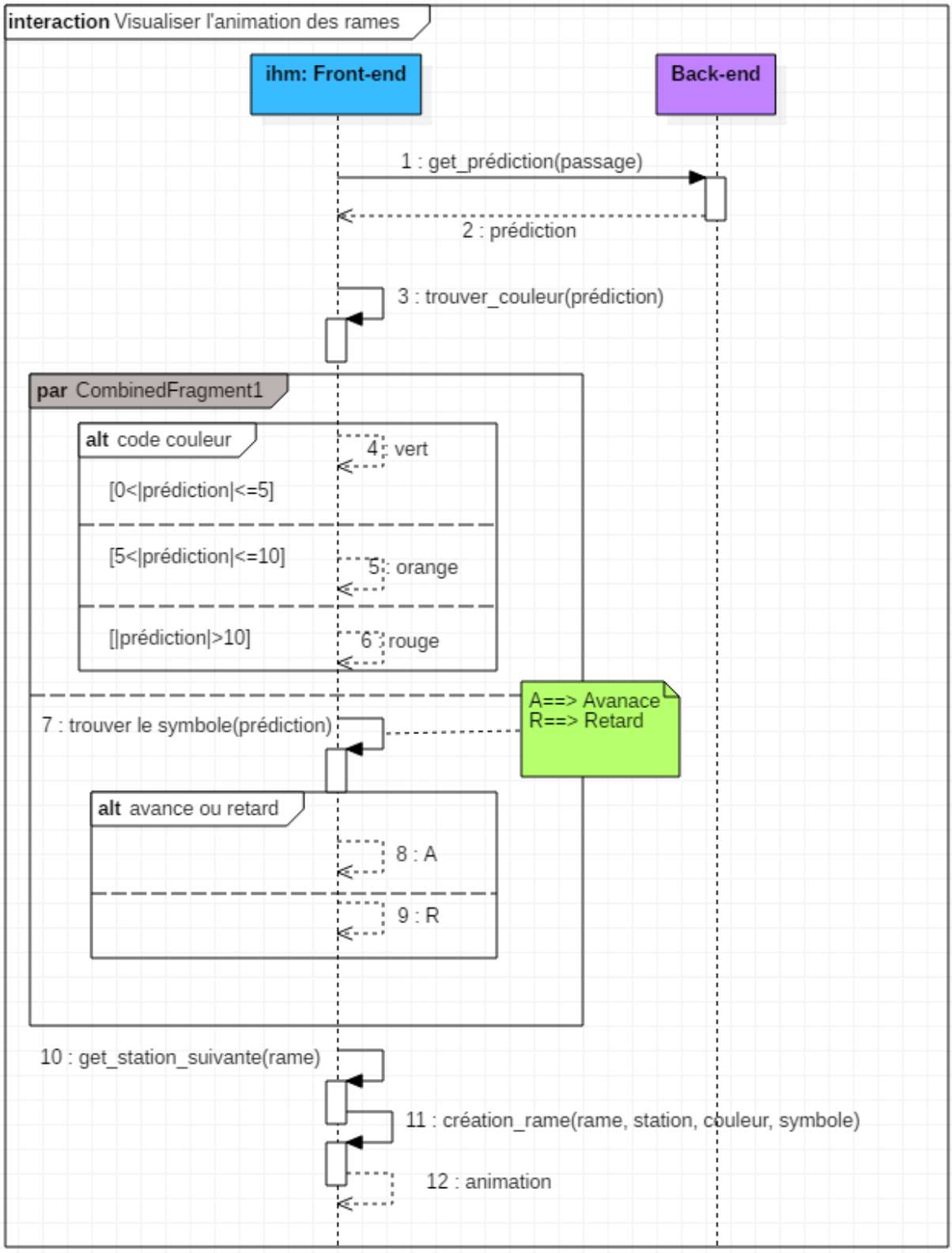


Figure 5.9 – Diagramme de séquence pour l'animation des rames avec un code couleur.

5.4 Mise en œuvre sur un cas test : Scénario simplifié

Les cas d'utilisation présentés dans cette section sont sélectionnés de manière à couvrir des journées et des heures différentes pour démontrer la capacité de notre

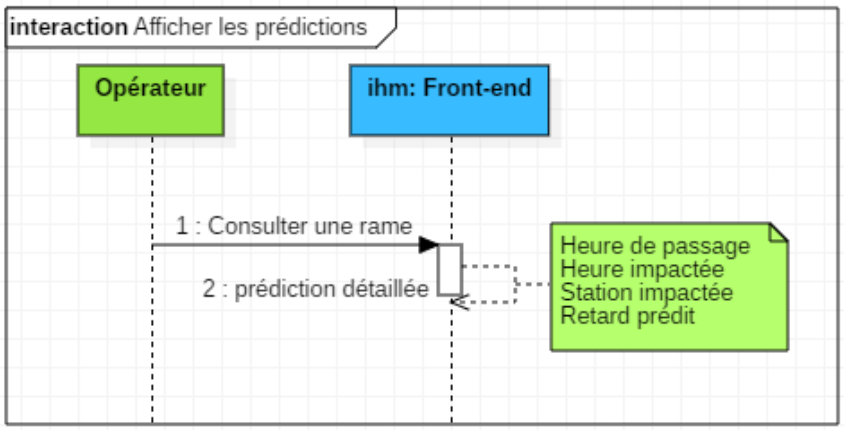


Figure 5.10 – Diagramme de séquence pour l'affichage des prédictions détaillés pour chaque rame.

interface à prédire les retards en temps réel. Les périodes d'analyse choisies s'étendent du 01 mai 2018 au 30 juin 2018, en utilisant les données de Tramway. En raison de la complexité d'intégrer et de visualiser l'ensemble de cette période sur un support papier, nous avons opté pour une approche ciblée en sélectionnant cinq plages horaires, chacune correspondant à une simulation d'une heure. Pour garantir une diversité représentative, les heures à tester sont choisies aléatoirement par ChatGPT [**WikipediaChatGPT**], tout en respectant la contrainte de couvrir cinq journées et cinq heures différentes. Chaque cas d'utilisation offre ainsi une compréhension claire et visuelle de l'apport de l'interface dans l'accès aux données et la prédiction en temps réel. Ce choix permet de montrer la flexibilité et la précision de l'interface dans divers contextes tout en maintenant une présentation compréhensible et concise.

5.4.1 Interface d'utilisation

Lors du lancement de l'application sur un PC, la première interface affichée correspond à celle décrite dans le diagramme 5.7. Cette interface permet de configurer les caractéristiques de la simulation en sélectionnant les paramètres appropriés. L'opérateur (utilisateur) peut choisir une date spécifique à simuler depuis la base de données (BDD) et définir une plage horaire précise en spécifiant une heure de début et une heure de fin pour la simulation, tout en respectant le format proposé sur l'écran d'accueil de l'interface. Comme illustré dans la figure 5.11, ces paramètres sont essentiels pour configurer la période de simulation et les détails temporels, garantissant une exécution conforme des scénarios de prédiction. De plus, un facteur d'accélération peut être défini, permettant de simuler les événements à une vitesse accrue, évitant ainsi d'attendre que la simulation se déroule en temps réel. Ce paramètre d'accélération offre une flexibilité et une personnalisation supplémentaires, permettant d'adapter la simulation aux besoins d'analyse spécifiques de l'utilisateur.

Une fois les configurations correctement définies, le bouton "Démarrer" s'active. En cliquant sur ce dernier, la simulation débute avec la plage horaire spécifiée. L'interface, décrite dans la figure 5.13, s'affiche alors à l'écran, et la simulation commence. Durant cette simulation, les stations ainsi que la circulation des rames entre les stations et au sein des stations sont représentées, avec un code couleur indiquant les prédictions de retard.

Bien que la simulation ait initialement été planifiée pour une durée de 2 heures, 2 minutes et 2 secondes, elle s'est déroulée en 30 minutes et 30 secondes grâce à l'application d'un facteur d'accélération de 30 (voir équation 5.1). La figure 5.14 illustre les informations détaillées sur les prédictions de retard pour chaque rame. Ces informations incluent l'ID de la rame (voiture), l'heure réelle, le retard prédit, la station impactée (station+15), ainsi que l'heure prévue pour l'arrivée de la rame à cette station estimée.

À la fin de la simulation, un onglet, illustré dans la figure 5.15, s'affiche pour indiquer que la simulation est terminée. Ce dernier offre également la possibilité de télécharger les données générées par la simulation. Une fois le téléchargement terminé, un autre onglet (figure 5.17) s'affiche pour confirmer que le téléchargement a été effectué avec succès.

Le fichier CSV 5.16 résultant contient plusieurs éléments à analyser, tels que :

— **DatePrévue** : La date de départ prévue.
— **DateRéelle** : La date de départ observée.
— **Sens** : Le sens de circulation de la rame.
— **Rame_ID** : L'identifiant unique de la rame.
— **Station_ID** : L'identifiant de la station actuelle.
— **Retard** : Le retard mesuré à la station actuelle.
— **Station_15** : La station estimée dans les 15 prochaines minutes au minimum.
— **HeurePrévue_15** : L'heure de départ prévue à la station_15.
— **HeureRéelle_15** : L'heure de départ observée à la station_15.
— **Retard_15** : Le retard mesuré à la station_15.
— **Retard_Prédit** : Le retard prédit par le modèle à la station_15.

Ces données permettent une analyse approfondie des performances du système, notamment la précision des prédictions de retard et l'efficacité des mesures de régulation appliquées.

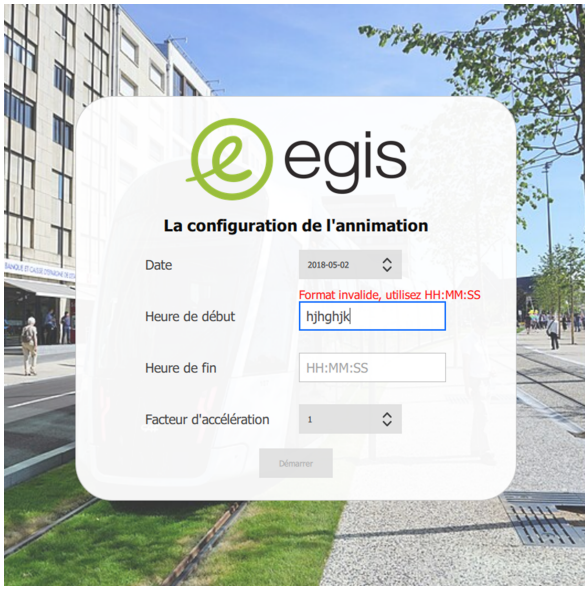


Figure 5.11 – Interface de configuration de la simulation avec des erreurs de saisie (format incorrect pour l'heure)



Figure 5.12 – Interface de configuration de la simulation avec des valeurs correctement saisies

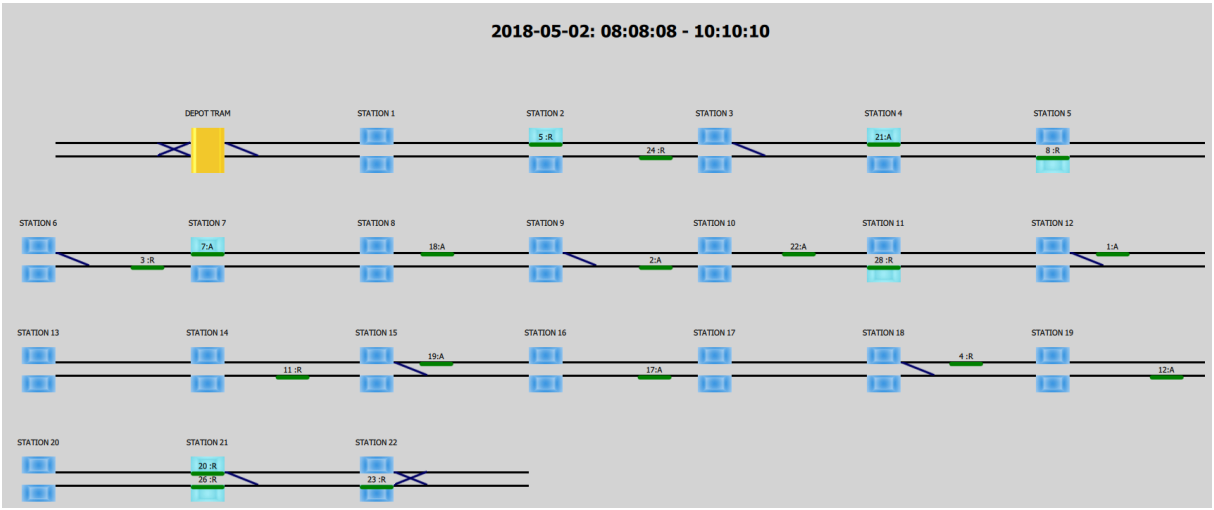


Figure 5.13 – Vue de la simulation en cours, illustrant les prédictions en temps réel

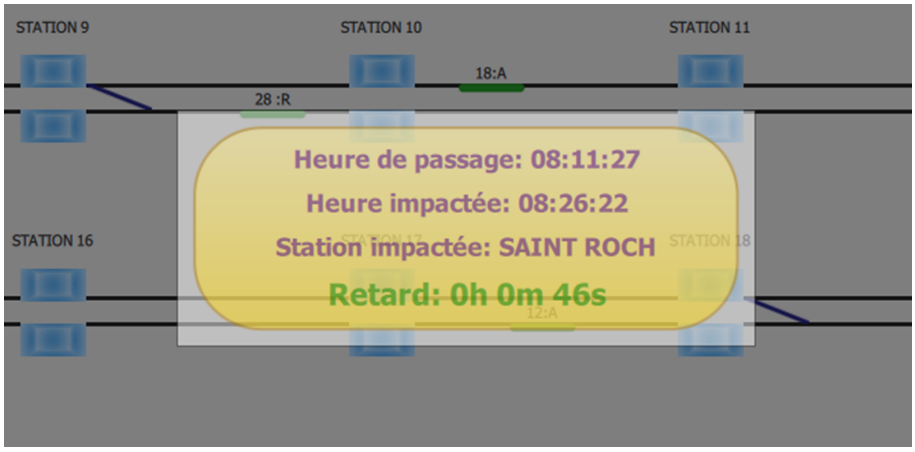


Figure 5.14 – Étiquette de prédiction affichée sur une rame retardée, indiquant les informations sur la station estimée et les détails du retard.

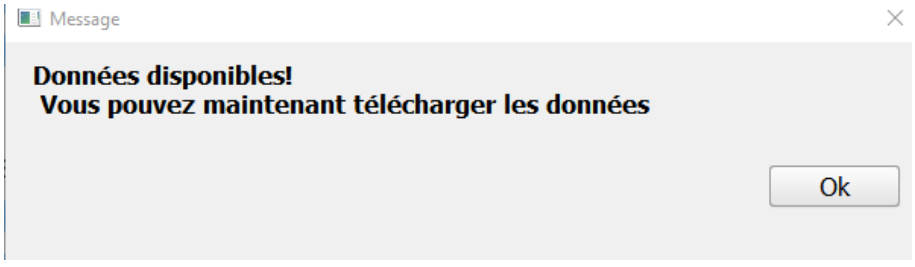


Figure 5.15 – Interface de téléchargement des données simulées

Date de creation: 2024-11-15 11:05:55

DatePrevue,DateReel,sens,rame_id,station_id,retard,station_15,HeurePrevue_15,retard_15,retard_predit

2018-05-02 08:47:00,2018-05-02 08:45:53,R,26,5,-67,8,08:57:24,-96,102.05241394042969
2018-05-02 08:46:00,2018-05-02 08:45:56,A,8,9,-4,9,08:45:56,-4,105.27818298339844
2018-05-02 08:45:19,2018-05-02 08:45:58,R,28,17,39,14,08:58:36,36,103.95191955566406
2018-05-02 08:47:00,2018-05-02 08:45:59,R,4,8,-61,8,08:45:59,-61,100.10867309570312
2018-05-02 08:46:00,2018-05-02 08:46:05,R,17,6,5,20,09:00:43,1,101.72291564941406
2018-05-02 08:46:00,2018-05-02 08:46:10,R,3,9,10,21,08:58:12,132,73.04798889160156
2018-05-02 08:44:00,2018-05-02 08:46:15,A,15,21,135,9,08:56:57,117,99.25243377685547
2018-05-02 08:44:00,2018-05-02 08:46:20,R,11,21,140,5,09:01:06,66,107.57481384277344
2018-05-02 08:45:00,2018-05-02 08:46:28,A,5,0,88,9,09:00:07,67,68.115234375
2018-05-02 08:45:00,2018-05-02 08:46:40,A,21,10,100,19,09:00:38,158,29.68746566772461
2018-05-02 08:45:49,2018-05-02 08:46:40,A,24,18,51,9,08:50:38,38,68.99723052978516
2018-05-02 08:46:11,2018-05-02 08:46:42,R,23,1,31,8,08:55:16,16,78.77777862548828
2018-05-02 08:44:00,2018-05-02 08:46:42,A,13,14,162,18,09:01:22,153,85.88560485839844
2018-05-02 08:47:00,2018-05-02 08:46:43,A,22,7,-17,14,09:00:50,-10,25.818330764770508
2018-05-02 08:47:00,2018-05-02 08:45:53,R,26,5,-67,8,08:57:24,-96,59.32469177246094
2018-05-02 08:46:00,2018-05-02 08:45:56,A,8,9,-4,9,08:45:56,-4,73.13421630859375
2018-05-02 08:45:19,2018-05-02 08:45:58,R,28,17,39,14,08:58:36,36,66.47593688964844
2018-05-02 08:47:00,2018-05-02 08:45:59,R,4,8,-61,8,08:45:59,-61,80.87897491455078

Figure 5.16 – Données générées par la simulation, prêtes à être analysées

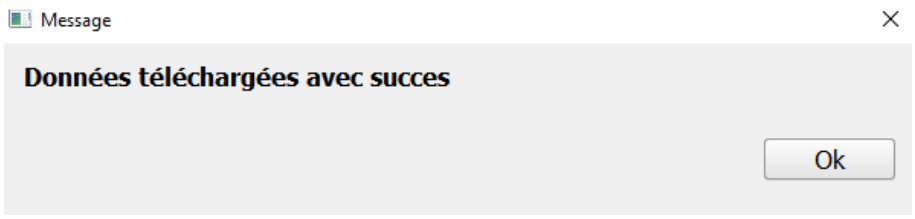


Figure 5.17 – Confirmation du téléchargement des données simulées

5.4.2 Analyse de la simulation ferroviaire issue des recommandations de ChatGPT

Cinq plages horaires sont choisies par ChatGPT pour tester la robustesse de l'application sur diverses configurations, en suivant la procédure expliquée précédemment. Le but est d'obtenir des fichiers CSV permettant d'analyser le contenu et les résultats.

Après avoir demandé à ChatGPT de sélectionner cinq plages horaires sur différents jours, en tenant compte des jours fériés, comme illustré dans la capture d'écran 5.18, la réponse présentée dans la capture 5.19 inclut deux jours fériés : le 8 mai (Fête de la Victoire) avec une plage horaire en matinée et le 27 mai (Pentecôte) couvrant une plage autour de midi. Les trois autres jours sélectionnés sont des jours ordinaires, dont deux tombent sur des week-ends (les 10 juin et 24 juin, qui sont des dimanches), avec des plages horaires variées telles que la matinée, la soirée et la nuit. Cette sélection garantit une diversité de scénarios pour évaluer l'IHM de manière exhaustive.

En général, même si le modèle présente certaines limitations en début de journée dues à l'absence de données historiques, il parvient tout de même à fournir des prédictions pertinentes. La simulation affiche également les codes couleur corrects pour indiquer si un train sera fortement en retard ou non. Globalement, les retards observés en début de journée ne sont pas significatifs, permettant au modèle de bien se comporter sur ces plages horaires.

Après avoir simulé et généré les cinq fichiers CSV correspondant à des plages horaires, les résultats obtenus sont résumés dans le tableau et la figure présentés. Le

tableau 5.1 met en évidence deux indicateurs clés : la RMSE, qui mesure l'écart moyen entre les prédictions et les valeurs réelles, et le pourcentage de précision (P%), calculé avec une marge d'erreur définie entre $[-150, +150]$. Chaque plage horaire simulée est associée à une date précise, où les simulations ont été effectuées avec un facteur d'accélération de 30. Ce facteur réduit significativement le temps de simulation à 15 minutes par heure tout en préservant une précision représentative des conditions réelles. Cette méthode accélérée permet d'obtenir des prédictions de retards fiables grâce au modèle XGBoost, qui s'est avéré performant sur les différentes plages horaires, avec une précision variant entre un minimum de 89% et un maximum de 95%.

La figure 5.20 complète les résultats en fournissant une visualisation détaillée des performances du modèle pour chaque plage horaire. Elle illustre la répartition des prédictions en pourcentage (P%) par rapport à l'intervalle d'erreur spécifié $[-150, +150]$. En début de journée, certaines limitations du modèle sont visibles, principalement dues à un manque de données historiques, ce qui entraîne une dispersion plus importante des prédictions. Toutefois, au fil des heures, la qualité des prédictions s'améliore, notamment grâce à l'accumulation des données en temps réel. Globalement, les retards prédits en début de journée sont souvent négligeables, ce qui permet au modèle de maintenir de bonnes performances et de fournir des informations pertinentes pour la gestion opérationnelle des retards ferroviaires.

Une attention particulière est portée à la journée du 15-05-2018, un mardi considéré comme une journée ordinaire. Durant la plage horaire du soir, on observe une dispersion significative des prédictions, avec un nombre élevé de points situés en dehors de l'intervalle d'erreur PCC. Ce comportement diffère des autres plages horaires et reflète des difficultés du modèle à prédire des retards importants. Néanmoins, cette anomalie reste négligeable lorsque l'on considère le nombre total de rames circulant à ce moment précis. La précision moyenne des prédictions pour cette plage est de 90%, bien qu'elle soit associée à la valeur la plus élevée de RMSE (109.55). Ce résultat souligne que le modèle rencontre des limites pour prédire correctement les retards extrêmes, en raison du manque d'exemples similaires dans les données d'entraînement. Ces difficultés pourraient également être dues à la nature spécifique des retards observés, tels que des incidents imprévus ou des congestions inhabituelles.

Il est également important de noter que cette journée figure parmi celles ayant une faible corrélation entre les données passées et futures, comme présenté dans la figure 4.20. Ces observations suggèrent que l'enrichissement du modèle avec des données supplémentaires, incluant des informations détaillées sur les types de retards et leurs causes, pourrait significativement améliorer les performances du modèle. Malgré ces limitations, le modèle reste fiable pour prédire les retards réguliers, offrant ainsi un outil efficace pour la gestion quotidienne des retards ferroviaires.

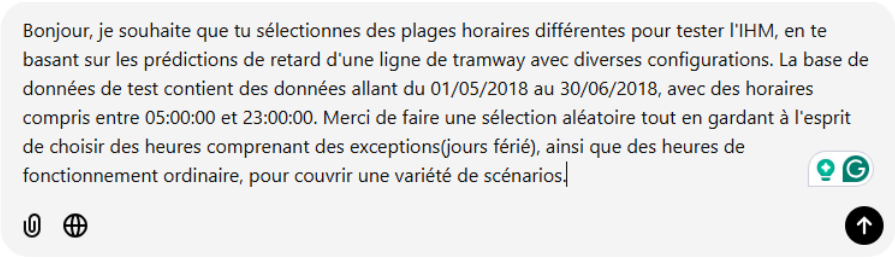
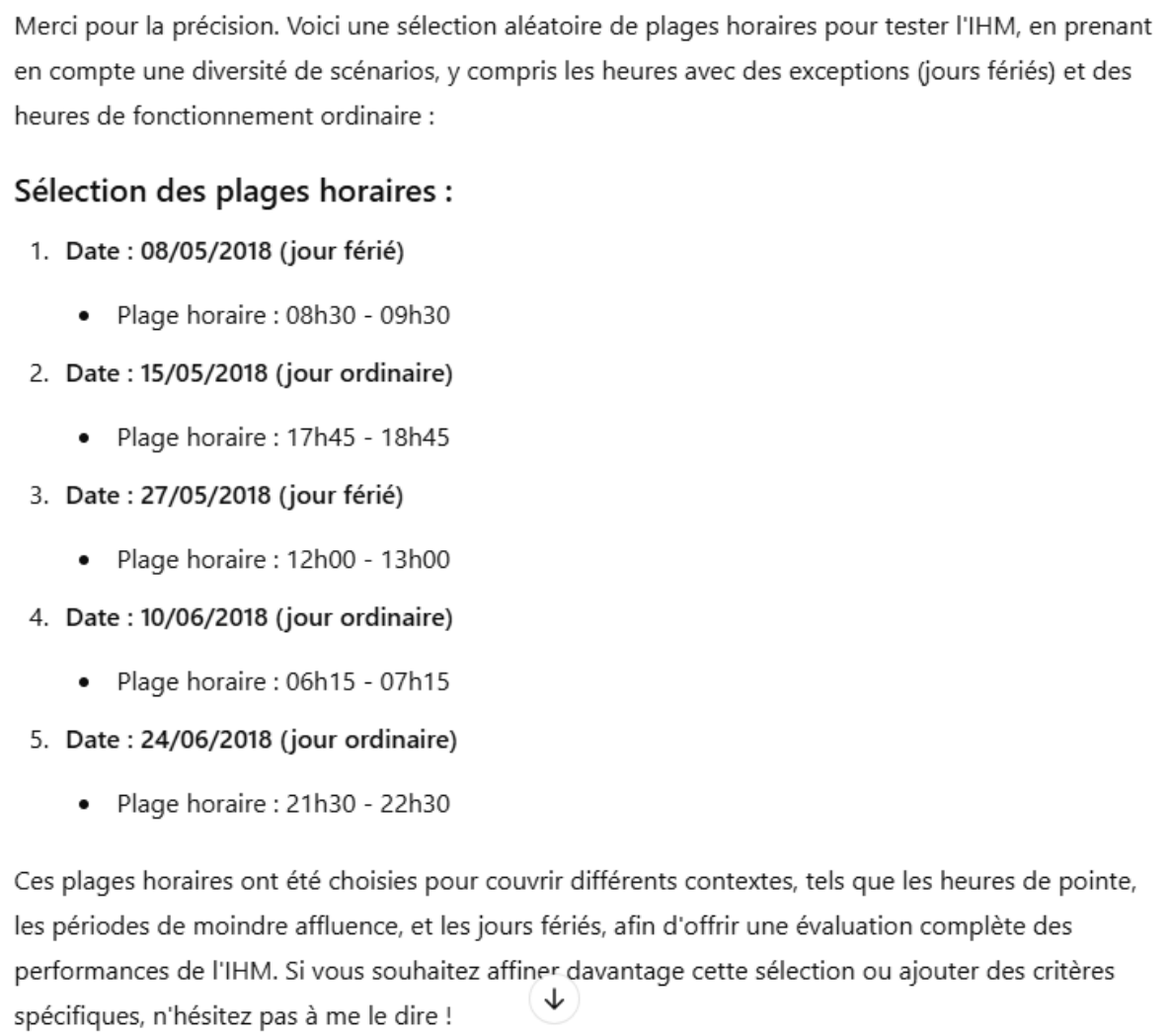


Figure 5.18 – Capture d'écran du message écrit à ChatGPT pour déterminer les plages horaires à analyser



Merci pour la précision. Voici une sélection aléatoire de plages horaires pour tester l'IHM, en prenant en compte une diversité de scénarios, y compris les heures avec des exceptions (jours fériés) et des heures de fonctionnement ordinaire :

Sélection des plages horaires :

1. **Date : 08/05/2018 (jour férié)**
   - Plage horaire : 08h30 - 09h30
2. **Date : 15/05/2018 (jour ordinaire)**
   - Plage horaire : 17h45 - 18h45
3. **Date : 27/05/2018 (jour férié)**
   - Plage horaire : 12h00 - 13h00
4. **Date : 10/06/2018 (jour ordinaire)**
   - Plage horaire : 06h15 - 07h15
5. **Date : 24/06/2018 (jour ordinaire)**
   - Plage horaire : 21h30 - 22h30

Ces plages horaires ont été choisies pour couvrir différents contextes, tels que les heures de pointe, les périodes de moindre affluence, et les jours fériés, afin d'offrir une évaluation complète des performances de l'IHM. Si vous souhaitez affiner davantage cette sélection ou ajouter des critères spécifiques, n'hésitez pas à me le dire !

Figure 5.19 – Capture d'écran de la réponse de ChatGPT pour la sélection des plages horaires

5.4.3 Analyse des contraintes et évaluation de l'IHM (Questionnaire destiné aux anciens opérateurs et experts en systèmes ferroviaires d'Egis Rail

L'évaluation de l'IHM, intégrant la conception d'un écran supplémentaire destiné aux PCCs, a été réalisée à l'aide d'un questionnaire adressé aux opérateurs d'Egis. Cette démarche a pour objectif principal d'évaluer l'efficacité de l'interface actuelle et de mettre en lumière les points faibles nécessitant des ajustements ou des améliorations.

L'objectif de cette évaluation est de recueillir des retours détaillés sur l'interface utilisateur et les fonctionnalités proposées par l'application. Les réponses obtenues permettent d'identifier les points forts à valoriser, les faiblesses à corriger, ainsi que les pistes d'amélioration pour mieux répondre aux attentes des utilisateurs et s'intégrer dans les processus opérationnels existants.

Le questionnaire a été conçu pour examiner plusieurs aspects cruciaux de l'IHM. Il évalue la facilité d'utilisation, la lisibilité et la clarté des informations affichées, notamment les codes couleur et les prédictions de retard. La précision des prédictions est

Plage Horaire	RMSE	P (%)
08-05-2018 08h30-09h30	82.95	95%
15-05-2018 17h45-18h45	109.55	90%
27-05-2018 12h00-13h00	94.58	91%
10-06-2018 06h15-07h15	79.69	89%
24-06-2018 21h30-22h30	84.74	89%

Table 5.1 – Les résultats avec plage horaire, RMSE et P (%)

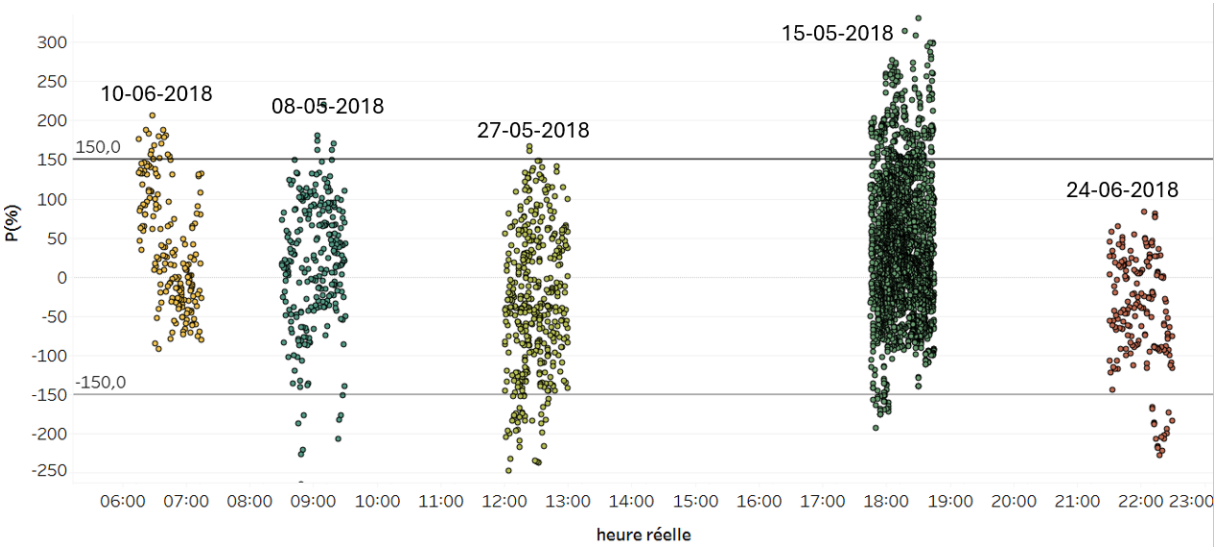


Figure 5.20 – Visualisation des prédictions dans l'intervalle d'erreur autorisé pour le PCC[-150 :150]

analysée pour juger de leur fiabilité et de leur utilité dans la prise de décision en temps réel.

Les retours obtenus grâce à cette évaluation permettent de prioriser les ajustements nécessaires. Parmi les améliorations identifiées, on peut noter l'optimisation de l'agencement graphique pour une meilleure lisibilité, des ajustements fonctionnels pour affiner la précision des prédictions, ou encore l'intégration de nouvelles fonctionnalités pour renforcer l'utilité opérationnelle de l'application. Cette démarche d'amélioration continue garantit que l'application répondra pleinement aux besoins des opérateurs tout en optimisant la gestion des retards.

la figure 5.21 met en évidence plusieurs aspects importants concernant l'évaluation de l'IHM par les opérateurs. Tout d'abord, il ressort que la majorité des opérateurs est familière avec les systèmes d'aide à la gestion des retards. Cette expérience préalable est un atout pour évaluer objectivement l'application. Cependant, peu d'entre eux ont déjà utilisé des outils similaires, ce qui pourrait influencer leur perception de l'intuitivité de l'interface.

En ce qui concerne l'intuitivité de l'interface, les réponses montrent une satisfaction modérée. Les opérateurs jugent l'IHM principalement "intuitive" mais rarement "très intuitive", ce qui laisse entrevoir un potentiel d'amélioration dans ce domaine. Par ailleurs, les codes couleur utilisés pour représenter les prédictions sont relativement

bien compris, avec des réponses majoritairement positives. Cependant, quelques ajustements pourraient encore renforcer leur lisibilité et leur compréhension immédiate.

La lisibilité globale des informations affichées sur l'IHM est identifiée comme un point faible. Plusieurs opérateurs considèrent l'affichage comme "moyennement lisible". Ce retour souligne la nécessité d'une refonte graphique ou d'une meilleure hiérarchisation visuelle des données pour améliorer l'ergonomie de l'interface.

La simulation est globalement perçue comme "plutôt réaliste", mais des écarts avec les conditions réelles sont signalés par certains opérateurs. Cela suggère que les paramètres de simulation doivent être ajustés. En ce qui concerne la précision des prédictions, les réactions ont été mitigées, certains opérateurs estimant qu'il est difficile de l'évaluer en l'absence d'essais dans un environnement opérationnel réel. Cela souligne l'importance d'une validation plus poussée pour affiner la fiabilité des résultats.

Un autre point soulevé concerne l'horizon de prédiction de 15 minutes, qui est jugé partiellement suffisant par plusieurs opérateurs. Certains suggèrent d'allonger cet horizon pour permettre des décisions plus efficaces dans un contexte opérationnel. Quant à l'interprétation des résultats, elle est jugée "assez facile" par la majorité, bien qu'une simplification supplémentaire pourrait encore en améliorer l'efficacité.

Enfin, en ce qui concerne la satisfaction globale, les opérateurs se déclarent majoritairement "satisfaits", mais les retours indiquent qu'ils ne recommanderaient pas immédiatement l'application sans ajustements. Cela met en lumière l'importance d'apporter des améliorations ciblées avant de considérer une implémentation opérationnelle à grande échelle.

En conclusion, cette analyse met en lumière des points forts et des axes d'amélioration, notamment en ce qui concerne la conception des codes couleur et la perception réaliste de la simulation. L'horizon de prédiction de 15 minutes, bien qu'acceptable, pourrait être optimisé avec l'intégration de nouvelles données. Par ailleurs, des améliorations sont nécessaires en matière de lisibilité des informations (design de l'interface) et de précision des prédictions. Il est important de préciser que l'évaluation ergonomique réalisée dans le cadre de ce projet constitue une première étape. Le questionnaire a été mené auprès d'anciens opérateurs et d'experts en systèmes ferroviaires d'Egis Rail, dont l'expérience a permis d'identifier des besoins et des pistes d'amélioration. Cependant, cette approche présente des limites, car elle ne permet pas à elle seule de garantir une intégration optimale de l'outil proposé au sein des postes opérateurs. Il sera donc nécessaire de poursuivre cette démarche par des évaluations ergonomiques plus approfondies, incluant l'analyse complète de l'activité des opérateurs en poste, de leurs méthodes de travail, de la co-activité et des procédures existantes. L'utilisation de scénarios opérationnels ou de cas d'usage sera également essentielle pour évaluer la pertinence de l'outil dans des conditions réalistes. En outre, le code couleur proposé devra être harmonisé avec celui des autres outils existants afin d'assurer une cohérence visuelle et fonctionnelle. L'ergonomie étant une discipline à part entière et ne constituant pas le cœur de cette thèse, ces approfondissements seront essentiels lors des phases ultérieures de développement pour garantir une intégration réussie dans les processus opérationnels

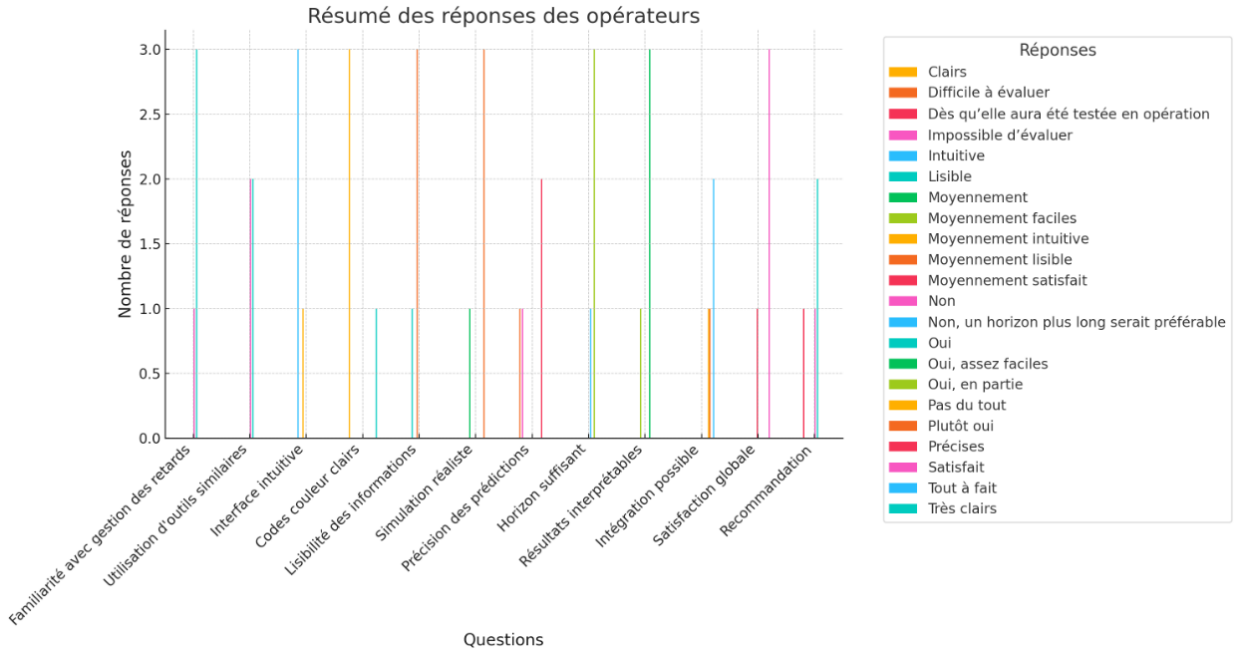


Figure 5.21 – Analyse des réponses des opérateurs sur l'évaluation de l'application de simulation et de prédiction des retards

5.5 Conclusion

La conception de l'IHM, incluant un écran supplémentaire destiné à être intégré dans le PCC, s'inspire des écrans existants dans ces centres afin de faciliter son intégration dans le système opérationnel. Cette IHM repose sur une architecture simple composée de trois éléments principaux : un front-end, qui représente l'interface utilisateur et son visuel ; un back-end, contenant les modèles d'apprentissage automatique pour générer les prédictions et les afficher avec un code couleur sur le front-end ; et une base de données, qui conserve les données de test issues des opérations tramway, permettant d'analyser les performances sur les deux derniers mois.

En raison de l'absence d'accès direct au système ferroviaire en temps réel, il a été nécessaire de développer cette IHM comme un premier prototype. Ce dernier recrée des conditions similaires à celles d'un PCC, notamment la fonctionnalité de récupération des passages des rames sur les stations toutes les deux secondes, garantissant ainsi une fluidité des flux de données proche de la réalité. Des diagrammes de séquence ont été conçus pour expliquer les principes et les flux d'information entre le front-end, le back-end, la base de données, et le modèle d'apprentissage utilisé XGBOOST. Plusieurs fonctions ont été développées pour retraiter les données, notamment pour alimenter le modèle en prenant en compte les retards précédents à chaque passage, ainsi que pour animer les rames afin de simuler des circulations basées sur des conditions réelles.

Une évaluation de cette IHM a été réalisée à l'aide d'une fiche d'évaluation auprès de quatre opérateurs d'Egis-Rail. Cette évaluation visait à analyser plusieurs aspects de l'interface : facilité d'utilisation, précision des prédictions, intuitivité, et efficacité. Elle a permis de mettre en évidence les points forts de l'IHM, ainsi que ses faiblesses, qui devront être corrigées dans le cadre des futurs travaux. Parmi les problèmes identifiés

figurent des erreurs dans la gestion des circulations des rames, le design de l'interface, ainsi que la lisibilité de certains éléments.

Ce premier prototype constitue une étape importante vers la conception d'une IHM fonctionnelle et optimisée, intégrant les retours des opérateurs pour répondre aux besoins des centres de contrôle ferroviaire.

6

Conclusion et Perspectives

Sommaire

5.1 Conception de l'interface homme machine 136

5.1.1 Maquette 137

5.1.2 Fonctionnalités clés avec diagramme des cas d'utilisation 138

5.2 Architecture générale du prototype complet (IHM) 139

5.3 Phase de Développement du Prototype 141

5.3.1 Diagramme de séquence : Création de la ligne à simuler 142

5.3.2 Diagramme de séquence : Caractéristiques de la simulation . . 144

5.3.3 Diagramme de séquence : Caractéristiques + Prédiction des retards 145

5.3.4 Diagramme de séquence : Animation des rames avec un code couleur 146

5.4 Mise en œuvre sur un cas test : Scénario simplifié 148

5.4.1 Interface d'utilisation 149

5.4.2 Analyse de la simulation ferroviaire issue des recommandations de ChatGPT 152

5.4.3 Analyse des contraintes et évaluation de l'IHM (Questionnaire destiné aux anciens opérateurs et experts en systèmes ferroviaires d'Egis Rail 154

5.5 Conclusion 157

Au cours de cette thèse, nous avons proposé une méthodologie reposant sur l'intégration d'un Système d'Aide à la Prise de Décision (SAPD) dans les postes de contrôle ferroviaires et les réseaux de transport public. Cette approche inclut la conception d'une interface homme-machine intuitive et facile à utiliser (prototype), s'inspirant des écrans déjà existants dans les PCCs, afin de garantir une transition fluide vers l'intégration de ce système, basé sur des modèles d'intelligence artificielle (IA) et les données d'exploitation.

6.1 Contributions

Pour répondre aux problématiques liées à l'aide à la décision dans les postes de commandement centralisés (PCCs), nous avons développé une méthodologie

innovante qui s'appuie sur les données d'exploitation liées à la circulation des rames. Cette approche vise à estimer et prédire les retards quotidiens des trains en intégrant des modèles d'apprentissage, associés à une méthode avancée d'ingénierie des caractéristiques. Afin de garantir une performance robuste des modèles, cette approche a été validée sur deux ensembles de données distincts, tout en tenant compte des spécificités de chacun. Les résultats montrent que cette combinaison de données opérationnelles, de modèles d'IA, et d'une ingénierie des caractéristiques adaptée permet d'améliorer la précision des prédictions et de renforcer l'efficacité des décisions prises dans les PCCs.

Données d'exploitation utilisées : Deux ensembles de données ont été utilisés, le premier issu de l'infrastructure Tramway et le second de l'infrastructure RER, chacun présentant des caractéristiques différentes. Ces données incluent principalement les heures de départ prévues et réelles, ainsi que l'écart entre les deux, représentant les retards. Pour enrichir ces ensembles, des données contextuelles (jours fériés, météo, week-ends, etc.) ont également été intégrées.

Principe du passé impacte le présent, qui impacte le futur : Parmi les contraintes rencontrées figurait la qualité des données brutes, qui n'intégraient pas directement les facteurs influençant les retards quotidiens. Pour résoudre ce problème, une méthode d'ingénierie des caractéristiques a été appliquée. Cette méthode s'appuie sur une fenêtre glissante qui capture le comportement des rames dans le passé récent, permettant d'analyser l'impact du passé sur le présent et le futur.

Horizon de prédiction à court terme : Après avoir validé l'influence du passé récent à travers des analyses de corrélation et de l'importance des variables, plusieurs tests ont été réalisés pour déterminer l'horizon temporel optimal pour les modèles d'apprentissage. Les résultats montrent que lorsque l'horizon de prédiction est élargi, la performance du modèle diminue, tandis qu'une réduction de cet horizon améliore les performances. Un horizon de 15 minutes a été fixé, ce délai étant suffisant pour permettre aux opérateurs (opérateurs) d'utiliser efficacement les prédictions pour la prise de décision.

Meilleur modèle d'apprentissage (XGBoost) : Une étude comparative a été menée pour sélectionner le meilleur modèle d'apprentissage sur les ensembles de données d'exploitation. Parmi les modèles de machine learning et de deep learning testés, XGBoost s'est avéré être le plus performant, avec une optimisation bayésienne de ses hyperparamètres.

Analyse de sensibilité : Une analyse de sensibilité a été effectuée pour valider la méthodologie sur différentes infrastructures et identifier ses limitations. Les caractéristiques spécifiques de chaque ensemble de données ont été comparées afin de valider le schéma global de l'analyse et de confirmer la robustesse de l'approche.

Interface homme-machine : Un prototype a été développé pour intégrer les prédictions à court terme du modèle XGBoost dans une interface utilisateur interactive. Cette interface inclut des animations en temps réel des rames avec un code couleur indiquant les retards. Conçue en réponse aux besoins des opérateurs

du PCC, elle combine les données d'exploitation en temps réel avec les prédictions du modèle, offrant ainsi une visualisation fidèle et lisible du système ferroviaire. Cette conception vise à faciliter l'intégration future de ce prototype dans les PCC.

6.2 Réflexion sur les défis rencontrés et les solutions apportées

L'intégration d'un SAPD dans un PCC a soulevé plusieurs défis au cours de cette thèse. Ces défis concernent divers domaines, notamment la quantité et la qualité des données disponibles dans le système ferroviaire, les problématiques spécifiques liées aux PCC, la résistance des OCCs à l'adoption d'un nouveau système, ainsi que les difficultés d'implémentation d'une interface répondant aux exigences des PCC.

Le premier défi, traité dans le chapitre 2, porte sur le manque de littérature scientifique spécifique à l'IA appliquée aux systèmes ferroviaires urbains. La majorité des articles disponibles se concentrent sur les systèmes ferroviaires régionaux, à grande vitesse, ou sur les réseaux de transport en commun. Cette lacune s'explique principalement par la sensibilité des données ferroviaires et leur accès limité. Ce contexte a rendu difficile l'identification d'approches adaptées à notre domaine d'étude.

Le deuxième défi concerne la qualité et la quantité des données disponibles. Dès le début de la thèse, deux objectifs principaux ont été définis : la gestion des retards et l'analyse des flux de voyageurs. Cependant, il est apparu que les données nécessaires, telles que celles relatives à la billetterie, au comptage des passagers et au poids des rames, étaient indisponibles (collaborateurs egis-rail). De plus, les données d'exploitation récupérées étaient limitées en quantité et présentaient des problèmes de qualité, nécessitant un prétraitement important. Des échanges avec les opérateurs ont été indispensables pour clarifier et résoudre les problèmes rencontrés avec ces ensembles de données brutes.

Le troisième défi réside dans l'absence d'accès direct à un PCC pour tester l'intégration du prototype d'interface, présenté dans le chapitre 5, avec des données en temps réel. Cette limitation a conduit à simuler un environnement réel en utilisant des données tests issues de l'entraînement du modèle d'apprentissage automatique. Cette approche a permis de reproduire des flux de données en temps réel afin de garantir des conditions d'évaluation réalistes.

Le quatrième défi, abordé dans la section 5.4.3, concerne l'évaluation du fonctionnement global de l'écran supplémentaire proposé. Quatre opérateurs ont été mobilisés pour tester et analyser l'interface développée. Les retours ont permis de mettre en évidence des points nécessitant des améliorations, notamment la finalisation de l'interface et une collaboration accrue avec les opérateurs afin d'assurer une satisfaction optimale avec le système proposé dans cette thèse.

Enfin, le cinquième défi concerne l'évaluation de l'efficacité de la méthodologie proposée dans cette thèse, qui repose sur le concept « le passé impacte le présent, et le présent impacte le futur ». Ce défi soulève la question de la validité de cette approche pour l'ensemble des infrastructures ferroviaires de transport public. Une étude de sensibilité a été réalisée à cet effet et est présentée dans la figure 4.28. Cependant, en raison des nombreux défis précédemment évoqués, il n'a pas été possible de finaliser ce travail dans son intégralité.

6.3 Perspectives

Notre méthodologie va au-delà de la simple gestion des retards ; elle peut être enrichie pour inclure des fonctionnalités supplémentaires, telles que la prédiction du flux de voyageurs sur les quais et à bord des rames, en réponse à divers travaux ou perturbations (une perspective future prometteuse). Cela permettrait d'offrir aux opérateurs une vision globale et optimisée du trafic, pour une gestion encore plus efficace du réseau ferroviaire. Toutefois, ces travaux ouvrent également la voie à de nombreuses perspectives de recherche dans ce domaine. Parmi ces perspectives, plusieurs axes futurs peuvent être explorés :

Développement d'un outil de prétraitement des données brutes : Vise à garantir une meilleure qualité des données utilisées dans les PCC. Les problèmes rencontrés dans cette thèse sur les deux ensembles de données étudiés ont mis en lumière des lacunes significatives dans la qualité des données stockées dans les centres de contrôle ferroviaires. Ces données, souvent sous-utilisées, nécessitent des traitements pour être exploitables dans des analyses statistiques. Ainsi, la création d'un outil dédié au traitement et à la préparation des données constitue la première perspective de ce travail.

Validation et amélioration de la méthodologie sur différentes infrastructures ferroviaires public : Est une étape essentielle pour évaluer son efficacité. L'analyse de sensibilité présentée dans cette thèse n'a pas pu être finalisée en raison d'un manque de données disponibles sur plusieurs infrastructures. Une collecte de données d'exploitation plus riche et diversifiée, couvrant un plus grand nombre d'infrastructures, est la deuxième perspective majeure. Cela permettra d'élargir les analyses et de valider la méthodologie proposée dans des contextes variés au sein du domaine ferroviaire.

Intégration de nouvelles sources de données : telles que les données de billetterie, les images des caméras de comptage, les données énergétiques et celles relatives à la circulation des rames, constitue une autre perspective clé. Ces données permettront d'étudier et d'analyser l'influence des retards sur l'ensemble des informations liées à un trajet tout au long d'une journée. Par exemple, les données issues des caméras des quais ou des rames ouvrent des opportunités pour intégrer des techniques de traitement et d'analyse d'images. Ces approches pourraient servir à développer des modèles de prédiction à court et long terme sur les flux de voyageurs. Ces modèles pourraient être combinés et intégrés dans l'IHM, enrichissant ainsi les informations affichées sur l'écran supplémentaire des PCC.

Information voyageurs : Une autre perspective consiste à développer une application dédiée à l'information voyageurs, en complément ou en remplacement des systèmes PCC. Cette application pourrait fournir en avance des informations et des prédictions sur les retards ainsi que sur les flux de voyageurs pendant les heures de pointe ou creuses. L'objectif serait d'évaluer et d'améliorer l'expérience des usagers. Une telle approche pourrait constituer une solution « gagnant-gagnant » en apportant des bénéfices tant pour les opérateurs que pour les voyageurs.

bibliography.bib

Bibliographie

[1] *Investissements nets de la Société Nationale des Chemins de fer Français (SNCF) de 2008 à 2023*. 2020. url : https://fr.statista.com/statistiques/506167/sncf-investissements/.

[2] Autorité de la Qualité de Service dans les Transports (AQST). *Analyse des faits marquants du bilan 2023*. 2023. url : https://www.qualitetransports.gouv.fr/analyse-des-faits-marquants-du-bilan-2023-de-la-a457.html.

[3] SNCF Open Data. *Régularité mensuelle des TER*. 2024. url : https://ressources.data.sncf.com/explore/dataset/regularite-mensuelle-ter/.

[4] Ministère de la Transition Écologique et Solidaire. *Analyse des données ferroviaires en France*. 2023. url : https://www.qualitetransports.gouv.fr/.

[5] SNCF Open Data. *Données de circulation et de régulation ferroviaire*. 2023. url : https://ressources.data.sncf.com/.

[6] Ramesh Sharda, Dursun Delen et Efraim Turban. *Business Intelligence and Analytics : Systems for Decision Support*. Pearson Education, 2020.

[7] Histoire Image. "Les premiers chemins de fer en France". In : *Histoire et Patrimoine* (2024). url : https://histoire-image.org/etudes/premiers-chemins-fer.

[8] Wikipedia Contributors. "Chemin de fer de Petite Ceinture". In : *Wikipedia* (2024). url : https://en.wikipedia.org/wiki/Chemin_de_fer_de_Petite_Ceinture.

[9] Wikipedia Contributors. "Paris Tramways in the Belle Époque". In : *Wikipedia* (2024). url : https://en.wikipedia.org/wiki/Paris_in_the_Belle_%C3%89poque.

[10] Wikipedia Contributors. "Paris Métro History". In : *Wikipedia* (2024). url : https://en.wikipedia.org/wiki/Paris_M%C3%A9tro.

[11] Wikipedia Contributors. "Strasbourg Tramway History". In : *Wikipedia* (2024). url : https://en.wikipedia.org/wiki/Strasbourg_tramway.

[12] Wikipedia Contributors. "Le Havre's Old Tramway". In : *Wikipedia* (2024). url : https://en.wikipedia.org/wiki/Le_Havre%27s_old_tramway.

[13] Wikipédia. *Contrôle de trafic centralisé*. 2024. url : https://fr.wikipedia.org/wiki/Contr%C3%B4le_de_trafic_centralis%C3%A9.

[14] Alstom. "Centralized Traffic Control System". In : *Alstom Official Documentation* (2022). url : https://www.alstom.com/fr/controle-des-trains-la-performance-et-la-fiabilite-pour-linteroperabilite-des-systemes-ferroviaires-ertms-et-la-compatibilite-avec-les-systemes-de-pilotage-automatique-ato.

[15] John Smith. *Railway Traffic Control : Concepts and Evolution*. 2nd Edition. RailTech Publishing, 2018.

[16] Railway Technology Magazine. “Overview of Centralized Traffic Control Systems”. In : *Railway Technology* 12.3 (2020), p. 45-50. url : https://www.railway-technology.com/ctc-overview.

[17] Fabien Belmonte et al. “Évaluation du facteur humain dans le domaine de la supervision de trafic ferroviaire : le projet SPICA-Rail”. In : *Lambda-Mu 16ème Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement*. Avignon, France, 2008, Session 7A. url : https://hal.science/hal-00335053/document.

[18] Sarah Miller. “Automatic Train Protection Systems : Enhancing Safety in Modern Railways”. In : *Railway Technology Review* 14.3 (2019), p. 75-85.

[19] David Johnson et Michael Lee. *Integration of ATS into PCC : Challenges and Benefits*. 2nd Edition. RailTech Publishing, 2018.

[20] John Smith. “Advances in CBTC Technology for Railway Systems”. In : *Journal of Railway Engineering* 15.3 (2019), p. 123-136.

[21] Sarah Miller. “Train Protection Systems and Their Role in Railway Automation”. In : *Railway Safety Review* 12.4 (2020), p. 45-56.

[22] David Johnson. *Safety in Automated Metro Systems : Lessons Learned*. 2nd Edition. RailTech Publishing, 2021.

[23] Michael Lee et Aisha Patel. “Key Features and Benefits of CBTC Systems”. In : *Urban Transit Innovations* 10.2 (2022), p. 75-85.

[24] Pierre Dupont et Sophie Martin. “Integration of CBTC with Automatic Train Control Systems”. In : *International Conference on Railway Technology*. 2020, p. 200-215.

[25] Thomas Harris. “Emergency Braking Systems in CBTC : A Technical Overview”. In : *Transportation Systems Review* 8.1 (2021), p. 50-65.

[26] John Smith et Maria Lopez. “ERTMS : A Path to Standardizing European Railways”. In : *Railway Interoperability Journal* 9.3 (2018), p. 95-110.

[27] László Lindenmaier, István Ferenc Lővétei et Szilárd Aradi. “Efficient Real-time Rail Traffic Optimization : Decomposition of Rerouting, Reordering, and Rescheduling Problem”. In : *arXiv preprint arXiv :2209.12689* (2022). url : https://arxiv.org/abs/2209.12689.

[28] A. Gupta, X. Wang et H. Zhang. “Energy-efficient Timetable Design for Urban Rail Transit”. In : *arXiv preprint arXiv :2309.05489* (2023). url : https://arxiv.org/abs/2309.05489.

[29] M. Sander, M. Belloni et M. Bierlaire. “Robust Rail Infrastructure Design under Uncertain Timetables”. In : *arXiv preprint arXiv :2308.00483* (2023). url : https://arxiv.org/abs/2308.00483.

[30] WalkMe Team. *Le système d’aide à la décision et sa définition : un peu d’histoire*. 2024. url : https://www.walkme.com/fr/glossaire/systeme-aide-decision/.

[31] Efraim Turban, Ramesh Sharda et Dursun Delen. *Decision Support and Business Intelligence Systems*. Pearson Education, 2010.

[32] Daniel J. Power. “Decision Support Systems : Concepts and Resources for Managers”. In : *Quorum Books* (2002).

[33] Serge Hervé Djoufack. “Systèmes d’Aide à la Décision dans les environnements complexes : Applications et perspectives”. Consulté le 4 décembre 2024. Thèse de doct. Université de Technologie de Compiègne, 2022. url : https://theses.hal.science/tel-03670871v1/file/these.pdf.

[34] Ralph H. Sprague et Eric D. Carlson. “Building Effective Decision Support Systems”. In : *Prentice Hall* (1980).

[35] Wikipedia contributors. *Communication-Based Train Control (CBTC)*. 2024. url : https://en.wikipedia.org/wiki/Communication-based_train_control (visité le 04/12/2024).

[36] Wikipedia contributors. *European Rail Traffic Management System (ERTMS)*. 2024. url : https://en.wikipedia.org/wiki/European_Rail_Traffic_Management_System (visité le 04/12/2024).

[37] François Belmonte. “Contributions à l’organisation et à la gestion des systèmes ferroviaires : une approche systémique appliquée à la supervision et au pilotage des trains”. Consulté le 4 décembre 2024. Thèse de doct. École Nationale Supérieure des Mines de Paris, 2008. url : https://theses.hal.science/tel-00359084/file/TheseSPICA-Rail_F._Belmonte_.pdf.

[38] Alan M. Turing. “Computing machinery and intelligence”. In : *Mind* 59.236 (1950), p. 433-460.

[39] John McCarthy et al. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”. In : Dartmouth College. 1955.

[40] Daniel Crevier. *AI : The Tumultuous History of the Search for Artificial Intelligence*. Basic Books, 1993.

[41] Stuart J. Russell et Peter Norvig. *Artificial Intelligence : A Modern Approach*. 3rd. Pearson Education Limited, 2016.

[42] Yann LeCun, Yoshua Bengio et Geoffrey Hinton. “Deep learning”. In : *Nature* 521.7553 (2015), p. 436-444.

[43] Laurent Gimazane. *Les différents algorithmes de l’IA*. Présentation PowerPoint. Disponible sur https://knowmap.org. Le document explique les concepts fondamentaux des algorithmes d’intelligence artificielle, en mettant en avant les distinctions entre apprentissage machine et apprentissage profond. 2024.

[44] Ian Goodfellow, Yoshua Bengio et Aaron Courville. *Deep Learning*. MIT Press, 2016.

[45] Andre Esteva et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In : *Nature* 542.7639 (2017), p. 115-118.

[46] Xu Bao et al. “Prediction of Train Arrival Delay Using Hybrid ELM-PSO Approach”. In : *Journal of Advanced Transportation* 2021 (2021), Article ID 7763126. doi : 10.1155/2021/7763126.

[47] Brent D. Mittelstadt et al. “The ethics of algorithms : Mapping the debate”. In : *Big Data & Society* 3.2 (2016), p. 1-21.

[48] Romain Couillet. *Machine Learning - Ensimag 3A IF : Principes de l'apprentissage statistique supervisé*. Document pédagogique décrivant les principes de l'apprentissage statistique supervisé, incluant des explications détaillées sur les SVM, les réseaux neuronaux et les méthodes de minimisation des risques. 2024.

[49] Richard S. Sutton et Andrew G. Barto. *Reinforcement Learning : An Introduction*. 2nd. MIT Press, 2018.

[50] Simon J.D. Prince. *Understanding Deep Learning*. 1st. Accessible via `http://udlbook.com`. Sous licence Creative Commons CC-BY-NC-ND. Cambridge, MA : The MIT Press, 2023. isbn : 9780262547994.

[51] Sara Akodad et al. "Modèle d'ensemble d'apprentissage profond basé sur la représentation au second ordre de descripteurs multi-couches d'un CNN". In : *Université de Bordeaux, CNRS, IMS, UMR 5218, Groupe Signal et Image*. F-33405 Talence, France, 2020.

[52] D. Camarazo et al. "Modélisation des systèmes ferroviaires - objectifs, approches et problématiques". In : *eduBIM 2023*. Présentation des objectifs, approches et problématiques des systèmes ferroviaires. Dijon, France, 2023. url : `https://hal.science/hal-04353184/document`.

[53] Stéphane Crozat. *Modélisation avancée en UML et en relationnel*. Concepts avancés de modélisation UML, notamment les compositions, agrégations et leur transformation en modèles relationnels. 2018. url : `http://stph.scenari-community.org` (visité le 10/12/2024).

[54] Grégory Claude. *UML et les Bases de Données*. Introduction à la transformation des diagrammes UML en modèles relationnels, avec un focus sur la normalisation et les clés primaires. 2010-2011.

[55] Wikipédia. *Capella (ingénierie)*. 2024. url : `https://fr.wikipedia.org/wiki/Capella_(ing%C3%A9nierie)`.

[56] Afafe Zehrouni. "Planification intégrée et robuste des ressources ferroviaires". Thèse co-dirigée par l'École des Mines de Saint-Étienne et la SNCF. Thèse de doct. Université de Lyon, 2016. url : `https://theses.hal.science/tel-03270916`.

[57] Philippe Rigaux. *Le Modèle Relationnel : Cours de Bases de Données*. Cours décrivant le modèle relationnel, les schémas relationnels, et leur implémentation en SQL. Université Paris-Sud. France, 2003.

[58] Mohamed Nemiche. *Modèle Relationnel*. Cours détaillant les bases du modèle relationnel, les notions fondamentales, la normalisation et les exercices pratiques. Faculté Polydisciplinaire d'Ouarzazate, Filière Informatique et Gestion d'Entreprise. 2012/2013.

[59] Florian Toqué. "Prévision et visualisation de l'affluence dans les transports en commun à l'aide de méthodes d'apprentissage automatique". Thèse de doctorat en informatique, IFSTTAR et Polytechnique Montréal. Thèse de doct. Université Paris-Est, 2019.

[60] Rémi Coulaud. "Modélisation et prévision de l'affluence et du temps d'échange à l'interface quai/train : application à la ligne H". Thèse co-dirigée par la SNCF et l'Université Paris-Sud. Thèse de doct. Université Paris-Sud, 2020.

[61] Nikola Markovic et al. “Analyzing passenger train arrival delays with support vector regression”. In : *Transportation Research Part C : Emerging Technologies* 56 (2015), p. 251-262. doi : 10.1016/j.trc.2015.04.004.

[62] P. Kecman et R. M. P. Goverde. “Adaptive real-time prediction of train events based on historical data”. In : *Transportation Research Part C : Emerging Technologies* 54 (2015), p. 60-85. doi : 10.1016/j.trc.2015.02.009.

[63] Trong-Ton Pham et al. “Modèle de graphe et modèle de langue pour la reconnaissance de scènes visuelles”. In : *RSTI - Document Numérique* 13.1 (2010), p. 211-227. doi : 10.3166/DN.13.1.211-227.

[64] Pierre Baron, Laurent Roche et Thibault Raffaillac. “Visualisation automatique de graphes pour l’apprentissage des algorithmes”. In : *10e Conférence sur les Environnements Informatiques pour l’Apprentissage Humain (EIAH 2021)*. Fribourg, Suisse : HAL, 2021, p. 373-376. url : https://hal.science/hal-03290328v1.

[65] Masoud Yaghini, Mohammad M. Khoshraftar et Masoud Seyedabadi. “Predicting Passenger Train Delays Using Neural Network”. In : *Iran University of Science and Technology, School of Railway Engineering* (2012).

[66] Roheen Qamar et Baqar Ali Zardari. “Artificial Neural Networks : An Overview”. In : *Mesopotamian Journal of Computer Science* 2023 (2023). Review article discussing the fundamentals of Artificial Neural Networks (ANNs), their architecture, types, and applications in machine learning and deep learning., p. 130-139. doi : 10.58496/MJCSC/2023/015. url : https://www.researchgate.net/publication/373700317.

[67] Chao Wen Weiwei Mou Zhaolan Cheng. “Predictive Model of Train Delays in a Railway System”. In : *8th International Conference on Railway Operations Modelling and Analysis - RailNorrköping 2019* (2019), p. 913-929.

[68] Sepp Hochreiter et Jürgen Schmidhuber. “Long Short-Term Memory”. In : *Neural Computation* 9.8 (1997), p. 1735-1780. doi : 10.1162/neco.1997.9.8.1735. url : https://www.researchgate.net/publication/13853244_Long_Short-Term_Memory.

[69] Benjamin Lindemann et al. “A Survey on Long Short-Term Memory Networks for Time Series Prediction”. In : *Procedia CIRP* 99 (2021), p. 650-655. doi : 10.1016/j.procir.2021.03.088. url : https://www.sciencedirect.com/science/article/pii/S2212827121000892.

[70] Zhibin Jiang et al. “Evaluating rail transit timetable using big passengers’ data”. In : *Journal of Computer and System Sciences* 82.1 (2016), p. 144-155. doi : 10.1016/j.jcss.2015.08.004.

[71] Martin Spielauer. *Qu’est-ce qu’une microsimulation dynamique en sciences sociales ?* Ce document offre une introduction détaillée à la microsimulation dynamique, ses principes, ses avantages, et ses applications dans le domaine des sciences sociales. Statistique Canada – Division de la modélisation. Ottawa, Canada, 2013. url : https://www.statcan.gc.ca/.

[72] Annabell Berger et al. “Stochastic Delay Prediction in Large Train Networks”. In : *11th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS’11)*. T. 20. OpenAccess Series in Informatics (OASIcs). Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2011, p. 100-111. doi : 10.4230/OASIcs.ATMOS.2011.100.

[73] H. Rouanet. “Les modèles stochastiques d’apprentissage”. In : *Mathématiques et sciences humaines* 5 (1964). Introduction aux modèles stochastiques dans le cadre de la psychologie behavioriste et des expériences d’apprentissage., p. 3-10. url : http://www.numdam.org/item?id=MSH_1964__5__3_0.

[74] Chunjiao Dong et al. “Flow rate and time mean speed predictions for the urban freeway network using state space models”. In : *Transportation Research Part C : Emerging Technologies* (2014). doi : 10.1016/j.trc.2014.02.014.

[75] Guillaume Costeseque. *Prévision court terme des temps de parcours : Synthèse bibliographique des différentes méthodes et expérimentations*. Rapp. tech. Rapport d’étude avec la participation de Nicolas Ditchi et Wilhémine Lecointre, contrôlé par Marie-Amélie Horvath et validé par Anne Gregoire. Nantes, France : Cerema - Centre d’Etudes et d’Expertise sur les Risques, l’Environnement, la Mobilité et l’Aménagement, 2019. url : https://hal.science/hal-03288122v1.

[76] Matthieu Lemoine et Florian Pelgrin. “Introduction aux modèles espace-état et au filtre de Kalman”. In : *Revue de l’OFCE* 86 (2003). Ce document présente les concepts fondamentaux des modèles espace-état, leur estimation par le filtre de Kalman et l’algorithme EM, ainsi que leurs applications à l’analyse économique, telles que la décomposition tendance-cycle et l’évaluation des indices coïncidents., p. 204-225. url : http://www.numdam.org/item?id=MSH_1964__5__3_0.

[77] Joseph Abdo. *Construire de nouvelles infrastructures routières : La solution à la congestion du trafic ?* Ce document discute des avantages et inconvénients de l’expansion des infrastructures routières pour réduire la congestion et ses impacts économiques, sociétaux, et environnementaux. Cimbéton. France, 2013. url : https://www.cimbeton.fr/.

[78] Michael F. Gorman. “Statistical estimation of railroad congestion delay”. In : *Transportation Research Part E : Logistics and Transportation Review* 45.4 (2009), p. 446-456. doi : 10.1016/j.tre.2008.08.004.

[79] Francesco Corman et Lingyun Meng. “A Review of Online Dynamic Models and Algorithms for Railway Traffic Management”. In : *IEEE Transactions on Intelligent Transportation Systems* 16.3 (2015), p. 1274-1284. doi : 10.1109/TITS.2014.2358392.

[80] Holger Flier et al. “Mining Railway Delay Dependencies in Large-Scale Real-World Delay Data”. In : *Robust and Online Large-Scale Optimization*. Springer, 2009, p. 354-368. doi : 10.1007/978-3-642-05465-5_18.

[81] Rui Shi et al. “Prediction and analysis of train arrival delay based on XGBoost and Bayesian optimization”. In : *Applied Soft Computing* 109 (2021), p. 107538.

[82] NVIDIA. *What is XGBoost ?* 2024. url : https://www.nvidia.com/en-us/glossary/xgboost/.

[83] Rémy Priem. “Optimisation bayésienne sous contraintes et en grande dimension appliquée à la conception avion avant projet”. Thèse de doctorat dirigée par Nathalie Bartoli et Youssef Diouane. Thèse de doct. Toulouse, France : Institut Supérieur de l’Aéronautique et de l’Espace (ISAE-SUPAERO), 2020. url : https://www.isae-supaero.fr/.

[84] Tianfeng Chai et R. R. Draxler. “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature”. In : *Geoscientific Model Development* 7 (2014), p. 1247-1250. doi : 10.5194/gmd-7-1247-2014. url : https://www.geosci-model-dev.net/7/1247/2014/.

[85] Shifei Ding, Xinzheng Xu et Ru Nie. “Extreme learning machine and its applications”. In : *Neural Computing and Applications* 25.3–4 (2014), p. 549-556. doi : 10.1007/s00521-013-1522-8. url : https://www.researchgate.net/publication/259633933.

[86] Allal El Moubarek Bouzid. *Optimisation par la méthode des essaims particulaires d’une fonction trigonométrique*. Ce document présente une introduction à l’optimisation par essaim particulaire (PSO), ses principes fondamentaux, ses variantes, et inclut des exemples d’application sur des fonctions mathématiques. École Normale Supérieure d’Enseignement Technique d’Oran. 2010. url : https://www.researchgate.net/publication/331354576.

[87] T. Cover et P. Hart. *Nearest Neighbor Pattern Classification*. T. 13. 1. IEEE, 1967, p. 21-27. doi : 10.1109/TIT.1967.1053964.

[88] Anna Veronika Dorogush, Vasily Ershov et Andrey Gulin. “CatBoost : gradient boosting with categorical features support”. In : *arXiv preprint* arXiv :1810.11363 (2018). url : https://arxiv.org/abs/1810.11363.

[89] Jerome H. Friedman. “Greedy Function Approximation : A Gradient Boosting Machine”. In : *The Annals of Statistics* 29.5 (2001), p. 1189-1232. doi : 10.1214/aos/1013203451.

[90] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 58.1 (1996), p. 267-288. url : https://www.jstor.org/stable/2346178.

[91] Chao Wen, Zhaolan Cheng et Weiwei Mou. “Predictive model of train delays in a railway system”. In : *8th International Conference on Railway Operations Modelling and Analysis* (2019), p. 913-926.

[92] William Marsh et George Bearfield. “Causal Modelling of Lower Consequence Rail Safety Incidents”. In : *RSSB* (2010). Presented at Rail Safety Conference, 2010.

[93] Judea Pearl. “Causal inference in statistics : An overview”. In : *Statistics Surveys* 3 (2009), p. 96-146. doi : 10.1214/09-SS057. url : https://projecteuclid.org/euclid.ssu/1268143831.

[94] Léon Sobrie et al. “Capturing complexity over space and time via deep learning : An application to real-time delay prediction in railways”. In : *European Journal of Operational Research* 310 (2023), p. 1201-1217. doi : 10.1016/j.ejor.2023.03.040.

[95] Ke Yu et al. “Delay prediction with spatial–temporal bi-directional LSTM in railway network”. In : *ICT Express* 9 (2023), p. 921-926. doi : 10.1016/j.icte.2023.05.002.

[96] S. CHTIOUI et al. “XGBoost in Public Transportation for Multi-Attribute Data : Delay Prediction in Railway Systems in Real-Time”. In : *IEEE Access* (2023). doi : 10.1109/ACCESS.2023.0322000.

[97] Cort J. Willmott et Kenji Matsuura. “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance”. In : *Climate Research* 30.1 (2005), p. 79-82. doi : 10.3354/cr030079. url : https://www.int-res.com/articles/cr2005/30/c030p079.pdf.

[98] Sylvain Arlot et Alain Celisse. “A survey of cross-validation procedures for model selection”. In : *Statistics Surveys* 4 (2010), p. 40-79. doi : 10.1214/09-SS054. url : https://projecteuclid.org/euclid.ssu/1268143839.

[99] Trevor Hastie, Robert Tibshirani et Jerome Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. 2nd edition. Springer, 2009. doi : 10.1007/978-0-387-84858-7.

[100] Jerome H. Friedman. “Greedy Function Approximation : A Gradient Boosting Machine”. In : *The Annals of Statistics* 29.5 (2001), p. 1189-1232. doi : 10.1214/aos/1013203451.

[101] George A. F. Seber et Alan J. Lee. *Linear Regression Analysis*. 2nd edition. Wiley, 2003. doi : 10.1002/0471722199.

[102] Andrew Gelman et al. *Bayesian Data Analysis*. 3rd edition. CRC Press, 2013. doi : 10.1201/b16018.

[103] Hugo Larochelle, Yoshua Bengio et Aaron Courville. “Decision Trees : From Basics to Modern Interpretations”. In : *Machine Learning Advances* 37 (2021), p. 45-67. doi : 10.1016/j.mla.2021.04.001.

[104] Jianxin Zhang, Yanmei Lu et Jianping Wu. “Random Forest : Theory and Application”. In : *Frontiers in Environmental Science* 9 (2021), Article 653034. doi : 10.3389/fenvs.2021.653034.

[105] Mike Schuster et Kuldip K. Paliwal. “Bidirectional Recurrent Neural Networks”. In : *IEEE Transactions on Signal Processing* 45.11 (1997), p. 2673-2681. doi : 10.1109/78.650093.

[106] James Bergstra et Yoshua Bengio. “Random Search for Hyper-Parameter Optimization”. In : *Journal of Machine Learning Research* 13 (2012), p. 281-305. url : http://jmlr.csail.mit.edu/papers/volume13/bergstra12a/bergstra12a.pdf.

[107] Wikipédia contributors. *Web scraping*. https://fr.wikipedia.org/wiki/Web_scraping. 2024.

[108] Jamell Ivor Samuels. “One-Hot Encoding and Two-Hot Encoding : An Introduction”. In : *Preprint* (jan. 2024). doi : 10.13140/RG.2.2.21459.76327. url : https://www.researchgate.net/publication/377159812_One-Hot_Encoding_and_Two-Hot_Encoding_An_Introduction.

[109] Dr. Messaoud Zerouti. *Statistique et Analyse de Données*. Polycopie de cours. Ecole Nationale Supérieure de Management (ENSM), Pôle universitaire de Koléâ, Tipaza, Année Universitaire 2018/2019. 2018. url : https://ensm.dz/wp-content/uploads/2023/06/Polycopie-ZEROUTI-Statistique-et-analyse-de-donnees.pdf.

[110] Alain Baccini et al. “Analyse statistique des données d’expression”. In : *Journal de la Société Française de Statistique* 146.4 (2005). To appear in a special issue on gene expression data, p. 345-360.

[111] Grégory Boissinot, Bertrand Lehurt et David Wursteisen. *Web API : Edition 2017*. Un livre blanc détaillant les concepts des Web APIs, y compris leur conception, leur architecture, et leurs applications. Paris, France : SOAT, 2017.

[112] Arnaud Brossard, Mourad Abed et Christophe Kolski. “Modélisation conceptuelle des IHM : Une approche globale s’appuyant sur les processus métier”. In : *Revue des Sciences et Technologies de l’Information - Série ISI : Ingénierie des Systèmes d’Information* 12.5 (2007). Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License, p. 69-108. doi : 10.3166/isi.12.5.69-108. url : https://uphf.hal.science/hal-03390656v1.

[113] Krishanu Dey. “Flask from Scratch : A Learning Platform for the Flask Web Framework”. A dissertation exploring a browser-based learning platform to teach Flask for web application development. Mém. de mast. School of Computer Science et Statistics, Trinity College Dublin, avr. 2022. url : https://github.com/krishanu-dey/pytch-releases.

[114] Michael Owens. *The Definitive Guide to SQLite*. A comprehensive guide on SQLite, detailing its architecture, usage, and integration with various programming languages. New York, NY, USA : Apress, 2006. isbn : 978-1-59059-673-9.

[115] Venkateshwaran Loganathan. *PySide GUI Application Development*. A comprehensive guide on developing dynamic and robust GUI applications using the PySide framework. Birmingham, UK : Packt Publishing, 2013. isbn : 978-1-84969-959-4.

[116] Jacques Malenfant. *Une étude sémantique du langage QML*. Rapp. tech. RR-4497. Available at HAL open archive : https://inria.hal.science/inria-00072091v1. IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex, France : INRIA, 2002.

Table des figures

2.1 Schéma général de l'intégration d'un SAPD dans un PCC. 24
2.2 Processus global d'utilisation du SAPD dans un PCC. 29
2.3 Classification des techniques d'apprentissage automatique et profond, incluant les approches supervisées, non supervisées et par renforcement 29
2.4 Comparaison des performances des modèles pour différents articles en termes de métriques RMSE, MAE et R^2 ([72], [65], [61], [67], [46], [81], [94], [95]). 37
2.5 Classification des méthodes d'apprentissage issues des contributions des travaux étudiés. 43
2.6 Schéma d'un arbre de décision dont l'objectif est de prévoir le retard . . 45
2.7 Répartition des contributions des articles sur les problématiques ferroviaire. 51

3.1 Conception de l'écran prototype destiné à être intégré dans les Postes de Commande Centralisée (PCC), illustrant l'affichage des retards en temps réel et les interactions possibles pour la gestion des incidents 57
3.2 Méthodologie pour l'intégration de l'IA dans les postes de contrôle ferroviaire, de l'analyse des données à l'évaluation des modèles de prédiction des retards. 59
3.3 Interaction des heures de départ des rames : relation entre les départs passés, présents, et futurs 61
3.4 Cartographie des ressources utilisées : synergie entre l'environnement académique (Estaca) et industriel (Egis-rail) au cœur de la thèse 62
3.5 Illustration du processus de modélisation des données relationnelles . . 64
3.6 Diagramme de classes final pour les données opérationnelles et de flux de passagers 66
3.7 Étapes de prétraitement des données pour l'analyse et l'apprentissage automatique 71
3.8 Algorithme de nettoyage des données pour les formats de date et les identifiants 72
3.9 Duplication des retards de la même rame par rapport à l'heure de départ prévue : comparaison des courbes bleues et rouges. 73
3.10 Algorithme de correction des incohérences temporelles affectant les dates de départ des rames 75
3.11 Incohérence des heures de départ : comparaison des heures de départ prévues et réelles pour chaque station. 77
3.12 Évolution des heures de départ prévues et réelles par station : après correction des incohérences 77
3.13 Example d'une distribution des retards observés sur trois périodes annuelles : 2017, 2018 et 2019 79

3.14 Processus de validation et correction des valeurs aberrantes dans les retards des rames 79
3.15 Processus d'enrichissement de la base de données avec des variables calculées et du Web Scraping pour l'analyse des retards ferroviaires . . . 81
3.16 Processus de la transformation des données d'exploitation quantitatives et qualitatives pour l'apprentissage automatique 83
3.17 Classification des types et méthodes d'analyse des données. 84
3.18 Processus d'entraînement des modèles d'apprentissage pour une analyse comparative 91

4.1 Packages et extensions utilisés pour les modèles d'apprentissage automatique 99
4.2 Évolution de la collecte des jeux de données pour le tramway et le RER au cours du projet de recherche 100
4.3 Schéma abstrait d'une ligne ferroviaire avec confidentialité des données 101
4.4 Structure des données brutes : RER vs Tramway 102
4.5 Évolution des retards selon les quartiles (Q1, Q2, Q3) au cours du temps pour infrastructure RER 105
4.6 Évolution des retards selon les quartiles (Q1, Q2, Q3) au cours du temps pour infrastructure Tramway 105
4.7 Visualisation des retards avec détection des valeurs aberrantes par isolation forest pour infrastructure tramway 106
4.8 Visualisation des retards avec détection des valeurs aberrantes par isolation forest pour infrastructure RER 106
4.9 Distribution des retards par jour de la semaine sur la ligne Tramway avec boîte à moustaches 109
4.10 Distribution des retards par jour de la semaine sur la ligne RER avec boîte à moustaches 109
4.11 Moyenne des retards quotidiens et identification des retards importants dans les données du tramway. 111
4.12 Moyenne des retards quotidiens et identification des retards importants dans les données du RER. 111
4.13 Évolution des retards moyens par heure et identification des retards importants. 111
4.14 Moyenne des retards par station de Tramway avec marge d'erreur de ±150 112
4.15 Moyenne des retards par station de RER avec marge d'erreur de ±150 . . 112
4.16 Corrélation entre la variable expliquée "retard" et les variables explicatives représentant les données d'exploitation "XOP" pour le tramway . . 113
4.17 Corrélation entre la variable expliquée "retard" et les variables explicatives représentant les données d'exploitation "XOP" pour le RER 113
4.18 Corrélation entre la variable expliquée "retard" et les 80 retards précedents pour le tramway 114
4.19 Corrélation entre la variable expliquée "retard" et les 100 retards précedents pour le pour le RER 114
4.20 Classification des jours en fonction de la corrélation entre retards historiques et anticipés pour le Tramway 115
4.21 Classification des jours en fonction de la corrélation entre retards historiques et anticipés pour le RER 115

4.22 Courbes de trajets pour différentes rames, avec un trajet incomplet visible pour la rame 28. 115
4.23 Répartition des données en entraînement, test et catégorisation (quantitative et qualitative) 117
4.24 Comparaison des temps d'entraînement des différents modèles 125
4.25 Comparaison des performances de différents modèles de prédiction des retards en fonction de l'heure de la journée 126
4.26 Comparaison des performances de différents modèles de prédiction des retards en fonction du jour 127
4.27 Comparaison des performances de différents modèles de prédiction des retards en fonction de la station 128
4.28 Processus d'évaluation de la robustesse et de la sensibilité du modèle. . 133

5.1 Modèle statique du prototype IHM 137
5.2 Maquette interactive du système ferroviaire 138
5.3 Diagramme de cas d'utilisation pour la simulation et l'affichage des prédictions dans l'IHM 140
5.4 Architecture globale de IHM : intégration du front-end, back-end, base de données et Xgboost.pkl pour la prédiction des retards 141
5.5 Diagramme de séquence illustrant le flux global de l'application, depuis le lancement de la simulation jusqu'à sa fermeture : la création de la ligne, la récupération et le traitement des données, la prédiction des retards, l'animation des rames. 143
5.6 Diagramme de séquence pour la création de la ligne à simuler : la réorganisation des stations, la génération des inter-stations et la visualisation des aiguilles. 144
5.7 Diagramme de séquence pour l'initialisation des caractéristiques de la simulation. 146
5.8 Diagramme de séquence du processus de prédiction des retards 147
5.9 Diagramme de séquence pour l'animation des rames avec un code couleur.148
5.10 Diagramme de séquence pour l'affichage des prédictions détaillés pour chaque rame. 149
5.11 Interface de configuration de la simulation avec des erreurs de saisie (format incorrect pour l'heure) 150
5.12 Interface de configuration de la simulation avec des valeurs correctement saisies 150
5.13 Vue de la simulation en cours, illustrant les prédictions en temps réel . . 151
5.14 Étiquette de prédiction affichée sur une rame retardée, indiquant les informations sur la station estimée et les détails du retard. 151
5.15 Interface de téléchargement des données simulées 151
5.16 Données générées par la simulation, prêtes à être analysées 152
5.17 Confirmation du téléchargement des données simulées 152
5.18 Capture d'écran du message écrit à ChatGPT pour déterminer les plages horaires à analyser 153
5.19 Capture d'écran de la réponse de ChatGPT pour la sélection des plages horaires 154
5.20 Visualisation des prédictions dans l'intervalle d'erreur autorisé pour le PCC[-150 :150] 155

5.21 Analyse des réponses des opérateurs sur l'évaluation de l'application de simulation et de prédiction des retards 157

Liste des tableaux

2.1 Synthèse des travaux sur l'intelligence artificielle appliquée aux transports publics . . . 39
2.2 Exploration des hyperparamètres par modèle de machine learning . . . 47
2.3 Exploration des hyperparamètres par modèle d'apprentissage profond . 49
2.4 Comparaison des travaux liés à la problématique des retards dans les réseaux ferroviaires. . . . 52

3.1 Exemple de données brutes pour n stations et m trains obtenues à partir du centre de contrôle avant l'utilisation du modèle relationnel expliqué en haut de cette section. . . . 68
3.2 Exemple de données pour la station s_3 et m trains après projection des données brutes présentées dans le tableau 3.1 sur le modèle relationnel. 68
3.3 Exemple de données pour la rame numéro 2 dans m stations après projection des données brutes présentées dans le tableau 3.1 sur le modèle relationnel. . . . 68
3.4 Exemple de données pour m stations et n rames après projection des données brutes présentées dans le tableau 3.1 sur le modèle relationnel. 69
3.5 Exemple de duplication des heures de départ réelles pour une même rame à différentes stations. . . . 74
3.6 Exemple de duplication des départs réels pour une même rame à la première station. . . . 74
3.7 Exemple de données d'exploitation avec incohérences temporelles (passage de 23h à minuit) . . . 76
3.8 Même exemple de le tableau 3.7 après correction des incohérences temporelles (passage de 23h à minuit) . . . 77
3.9 Exploration des caractéristiques pour la prédiction de retard y^{+f} . . . 90

4.1 Nombre de lignes impactées par la phase de prétraitement . . . 103
4.2 Statistiques des retards et valeurs aberrantes pour Tramway et RER . . . 108
4.3 Les taux statistiques pour Tramway et RER. . . . 114
4.4 Exploration des modèles de machine learning et paramètres associés . . 118
4.5 Exploration des hyperparamètres - À base des arbres . . . 120
4.6 Exploration des hyperparamètres - Apprentissage profond . . . 121
4.7 Métriques de performance des modèles sur l'ensemble d'entraînement . 122
4.8 Métriques de performance des modèles sur l'ensemble de test . . . 123
4.9 Résultats du modèle XGBOOST pour les données RER . . . 129
4.10 Comparaison des performances des modèles pour RER et Tramway . . . 131
4.11 Performances du modèle XGBOOST en fonction de la fenêtre temporelle capturant les retards précédents pour le RER . . . 131

4.12 Performances du modèle XGBOOST en fonction de l'horizon temporel pour le RER 132

5.1 Les résultats avec plage horaire, RMSE et P (%) 155

Liste des mots-clés

PCC Poste Contrôle Centralisé
SAPD Systèmes d'Aide à la Prise de Décision
RER Réseau Express Régional
IHM Interface Homme-Machine
CTC Contrôle de Trafic Centralisé
CCT Contrôle Centralisé de Trafic
ATS Automatic Train Supervision
ATP Automatic Train Protection
ATC Automatic Train Control
ATO Automatic Train Operation
CBTC Communication Based Train Control
ERTMS European Rail Traffic Management System
TTM Timetable Management
TTP Timetable Prévu
TTR Timetable Réel
CSS Conditions de Trafic Congestionné
NSS Conditions de Trafic Non congestionné
SMIS Safety Management Information System
DSS Decision Support Systems
GDSS Group Decision Support Systems
ODSS Organizational Decision Support Systems
OLAP Online Analytical Processing
IA Intelligence Artificielle
CNN Réseaux Neuronaux Convolutionnels
RNN Réseaux Récurrents
ML Machine Learning
SVM Machines à Vecteurs de Support
XGBOOST eXtreme Gradient Boosting
LightGBM Light Gradient Boosting Machine
ACP Analyse en Composantes Principales

LSTM Long Short-Term Memory

UML Langage de Modélisation Unifié

RMSE Erreur Quadratique Moyenne Racin

MAE Erreur Absolue Moyenne

OB Optimisation Bayésienne

ELM Extreme Learning Machine

PSO Particle Swarm Optimization

BiLSTM Bi-directional Long Short-Term Memory

TI True In (Prédictions Correctes)

FI False In (Prédictions Incorrectes)

CSV Valeurs Séparées par des Virgules

API Application Programming Interface

HTTP Hypertext Transfer Protocol

QML Qt Modeling Language