Développement et implémentation de modèles apprenants pour l'exploitation des grandes gares

Marie Milliet de Faverges

le cnam

École Doctorale Informatique, Télécommunications et Électronique
Centre d'Etudes et de Recherche en Informatique et Communication

THÈSE DE DOCTORAT

présentée par : **Marie MILLIET DE FAVERGES**
soutenue le : **16 octobre 2020**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline : **Sciences et technologies de l'information et de la communication**

Spécialité : **Informatique**

Développement et implémentation de modèles apprenants pour l'exploitation des grandes gares

THÈSE dirigée par

Picouleau Christophe	*Professeur, Conservatoire National des Arts et Métiers*

et co-encadrée par

Russolillo Giorgio	*Maître de Conférence, Conservatoire National des Arts et Métiers*
Merabet Boubekeur	*Docteur, SNCF Réseau*

RAPPORTEURS

Meunier Frédéric	*Professeur, École Nationale des Ponts et Chaussées*
Ghattas Badih	*Professeur, Université d'Aix-Marseille*

PRÉSIDENT DU JURY

Rossi André	*Professeur, Université Paris-Dauphine*

EXAMINATEURS

Latouche Aurélien	*Professeur, Institut Curie et Conservatoire National des Arts et Métiers*
Rodriguez Joaquin	*Professeur, Institut français des sciences et technologies des transports, de l'aménagement et des réseaux*

Remerciements

Je souhaite avant tout remercier chaleureusement mon directeur de thèse et mes encadrants qui ont su guider ces travaux et être présents tout au long de ces trois années. Mes premiers remerciements vont à mon directeur de thèse Christophe Picouleau pour ses nombreux conseils et son encadrement. C'était un plaisir de travailler avec lui. J'ai une profonde gratitude pour Giorgio Russolillo, qui s'est beaucoup investi et a su me stimuler tout au long de ma thèse. J'ai beaucoup appris et évolué grâce à lui. Je remercie également Boubekeur Merabet pour sa confiance et ses encouragements. Je lui suis particulièrement reconnaissante de s'être rendu aussi disponible sur la fin.

Je remercie Frédéric Meunier et Badih Ghattas d'avoir rapporté cette thèse, ainsi qu'Aurelien Latouche, Joaquin Rodriguez et André Rossi qui ont accepté de faire partie du jury pour ma soutenance. Leur lecture et leur retour sont un précieux regard extérieur sur mon travail.

Je remercie également les membres des équipes OC et MSDMA du CEDRIC pour leur accueil.

Merci à toute l'équipe de la DGEX Solutions au sein de laquelle cela a été un plaisir de travailler et d'apprendre au quotidien. Je pense en particulier à Bertrand Houzel sans qui cette thèse n'aurait pas eu lieu, merci pour sa confiance, son engagement et les moyens qu'il a mis à ma disposition. Je remercie également tous ceux qui ont participé de près ou de loin à cette thèse par leurs conseils et leurs expertises, en particulier Rémi Parel qui m'a aidée avec beaucoup de patience et de gentillesse, ainsi qu'Antoine Robin et Alexandre Mani qui ont travaillé avec moi lors de leur stage. Merci à Franck grâce à qui l'expérience du doctorat à la DGEX était moins solitaire.

Je ne peux citer toutes les personnes qui ont contribué à l'atmosphère incroyable qui règne dans cette cellule, mais je souhaite tout de même adresser un remerciement spécial à Mélanie, Clément, Mingzhi, Lucile, Jean-Patrick, Yvonnig, Sophia, Rudy, Marion et tous les autres que j'ai rencontrés ces dernières années. J'ai noué des amitiés et des relations très riches lors de mon passage chez SNCF Réseau.

Enfin je souhaite remercier ma famille et mes amis qui m'ont si bien entourée. Ils ont vu ma disponibilité et mon attention s'amenuiser sur la fin et j'espère qu'ils ne m'en tiendront pas rigueur ! J'ai une pensée particulière pour mes amis en doctorat. Le partage de nos expériences respectives était toujours réconfortant, je leur souhaite beaucoup de courage pour la fin !

Je saisis également cette occasion pour remercier encore une fois mes parents qui m'ont toujours encouragée, conseillée et soutenue dans mes différents projets, dont celui de faire une thèse. Je pense aussi à mes soeurs Aliénor, Constance, Héloïse et Victoire, ainsi que Aurore, Eduardo et maintenant Vasco, qui sont toujours là quand il le faut !

Je suis enfin très reconnaissante envers Mayeul qui a vécu cette thèse très personnellement et sur qui j'ai pu beaucoup me reposer. Ces trois années ne se seraient pas déroulées aussi bien sans ses encouragements et sa compréhension.

J'espère que chacun et chacune mesure l'impact très personnel et précieux qu'il/elle a eu dans mon parcours, je n'aurais pas pu être mieux entourée !

Résumé

Les acteurs ferroviaires sont aujourd'hui confrontés à d'importantes difficultés liées à la hausse de la demande en transport et à la congestion du réseau qui en découle. Les ressources sont plus difficiles à planifier, et leur forte utilisation augmente le risque de propagation des retards. Cela nuit à la qualité de service et met les agents en opérationnel sous tension.

Cette thèse cherche à exploiter les archives de retards de trains afin d'anticiper de tels conflits de ressources et de produire des planifications plus robustes. Elle traite en particulier du problème d'occupations des voies en gare, qui consiste à affecter à chaque train planifié une voie à quai et un itinéraire à travers la zone de gare pour s'y rendre. Les gares sont des points critiques du réseau, et les retards s'y propagent plus qu'ailleurs en raison des nombreuses ressources partagées. La robustesse des planifications y est cruciale. La stratégie proposée dans ces travaux consiste à utiliser un algorithme de routage de trains en gare qui exploite des prédictions sur les retards acquises grâce à des méthodes de modélisation et d'apprentissage statistique.

Ces méthodes visent à estimer pour chaque train sa distribution de probabilité de retards conditionnellement à des variables explicatives. Celles-ci sont construites à partir de plusieurs facteurs connus pour être corrélés au retard, tels que le contexte temporel, la mission ou la densité du trafic. Deux approches sont testées pour estimer ces probabilités conditionnelles, une paramétrique avec un modèle linéaire généralisé, et une non paramétrique avec une forêt aléatoire. L'étape cruciale de ce module a consisté en l'élaboration d'une méthodologie de sélection et évaluation des modèles proposés. Une méthode de sélection utilisant un score quadratique et une évaluation de la qualité basée sur la calibration et la discrimination des prédictions est recommandée.

Une fois que la fiabilité des distributions prédites a été attestée, elles sont utilisées pour quantifier le risque de conflit de ressources, et donc de propagation de retards. L'approche adoptée utilise un graphe dont les sommets correspondent à des affectations train/itinéraires et dont les arêtes symbolisent la compatibilité entre les affectations. Des poids positifs sont associés à ces arêtes quand elles sont soumises à une incertitude, qui est calculée à partir des probabilités estimées. Des approches de recherche locale sont proposées pour optimiser l'adaptation des graphiques d'occupation des voies en cherchant une clique de poids minimum sur ce graphe. En confrontant les solutions proposées par ces algorithmes avec les retards réels rencontrés, une importante baisse de la fréquence et de l'amplitude des conflits est observée.

Mots-clés : Recherche Opérationnelle, Apprentissage statistique, Opérations ferroviaires, Modélisation des retards

Abstract

Railway actors are currently facing major challenges due to the increase in public transport demand and the resulting network congestion. Resources are more difficult to plan, and their important use increases the delay propagation risk. It affects the quality of service and puts the operational agents under pressure.

This thesis aims to analyze train delay records in order to anticipate resource allocation conflicts and to produce more robust planning. It deals specifically with the train platforming problem, which consists in assigning to each planned train a platform track and a route through the station area to get there. Stations are critical points in the network, and delays are spreading more there than elsewhere due to the many shared resources. Robust planning is crucial. The strategy proposed in this work consists in supplying a train routing and platform allocation algorithm with delay predictions acquired with modeling and statistical learning methods.

These methods aim to estimate for each train its delay probability distribution, conditionally to explanatory variables. These variables are constructed using several factors known to be correlated with delay, such as the time, the mission or the traffic density. Two approaches are used to estimate these conditional probabilities, a parametric one, with a generalized linear model, and a non-parametric one, with a random forest. The central part of this module was the development of a methodology for the selection and evaluation of the proposed models. A selection method using a quadratic score and a quality assessment based on calibration and discrimination of predictions is recommended.

Once the reliability of predicted distributions has been confirmed, they are used to quantify the risk of resource conflicts, and therefore of propagation of delays. The approach adopted uses a graph whose vertices correspond to train / route assignments, and whose edges represent compatibility between the assignments. Positive weights are associated with these edges when they are subject to an uncertainty, which is calculated from the estimated probabilities. Local search approaches are proposed to adapt the platform allocations by looking for a minimum weight click on this graph. A significant drop in the frequency and amplitude of conflicts is observed by comparing the proposed solutions with the real observed delays.

Keywords : Operations Research, Marchine Learning, Railway Operations, Delay Modeling

Table des matières

Liste des tableaux

Table des figures

Acronymes

AFC Automatic Fare Collection
APC Automatic Passenger Counting
AVL Automatic Vehicle Location
CART Classification and Regression Trees
CDF Cumulative Distribution Function
EF Entreprise Ferroviaire
GAMLSS Generalized Additive Models for Location, Scale and Shape
GI Gestionnaire d'Infrastructure
GLM Modèle linéaire généralisé
GOV Graphique d'occupation des voies
HL Test d'Hosmer-Lemeshow
KS Test de Kolmogorov-Smirnov
MAE Mean Absolute Error
MLE Maximum Likelihood Estimate
MSE Mean Squarre error
NBI Distribution Negative Binomiale
PET Probability Estimation Tree
PIT Probability Integral Transform
PLNE Programme Linéaire en Nombres Entiers
PR Point Remarquable
RF Random Forest
RPS Ranked Probability Score
RQR Randomized Quantile Residuals
TN Trains transiliens
TPP Train Platforming Problem
VNS Variable Neighborhood Search

Chapitre 1

Problématique

Ces dernières décennies ont marqué un point de rupture dans l'organisation de la mobilité, avec une croissance forte de la demande en transport, une concentration démographique dans les villes et le développement de nouveaux besoins et usages. Les acteurs de transport doivent s'adapter à ces changements tout en assurant la sécurité et le confort des voyageurs. Le sujet est particulièrement critique dans les villes où l'étalement urbain conduit les usagers à parcourir des distances plus importantes pour les trajets du quotidien et où la mutualisation des moyens de transport est vivement encouragée. Les transports en commun sont donc de plus en plus sollicités.

Dans le cadre du transport ferroviaire, la réponse à cette demande croissante s'accompagne d'une importante congestion du réseau. En effet, les ressources ferroviaires sont fortement contraintes, soit par leur coût de déploiement soit par leur environnement qui empêche ou limite les possibilités d'aménagements. Les acteurs ferroviaires doivent optimiser l'utilisation des différentes ressources humaines et matérielles pour pouvoir assurer l'offre de transport. En cas de congestion, elles sont en effet sollicitées plus que la normale ce qui fragilise la coordination des différents systèmes. Le moindre retard peut compromettre l'enchaînement des opérations et se propager du fait du partage des ressources.

La saturation de l'infrastructure est particulièrement difficile à gérer dans les gares. Les travaux d'expansion sont impossibles en raison de la place centrale qu'elles occupent dans les villes. Une gare dispose ainsi d'un nombre limité de quais sur lesquels réceptionner les trains, et l'accès à ces quais est complexe car les lignes ferroviaires forment un goulot d'étranglement en se rejoignant en avant gare. La densification des circulations requiert des mouvements fréquents dans le périmètre de gare et une réutilisation rapide des voies, ce qui favorise la propagation des retards en cas de perturbations.

Ces travaux se concentreront sur ces problématiques ferroviaires avec comme objectif d'exploiter les historiques de circulations pour quantifier l'incertitude liée aux horaires des trains afin de proposer des adaptations des planifications en gare qui soient plus robustes aux perturbations courantes.

1.1 Contexte industriel

Cette thèse est menée dans le cadre d'une convention CIFRE avec SNCF Réseau. Des rappels sur l'organisation de la production ferroviaire en France sont donnés ici.

1.1.1 Notions préliminaires

Les acteurs : le groupe public ferroviaire est organisé en cinq SA (sociétés anonymes) :

— SNCF Voyageurs est l'entreprise ferroviaire (EF) de SNCF qui gère l'exploitation des trains de voyageurs et de marchandises. D'autres entreprises ferroviaires que la SNCF circulent sur le réseau ferré français.

— SNCF Réseau est le gestionnaire d'infrastructure (GI), dont le rôle est d'entretenir et développer le réseau ferré, et d'en organiser l'utilisation par les différentes entreprises ferroviaires. Le GI prend également en charge la gestion des opérations en temps réel en cas d'aléas.

— Fret SNCF est une filiale de la SNCF en charge du transport fret.

— SNCF Gares & Connexions gère les gares voyageurs du réseau français.

— SNCF est l'organisme de tête qui assure la cohérence du groupe et le pilotage de la stratégie.

Plusieurs autres acteurs interviennent dans la prise de décision, la stratégie et la réglementation, comme les usagers, les pouvoirs publics ou l'Union Européenne.

Organisation de l'infrastructure : l'infrastructure ferroviaire est constituée des installations et équipements permettant la circulation des trains : voies ferrées, caténaires, signalisation, communication, etc. Étant donné que la distance de freinage est largement supérieure à la distance de visibilité, la sécurité des circulations sur l'infrastructure est assurée par un système de cantonnement, c'est-à-dire de découpage des voies en différents cantons. Un seul train est autorisé à la fois sur chaque canton, et la signalisation ferroviaire permet d'assurer le respect des contraintes d'occupation des cantons, ainsi que la gestion des autres risques ferroviaires.

Composition d'un train : pour faire circuler un train, l'entreprise ferroviaire en charge doit rassembler les ressources suivantes :

— un sillon : c'est une autorisation d'utilisation du réseau à un horaire et sur un parcours précis. La demande de sillon est effectuée par l'entreprise ferroviaire auprès du gestionnaire d'infrastructure.

— le matériel roulant : il s'agit d'une ou plusieurs rames dont le type peut varier en fonction de la mission (fret, TGV, TER, etc.). Les rames sont soumises à des contraintes de maintenances et de révision régulières.

— ressources humaines : conducteurs, éventuellement contrôleurs habilités sur le matériel roulant et la ligne, mais également les agents d'escale qui assurent le bon déroulement des opérations en gare.

— les voyageurs ou les marchandises.

Le partage de ces ressources par différents trains favorise la propagation des retards dans le réseau.

1.1.2 La production ferroviaire

Trois phases se distinguent dans la production ferroviaire :

— La conception et adaptation du plan de transport : cette phase débute près de 3 ans avant la mise en production et s'organise de manière séquentielle : les différentes ressources sont planifiées les unes après les autres pour construire le plan de transport. Celui-ci correspond à l'organisation des ressources humaines et matérielles nécessaires pour répondre à la demande de manière sécuritaire.

Les ressources sont en général planifiées dans l'ordre suivant : les horaires des trains sont établis en premier de manière à respecter les dessertes et fréquences imposées en amont via un processus de commande de sillons des entreprises ferroviaires auprès du gestionnaire d'infrastructure, puis le matériel roulant est affecté à chaque circulation, et enfin le personnel nécessaire est mobilisé. Ce

plan de transport est ensuite adapté dans les derniers mois, et ce jusqu'à la veille des opérations à 17h afin d'intégrer des demandes des différents acteurs.

— La phase opérationnelle : c'est la gestion en temps réel de la production ferroviaire. Le gestionnaire d'infrastructure a la responsabilité de l'organisation des circulations, et peut arbitrer en cas de conflits. Il est en relation constante avec les entreprises ferroviaires.

— L'analyse : les relevés d'opérations sont ensuite étudiés, en premier lieu pour remonter les évènements et identifier la responsabilité de chacun en cas d'incident, mais aussi pour mesurer la performance générale ou identifier des leviers d'amélioration potentiels.

Modernisation des processus : la construction du plan de transport repose encore largement sur l'expertise humaine, notamment par réutilisation et adaptation des plans de transports des années précédentes. L'adaptation en temps réel repose également sur l'appréciation et la réactivité des agents dans les centres opérationnels. Une volonté de modernisation des différentes étapes est présente dans le groupe public ferroviaire et plusieurs axes sont explorés. Deux nous intéressent particulièrement ici.

Tout d'abord un effort conséquent a été placé ces dernières années sur la remontée, l'archivage et l'accessibilité des données opérationnelles. Elles contiennent notamment les horaires de chaque évènement sur des points stratégiques du réseau (passage, arrêt, départ), les comptes rendus d'incidents, l'affluence en gare ou à bord des trains, etc. Elles peuvent être utilisées en temps réel, donnant ainsi plus d'informations pour aider les agents dans la gestion opérationnelle ou pour fluidifier l'information voyageur, ou être utilisées hors ligne pour analyser les performances ou adapter les planifications en fonction de ce qui a été observé.

Un autre axe est celui de la conception d'outils d'aide à la décision pour appuyer l'organisation de la production ferroviaire. La combinatoire du système ferroviaire est telle que la vérification et l'optimisation des différentes planifications est très complexe et sollicite de nombreuses ressources humaines. Une partie de ce travail peut être automatisée ou assistée à l'aide de programmes mathématiques adaptés. La recherche opérationnelle s'est en particulier imposée comme un outil efficace pour cette tâche.

Performance et robustesse : les performances des opérations ferroviaires sont mesurées par plusieurs indices qui permettent de quantifier le respect du niveau de service annoncé. Les indicateurs principaux utilisés sont la ponctualité et la régularité des trains en circulation. Étant donné l'impact qu'a la congestion du réseau sur le trafic, la robustesse des services ferroviaires, qui se définit comme "la capacité effective à réaliser les services promis aux clients", a été placée au coeur de la stratégie du groupe public ferroviaire [5]. Plusieurs points d'action ont été recommandés : la robustesse de conception des services, l'efficacité de l'assemblage des ressources, la gestion des restrictions de capacité et la maîtrise des évènements externes.

La robustesse de conception des services concerne notamment les différentes étapes de planification des ressources et du plan de transport : la construction de ces planifications doit intégrer un objectif de protection face aux perturbations afin de contribuer localement à la qualité globale de l'offre ferroviaire.

1.1.3 Occupation des voies en gare

Les gares jouent un rôle unique dans l'organisation du système ferroviaire, tant par l'interface qu'elles forment avec les usagers que par la place centrale qu'elles tiennent dans la coordination des différentes ressources. En particulier, l'infrastructure est l'une des ressources les plus contraintes et son utilisation doit être planifiée en amont.

Un Graphique d'Occupation des Voies, ou GOV, est une planification des affectations de voies en gare pour chaque train en fonction de ses horaires d'arrivée et de départ. Les GOV sont construits

par le gestionnaire d'infrastructure environ un an avant les opérations, en parallèle de la grille horaire afin d'en évaluer la faisabilité et d'estimer la capacité résiduelle. Ils sont ensuite adaptés dans les mois qui suivent jusqu'à la veille pour intégrer les divers changements.

Cette organisation des voies en gare est cependant très sensible aux aléas opérationnels. En particulier, des retards, même de quelques minutes peuvent générer des conflits en raison de la congestion de l'infrastructure. Ces retards se propagent vite, par exemple si un train est maintenu sur sa voie pour laisser passer la circulation retardée. Ces aléas nécessitent alors l'intervention d'agents pour replanifier, soit en modifiant les affectations, soit en propageant les retards. Chaque minute de retard générée représente un coût important pour les différents acteurs et fragilise un peu plus le système puisque ces retards peuvent avoir des répercussions sur la suite du trajet et dans les gares suivantes. Travailler sur la conception robuste des affectations de voies en gare contribue donc pleinement à améliorer la qualité de service au global.

1.2 Présentation du sujet

Ces travaux tenteront de répondre à la question suivante : dans quelle mesure l'analyse des archives de retards observés peut-elle permettre d'améliorer la résistance aux perturbations des planifications d'occupations des voies en gare ?

La stratégie adoptée dans ces recherches consiste à travailler en deux temps, avec en premier lieu l'étude des archives de retards puis l'utilisation des motifs identifiés pour alimenter un algorithme de placement de trains. La méthodologie est dédiée à l'adaptation à moyen terme de ces placements, soit quelques semaines avant les opérations. Cela permet d'avoir à disposition des données récentes ainsi qu'une solution initiale qui sera adaptée pour gagner en résistance aux retards.

Cette problématique transdisciplinaire soulève plusieurs enjeux, à la fois industriels et académiques qui sont détaillés ici.

1.2.1 Objectifs industriels

L'analyse par apprentissage statistique des données de retards est encore récente chez SNCF Réseau, et de manière générale dans la communauté ferroviaire. Ces données servent principalement aujourd'hui à l'organisation en temps réel des circulations, à l'identification des responsabilités de chaque acteur en cas de conflit ou à l'analyse des performances. L'accès à ces données s'est plus largement ouvert au sein du groupe ces dernières années, permettant l'émergence de nouveaux sujets. Actuellement, les sujets d'analyse prédictive des retards à la SNCF sont principalement liés à l'estimation des retards en temps réels, afin d'évaluer les heures d'arrivées aux futurs arrêts connaissant les retards et incidents courants sur le réseau. Cette information n'est pour l'instant pas utilisée pour aider à la mesure ou l'optimisation de la robustesse ferroviaire qui sont encore largement menées de manière déterministe.

Dans ces travaux, on cherche à identifier les profils de risque de retard conditionnellement au contexte de circulation pour les trains commerciaux prévus dans une gare donnée. L'horizon de prédiction étant d'au moins un jour avant les opérations à quelques semaines, le contexte est encore peu connu (incidents en cours, échanges de matériels, routage, etc.), ce qui limite la connaissance qu'on peut extraire des observations. Par ailleurs, la congestion des gares est telle qu'il n'y a pas d'assurance que la connaissance du risque de retard permette effectivement de protéger les circulations. On cherchera à apporter une réponse aux questions suivantes pour SNCF Réseau :

— Y-a-t-il des motifs récurrents identifiables dans les données de retards de train en phase pré-opérationnelle ?

— Cette information est-elle fiable et stable au cours du temps ?

— Une fois des motifs identifiés, quelles sont les pistes pour les intégrer dans le processus d'adaptation des planifications ?

— Cette nouvelle information est-elle passible d'améliorer la robustesse des planifications ?

On accompagnera ces réponses d'analyses et de recommandations concernant les cas d'études, les types de circulations ou hypothèses de modélisation. Des considérations seront également apportées sur la stabilité au cours du temps des prescriptions proposées.

1.2.2 Positionnement scientifique

Ces travaux se penchent sur la question de l'incertitude des données dans le cadre d'une résolution par des méthodes de recherche opérationnelle. Traditionnellement, deux approches permettent de résoudre des problèmes : l'optimisation robuste et l'optimisation stochastique. La première ne connaît pas la distribution des données mais introduit un ensemble de valeurs possibles, et va chercher à produire une solution réalisable malgré des variations des paramètres à l'intérieur de ces ensembles. L'optimisation stochastique nécessite la connaissance de la distribution de probabilité des évènements, et cherche une solution optimale en moyenne. Ces dernières années ont vu l'émergence des approches prescriptives qui visent à intégrer des techniques de différents domaines, comme les statistiques, l'apprentissage ou l'optimisation combinatoire, afin de répondre avec plus d'efficience aux problèmes de décision en tirant parti des masses de données collectées. Un objectif est notamment de compléter les connaissances préalables données par des experts avec des paramétrages déterminés automatiquement à l'aide d'analyses et de prédictions basées sur des données.

Ces travaux s'inscrivent dans ce contexte scientifique, et on propose ici de quantifier l'incertitude autour de la structure d'un problème combinatoire en utilisant des méthodes d'apprentissage statistique appliquées sur les réalisations des aléas dans le passé. La variable cible est étudiée conditionnellement à un ensemble d'autres variables afin d'en extraire des informations sur la distribution sous-jacente des paramètres incertains. Cette méthodologie d'aide à la décision repose ainsi sur la coordination d'un modèle d'apprentissage statistique et d'un modèle de Recherche Opérationnelle. Cet interfaçage impose des contraintes supplémentaires, tant sur le format des prédictions et l'évaluation de leur qualité que sur la complexité de l'intégration des données ou la taille des instances.

Ces recherches soulèvent un enjeu double quant à la validation et l'évaluation des modèles. Le premier pan de la méthodologie repose sur la construction de modèles d'estimation de probabilités conditionnelles : il s'agit donc d'être en mesure de construire de telles probabilités individuelles pour chaque combinaison de variables explicatives, et d'attester ensuite qu'elles apportent une représentation fiable d'un processus dont on ne connaît qu'un nombre fini d'observations passées. L'évaluation de ces probabilités est d'autant plus complexe que le contexte peut être unique.

Le second pan utilise les résultats de ces analyses prescriptives pour proposer des solutions intégrant le risque de variation des valeurs de ses paramètres d'entrée. On se heurte cependant au problème d'évaluation des solutions proposées, que ce soit d'une manière déterministe ou stochastique. Selon le problème étudié, l'impact des aléas n'est pas toujours facile à mesurer, et il ne peut pas toujours être pris en compte dans l'optimisation des solutions, par exemple en raison d'une trop grande combinatoire ou de difficultés de modélisation. Ainsi, même en disposant d'une connaissance de la distribution des paramètres du problème, on n'est pas forcément en mesure de l'exploiter stratégiquement, ni d'attester de manière fiable si une amélioration a été permise par cette information.

1.3 Structure de la thèse

Les travaux de référence du domaine ferroviaire qui ont guidé ces recherches sont exposés dans le chapitre 2. Ils s'organisent selon plusieurs axes : l'utilisation de Recherche Opérationnelle pour assister

l'organisation de l'offre ferroviaire, la mesure et l'optimisation de la robustesse aux retards, le potentiel des données ferroviaires et enfin une revue détaillée des analyses statistiques de retards de trains.

Le chapitre 3 propose deux méthodes alternatives pour estimer la distribution de probabilité d'une variable cible conditionnellement à une ou plusieurs variables explicatives. Une revue détaillée des approches de sélection et d'évaluation des modèles construits est fournie, ainsi que des recommandations pour choisir l'approche adéquate. Ces techniques sont ensuite appliquées sur le cas réel des données de retards de trains dans le chapitre 4. Les différentes étapes de construction, sélection et évaluation du modèle final sont explicitées, permettant d'identifier des caractéristiques propres aux données de retards de train et de proposer des hypothèses et restrictions de modélisation adaptées.

Ces prédictions sont ensuite intégrées dans un module de recherche locale pour l'adaptation robuste de graphiques d'occupation des voies. Le chapitre 5 recense les différents algorithmes construits dans ce but.

Enfin, le chapitre 6 synthétise les contributions et principaux résultats de ces travaux. Les limites, perspectives et recommandations sur cette méthodologie sont également décrites.

Chapitre 2

Planification et données du système ferroviaire

Ces travaux s'inscrivent dans un cadre d'aide à la décision pour le transport ferroviaire, à la frontière des statistiques, de l'apprentissage automatique et de la recherche opérationnelle. Les premiers articles d'application de la Recherche Opérationnelle aux problèmes ferroviaires ont été publiés il y a plusieurs dizaines d'années, et le domaine ne cesse de s'enrichir et de se structurer depuis. Les sujets liés aux données, et en particulier pour l'apprentissage statistique, sont beaucoup plus récents, et n'ont été rendus possibles que par l'automatisation de nombreux types d'enregistrements et de leur archivage, et par les progrès techniques permettant leur traitement. Ce domaine en pleine expansion représente un enjeu certain pour la compréhension des aléas et l'analyse des performances, mais également pour l'anticipation et l'amélioration en amont de l'organisation du système ferroviaire.

Ce chapitre illustre ces problématiques par une revue de littérature des éléments actuels en planification ferroviaire et des perspectives que peuvent avoir les données pour celle-ci. La première section résume les grands axes de la recherche opérationnelle ferroviaire, avec une introduction rapide des problèmes les plus classiques, ainsi qu'une réflexion sur la qualité des solutions et l'optimisation de la robustesse. Le problème d'occupation des voies en gare est ensuite étudié plus en détail en section 2.2. La section 2.3 énumère les bases de données classiques en transport public et ferroviaire, ainsi que les cas d'usage les plus importants. Enfin, la section 2.4 détaille les travaux dédiés à l'analyse et à la prédiction des retards de trains.

2.1 La Recherche opérationnelle ferroviaire

Le domaine de la Recherche Opérationnelle s'est rapidement imposé comme un outil clé pour assister l'organisation de l'offre ferroviaire en raison de sa forte combinatoire et de sa complexité d'appréhension. Une classification des principaux problèmes d'optimisation ferroviaire est rappelée, puis la notion de qualité d'une solution est discutée. Enfin la question de la robustesse ferroviaire est introduite par le problème des grilles horaires.

2.1.1 Quelques problèmes classiques

La planification ferroviaire est construite de manière séquentielle, c'est-à-dire en organisant les différents sous-systèmes les uns après les autres, avec éventuellement des boucles pour corriger et adapter. En particulier, trois horizons temporels se distinguent :

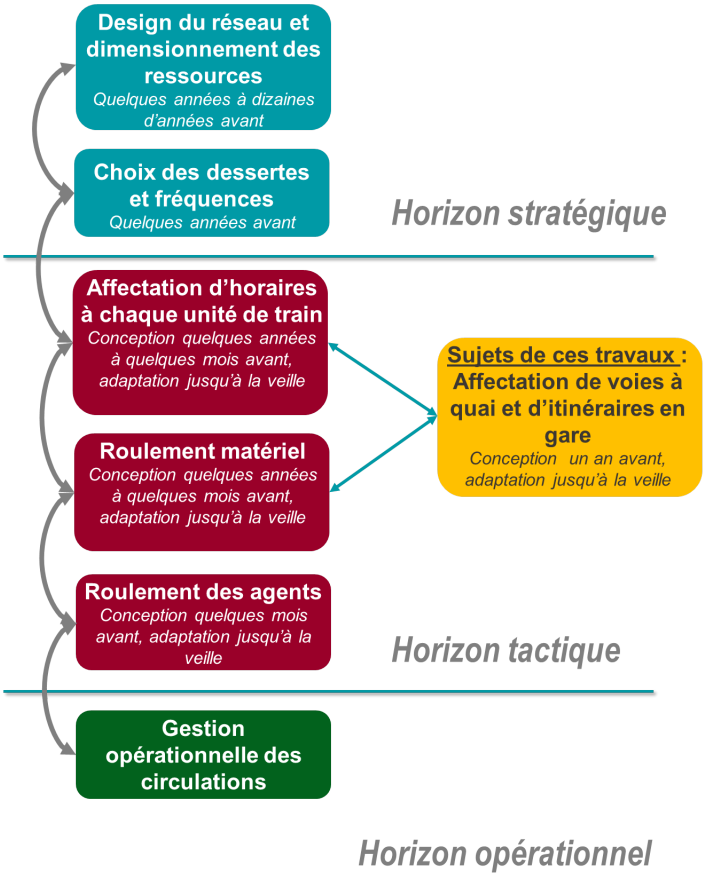

FIGURE 2.1 – Chronologie de la production ferroviaire

— l'horizon stratégique : cette phase se situe plusieurs années avant les opérations, et correspond aux décisions structurantes pour la production ferroviaire

— l'horizon tactique : c'est pendant cette phase que sont construites les planifications sous contraintes des ressources définies pendant l'horizon stratégique. D'un point de vue industriel, l'horizon tactique intègre les phases de conception du plan de transport (quelques années à un an en avance) et d'adaptation du plan (année en cours jusqu'à la veille).

— l'horizon opérationnel : cette étape comprend les décisions à court-terme, c'est-à-dire le jour même des opérations, afin d'ajuster le plan de transport en fonction des aléas rencontrés.

Cette sous-section présente les étapes les plus importantes de la production ferroviaires, et en particulier le problème de construction des grilles horaires qui est le plus étudié dans le domaine. Ces différentes étapes sont également résumées dans la figure 2.1. Une revue plus détaillée des différents problèmes et modèles classiques peut être trouvée dans l'article de Huisman et al [103].

2.1.1.1 Horizon stratégique et tactique

Design du réseau et des ressources : les premières décisions prises à un horizon stratégique sont le dimensionnement des ressources ainsi que les niveaux de service à fournir et changements de

politiques. Par exemple, les grands chantiers de modification de l'infrastructure, les plans de recrutement de personnel ou l'introduction de nouveaux matériels roulants sont décidés plusieurs années en avance. Ces sujets sont peu modélisés mathématiquement mais ils conditionnent fortement la suite.

Choix des dessertes : ce problème stratégique, appelé aussi *line planning*, consiste pour une infrastructure donnée à déterminer un ensemble de missions, c'est-à-dire une station d'origine, une station terminale et des stations intermédiaires où le train marquera l'arrêt, ainsi qu'un type de train et une fréquence de service requise. Cette planification est en général produite à l'aide d'une estimation du nombre de passagers entre chaque paires de stations pour une plage horaire donnée. Plusieurs critères doivent être pris en compte comme les coûts engendrés, la robustesse du système, le temps de trajet et le nombre de correspondances nécessaires pour les passagers. Une revue détaillée a été proposée par Schöbel [158]. La suite des problèmes présentés appartiennent à l'horizon tactique.

Construction des grilles horaires : une fois l'infrastructure, les ressources, les lignes et les fréquences imposées, la construction des grilles horaires, ou *train timetabling problem*, vise à affecter un horaire à chaque unité de train pour répondre aux objectifs fixés à l'étape précédente [40]. Cette étape est cruciale car elle conditionne beaucoup la suite des planifications (allocations de quais, emplois du temps du personnel,...). Il faut anticiper ces planifications futures, par exemple lors du choix des temps de stationnement, mais aussi assurer la robustesse de la grille pour éviter les phénomènes de propagation des retards et prendre en compte la perspective des passagers. La conception des grilles horaires se fait sur une vision macroscopique du réseau, sans considérer les choix d'itinéraires et occupations des voies. Plusieurs objectifs se contredisent alors, puisque par exemple un allongement du temps de trajet peut faire gagner en stabilité puisque les trains peuvent rattraper leur retard. Cependant cela crée aussi une baisse du niveau de service pour le passager et augmente la consommation de capacité, rendant le routage des trains plus difficile.

Roulement du matériel roulant : un roulement de matériel affecte à chaque circulation planifiée une rame physique. Cette planification est soumise à plusieurs contraintes, comme la maintenance des rames à intervalles réguliers, la compatibilité du matériel pour la mission demandée, etc. Dans certains cas, il n'est pas possible d'enchaîner tous les mouvements planifiés, et il est alors nécessaire d'introduire des trajets dits "à vide" qui font circuler une ou plusieurs rames entre deux missions se terminant et commençant à des gares différentes.

Roulement des agents : une fois que les circulations commerciales et techniques sont planifiées, des conducteurs et contrôleurs doivent être affectés à chaque train. Ce problème est généralement appelé *crew scheduling problem* dans la littérature, ou construction des roulements des agents. L'objectif cherché est en général de minimiser les coûts et le personnel nécessaire en intégrant les contraintes de réglementation du travail et de compatibilité avec les compétences des conducteurs sur les lignes et les matériels considérés. Un exemple sur le cas français a été traité par Froger et al [75].

Autres problématiques locales : plusieurs autres planifications peuvent enrichir cette liste, cependant celles qui sont présentées ci-dessus sont plus structurantes car elles concernent une grande partie ou la totalité du réseau ferroviaire. Ces problèmes sont appelés *centraux*, par opposition aux planifications *locales* qui n'ont un impact que sur une zone géographique limitée [103]. C'est par exemple le cas du problème de routage et d'allocation de voies en gare (ou *train platforming problem*), qui est le sujet de ce travail, et qui est présenté plus en détail dans la section 2.2. D'autres planifications locales peuvent s'ajouter à ce problème, comme la gestion des sites de maintenance, les garages de trains ou les emplois du temps du personnel sédentaire.

Problématiques d'assemblage : la coordination des différents pans de la production suscite également de l'attention. De plus en plus de travaux se consacrent par exemple à la conception de modèles intégrés pour optimiser plusieurs problèmes simultanément et éviter les pertes dues à une organisation séquentielle. C'est par exemple ce que proposent Dewilde et al [67] et Burggraeve et al [36] qui étudient le problème de conception de grilles horaires conjointement à celui de routage en gare afin de gagner en robustesse.

2.1.1.2 Horizon opérationnel :

Les retards primaires : un retard primaire, appelé aussi retard exogène ou initial, est une déviation entre l'horaire prévu et l'horaire réalisé due à un facteur externe aux circulations. Il s'oppose à un retard secondaire qui est causé par un train perturbé sur un autre train. Les retards primaires ont des causes très variables, dont les principales sont énumérées ci-dessous [7, 9, 81, 171, 175] :

- Dysfonctionnements de l'infrastructure :
  - Avaries de la voie
  - Dérangements d'installations (système d'aiguillage, feu de signalisation,...)
  - Problèmes de caténaires et d'alimentation
  - Dysfonctionnement de passage à niveau
- Incidents liés aux entreprises ferroviaires :
  - Panne de matériel roulant
  - Disponibilité des ressources humaines (maladies, retards,...) et matérielles (retour de maintenance tardif, réutilisation)
- Aléas opérationnels :
  - Problèmes de communication
- Incidents liés aux passagers :
  - Actionnement intempestif du signal d'alarme
  - Accidents de personne, malaises voyageurs
- Causes externes :
  - Colis suspects
  - Malveillance : vols de câbles, intrusions, vandalisme
  - Environnement : fortes chaleurs, grands froids, crues, feuilles, incendies aux abords des voies, etc
  - Heurts d'animaux et obstacles sur la voie
  - Accidents de passage à niveau
- Autres facteurs d'instabilité du réseau :
  - Habitudes de conduite des agents
  - Temps d'occupation à quai excessif (embarquement/débarquement des passagers)
  - Hétérogénéité des lignes
  - Qualité des planifications

Les retards secondaires : ces retards sont générés par des phénomènes de propagation interne au système ferroviaire. Un retard se propage d'un train à un autre par partage des ressources, comme l'infrastructure, le matériel roulant ou les ressources humaines (le second train doit attendre que la ressource se libère pour l'utiliser), ou encore pour respecter des correspondances entre plusieurs circulations. La qualité des planifications a un rôle majeur dans la propagation puisque le placement stratégique de marges entre deux utilisations d'une même ressource peut permettre de limiter les retards secondaires créés. Les décisions prises en opérationnel pour réguler le trafic vont contrôler la propagation et arbitrer entre différents cas de figure, par exemple selon la priorité des circulations.

Gestion du trafic en temps réel : les conséquences sur le réseau de ces incidents peuvent aller du simple retard de quelques minutes à la réduction de capacité de plusieurs heures (les circulations doivent ralentir, voire être complètement interrompues le temps de l'incident). Suite à ces perturbations, le plan de transport n'est potentiellement plus faisable et doit donc être adapté pour assurer la circulation des trains. Les problématiques de replanifications se posent à chaque niveau de la production ferroviaire : il faut adapter les horaires et temps d'occupation, éventuellement modifier les roulements matériels, changer les affectations de voies, etc. De manière générale, deux trains sont considérés en conflit quand ils cherchent à utiliser la même ressource simultanément (quai, canton, matériel, personnel, etc.).

Contraintes de faisabilité : plusieurs contraintes sont à respecter obligatoirement pour permettre la circulation des trains, à savoir que chaque train doit être seul par canton, les temps de marche et d'arrêts doivent être cohérents, avec des temps de stationnement en gare suffisants, et il faut s'assurer de la disponibilité du matériel et des agents.

Régulation du trafic : Plusieurs règles de replanification (*dispatching rules*) peuvent être appliquées en cas de conflits afin d'assurer la faisabilité des circulations [3, 55] :

- Ajustement des horaires : cela inclut la propagation des retards, mais aussi le fait de faire partir des trains en avance (trains techniques) ou d'ajuster la vitesse des trains. Dans le cas de la propagation des retards, la planification d'origine est conservée (ordre des trains, voie assignée, etc.), et en cas de conflit les trains sont retenus jusqu'à ce que la ressource se libère.
- Dépassements et changements dans l'ordre des circulations : si un train est retardé, il peut y avoir un conflit avec le train suivant passant par la même ligne. Il faut donc arbitrer lequel passera en premier, et éventuellement changer l'ordre pour ne pas retarder le second train.
- Suppression de correspondance : le second train de la correspondance n'attendra pas le train précédent afin d'éviter l'apparition de nouveaux retards sur une ligne.
- Modification des routes et voies affectées : quand un train est retardé en gare et occupe un quai plus longtemps que prévu, une autre voie sera assignée aux trains suivants. En ligne, si la capacité est fortement réduite, par exemple sur une ligne dédiée aux trains grande vitesse, il peut arriver que les trains soient déroutés vers la ligne classique.
- Suppression complète ou partielle (terminus prématuré) des circulations : ce scénario est courant dans le cas de transport en zone dense où supprimer un train permet de retourner plus rapidement en situation nominale car il y a d'autres trains pour absorber la charge.
- Modification de la mission : des arrêts peuvent être supprimés ou être rajoutés
- Changement d'équilibre : si un train est en retard et ne peut pas assurer sa mission suivante, les réutilisations de rames pourront être modifiées en utilisant une nouvelle rame (par exemple provenant d'un site de maintenance) pour assurer la mission.

Ces stratégies ont pour but de sortir de situation perturbée et de faire circuler au maximum les trains prévus. La planification en temps réel est menée selon différents objectifs. Par exemple il est préférable de ne pas trop s'éloigner du plan d'origine, minimiser les conséquences des retards, etc. Des règles de priorité peuvent être mises en places pour arbitrer les conflits, privilégiant par exemple les trains à l'heure, ceux transportant des passagers ou les trains à grande vitesse [3].

Résolution : plusieurs modèles ont été proposés pour répondre à ce problème, et le détails des approches mathématiques de replanification peut être trouvé dans la revue de littérature de Cacchiani et al [37]. Les replanifications des circulations ainsi que celles des roulements de matériels et d'agents y sont abordées. Pour les grilles horaires, la modélisation la plus classique représente les circulations par des graphes d'évènements (*timed-event graph*) qui intègrent à la fois le réseau, les circulations et leurs interactions. Corman [55] distingue les approches réactives qui n'évaluent pas ce qu'il se passera dans le futur et les approches proactives qui anticipent. Étant donné la complexité du système ferroviaire, les modélisations proposées contraignent les règles de gestion opérationnelle. Par exemple Kecman et al considèrent les correspondances comme fixées [108]. Corman [55] prévient qu'un grand nombre de degrés de liberté donne de meilleures solutions mais demande une trop grande complexité.

2.1.2 Qualité des planifications ferroviaires

Cette partie présente plusieurs aspects qui peuvent être optimisés dans les différentes étapes de la production. La plupart des indicateurs sont développés pour le problème de conception des grilles horaires. Le cas spécifique du problème des affectations de voies en gare et de sa robustesse est traité séparément dans la section 2.2. Les questions de sécurité et de dimensionnement des ressources ne sont pas abordées autrement que par la faisabilité. Ils sont pensés en amont des planifications et n'interviennent pas dans les problèmes présentés ici mais constituent des prérequis au bon fonctionnement.

Principaux aspects : Goverde et Hansen ont proposé un panorama des attributs cherchés d'une grille horaire [84].

- la faisabilité : indicateur fondamental de la qualité d'une grille horaire, et de toute autre solution. Il faut que les trains soient en mesure de respecter les horaires et routes choisis tout en satisfaisant les contraintes de sécurité du système.
- la stabilité : la capacité d'un train à absorber des retards exogènes et à revenir à son état d'origine sans replanification. La stabilité concerne particulièrement la conception des grilles horaires où on souhaite évaluer les marges additionnelles à ajouter aux temps de marche afin de pouvoir absorber les variations de temps de trajet.
- la robustesse, qui est fortement liée à la stabilité, prévient l'apparition des retards secondaires en utilisant à la fois des marges sur les temps de trajet et en jouant sur les espacements entre les trains pour éviter la propagation des retards. Goverde et Hansen différencient entre autre la stabilité de la robustesse par leurs caractères respectivement déterministe et stochastique.
- la résilience : la capacité à pouvoir prévenir et réduire les retards en temps réel en utilisant de la replanification.

Ces notions s'étendent à d'autres problèmes que celui des grille horaires. En particulier la faisabilité et la robustesse sont des atouts majeurs du système ferroviaire en raison de son organisation séquentielle. Les différentes étapes de la production, comme montré dans la figure 2.1, sont construites de manière à permettre la faisabilité des planifications effectuées en aval. Par exemple une grille horaire est ajustée si elle ne permet pas la réalisation d'un planning des occupations en gare faisable. Au delà de la faisabilité, la robustesse de chaque sous-système contribue au bon fonctionnement global. De

nombreuses recherches travaillent maintenant sur des formulation intégrées pour résoudre simultanément plusieurs pans de la planification et assurer ainsi un meilleur fonctionnement des opérations.

Objectifs complémentaires : d'autres aspects sont parfois cherchés dans une solution :

- Mesures industrielles a posteriori : les autorités de régulation, entreprises ferroviaires et gestionnaires d'infrastructure jugent le respect des objectifs de qualité de service via plusieurs indices [8, 9]. Les plus communs sont la ponctualité et la régularité. Leurs définitions varient selon les périmètres mais la ponctualité mesure les retards, en général via le pourcentage de trains dont le retard est inférieur à un seuil donné, et la régularité intègre les annulations par exemple avec le taux d'annulation de circulations programmées. D'autres indicateurs comme l'information voyageur, les réclamations ou le confort sont évalués pour mesurer le niveau de service.
- Qualité de l'offre de transport : temps de trajet, trajets directs, fréquence, etc. [171].
- Optimisation de la capacité résiduelle : la capacité représente l'utilisation de l'infrastructure et se définit comme le nombre de trains qu'on peut faire circuler sur une infrastructure donnée pendant une certaine période de temps. L'optimisation de la capacité résiduelle est proche des questions de stabilité, de robustesse et de résilience car elle permet de faciliter le reroutage ou l'absorption des retards en cas de perturbations. Plusieurs études se sont intéressées aux interdépendances entre la capacité et les autres indicateurs de performance et montrent qu'une forte utilisation de la capacité (nombre de trains, circulations hétérogènes, etc.) a un impact négatif sur la fiabilité et les retards, et à l'inverse, augmenter la robustesse, par exemple en imposant des espacements ou des temps de trajets plus importants, baisse directement la capacité comme moins de trains peuvent circuler. [11, 117, 131]. La capacité résiduelle est régulièrement utilisée comme mesure de stabilité [22, 84].
- Retards des passagers : le retard des trains est le plus souvent étudié, cependant il diffère du retard perçu par les passagers. Ainsi, une correspondance manquée pourra faire perdre plusieurs heures supplémentaires à un passager alors qu'un retard systématique sur tous les trains d'une ligne desservie à fréquence haute ne le retardera pas ou peu. C'est par exemple ce que font Sels et al [160] en modélisant les retards primaires des trains et en déduisant la valeur des retards des passagers. Takeuchi et al [165] évaluent la robustesse d'un plan en fonction de l'impact moyen qu'il aura sur le confort des voyageurs.
- Equité : dans un contexte d'ouverture à la concurrence, les gestionnaires d'infrastructure doivent garantir un accès au réseau équitable pour les différentes entreprises ferroviaires.
- Cadencement : les acteurs industriels recherchent parfois la praticité des planifications. L'exemple le plus important est celui du cadencement des horaires, mais on peut aussi avoir des voies dédiées et des affectations systématiques de routes.
- Consommation énergétique : des profils de vitesse sont optimisés conjointement à la conception des grilles horaires pour réduire la consommation électrique en ligne [180].

2.1.3 Robustesse des grilles horaires

Dans un contexte de recherche opérationnelle, une solution robuste d'un problème est définie comme une solution qui restera faisable si les paramètres du problèmes subissent de petites variations. La robustesse est particulièrement étudiée dans le cadre de problèmes d'optimisation où on cherche à atteindre un équilibre entre robustesse et qualité de la solution [21].

La notion de robustesse en recherche opérationnelle ferroviaire diffère quelque peu de cette définition, et est parfois spécifique au cas d'étude et au problème considéré. La plupart des travaux du

domaine ont été effectués dans le cadre des grilles horaires, cependant d'autres références sur la robustesse pour les conceptions de ligne, les roulements de matériel ou les roulements d'agents peuvent être trouvées dans l'état de l'art de Lusby, Larsen et Bull [124].

On retiendra la définition générale de Andersson et al [15] qui considèrent une grille horaire comme robuste si les trains peuvent conserver leur créneau d'origine malgré quelques petits retards. Cette définition contient deux éléments clés présents dans de nombreux autres travaux : la robustesse ne doit pas ou peu inclure de déformation de la grille (suppression d'arrêts ou autres modifications liées à la gestion opérationnelle), et la robustesse ne concerne que les petites perturbations, les retards importants nécessitant des ajustements trop complexes et spécifiques pour être raisonnablement anticipés.

Khadilkar [110] donne une définition alternative plus proche de la définition industrielle en considérant que la robustesse est une fonction de l'ensemble des décisions prises, que ce soit au niveau stratégique par l'infrastructure, au niveau tactique avec le choix de la grille ou au niveau opérationnel avec les politiques de gestion des circulations, en gardant toujours cet objectif de limiter la propagation des retards primaires. Il complète cette définition par deux perspectives de robustesse : la robustesse individuelle, qui consiste à limiter les risques de propagation des retards primaires de chaque train vu seul sur les circulations qui l'entourent, et la robustesse collective qui évalue les conséquences globales quand plusieurs trains sont soumis à des retards primaires.

2.1.3.1 Mesure

La robustesse d'une solution peut s'évaluer de deux manières : analytiquement par des indicateurs ou par de la simulation. Les indicateurs de performance se calculent à partir des caractéristiques de la solution. Ils sont accessibles mais ne donnent qu'une vision partielle de ce qui se passerait avec des perturbations, et souvent ne qualifient que la robustesse individuelle. La simulation fonctionne sur un échantillonage aléatoire des données d'entrée qu'on applique à la solution pour étudier son comportement en situation perturbée. Elle donne une évaluation plus complète et réaliste mais requiert un important travail de modélisation en amont. Dans le cadre de la simulation ferroviaire, il faut décrire les perturbations, par exemple avec la distribution de probabilité des retards, mais il faut aussi modéliser le fonctionnement et les décisions prises par les agents en opérationnel. Différentes stratégies d'évaluation de la robustesse ferroviaire et en particulier pour les grilles horaires, sont présentées ici.

Méthodes analytiques : une des mesures les plus utilisées, en particulier chez les acteurs industriels, est la formule de compression des grilles horaires proposée par l'UIC 406 [2] qui permet d'estimer la capacité résiduelle sur les lignes. L'indicateur se construit à partir du taux d'occupation de l'infrastructure sur une période donnée.

Salido et al [157] définissent la *t-robustesse* comme le pourcentage des perturbations de moins de t unités de temps qu'une grille horaire peut absorber.

Andersson et al [15] construisent une base d'indices de robustesse de référence, avec notamment le nombre de trains par section et par heure, la quantité total de marge par train, la somme des inverses des plus petits espacements, le pourcentage d'espacement inférieurs à un seuil donné, etc. Ils comparent ces indices avec un nouvel indicateur basé sur les marges aux points critiques des trajets de chaque train.

Jensen et al [104] énumèrent d'autres mesures complémentaires. L'indice d'hétérogénéité quantifie indirectement la dispersion des espacements en étudiant les valeurs minimales entre les trains sur une ligne. L'indice de complexité correspond à la fois à la complexité de l'infrastructure (nombre de routes qui se croisent) et à la complexité des opérations en calculant la proportion de couples de trains utilisant des routes en conflit.

Burdett et Kozan [35] mesurent la robustesse d'une grille par une analyse de sensitivité. Ils la décomposent pour cela en opérations élémentaires, et identifient pour chacune d'elle et pour chaque

valeur de retard la liste des opérations qui seraient impactées. La robustesse est alors vue comme la valeur de sensibilité maximale, soit, pour un niveau de retard donné, la taille de la liste maximale d'opérations impactées.

Une alternative à ces indicateurs est l'utilisation d'algèbres max-plus qui permet d'intégrer efficacement les contraintes de précédence pour évaluer la qualité d'une grille horaire cyclique [171]. En particulier, ces algèbres permettent de calculer le temps minimal nécessaire pour réaliser une période de la grille, ou la marge de stabilité, c'est-à-dire la quantité de suppléments de temps qu'on peut ajouter aux processus sans excéder la durée du cycle.

Simulation : Takeuchi et al [165] estiment la robustesse en mesurant de manière probabiliste le niveau d'inconfort pour les usagers. Étant donné que la probabilité associée à l'inconfort est inconnue, l'indice est calculé par simulation en générant des retards et en évaluant leur impact sur les usagers.

Liebchen et al [121] construisent des modèles d'optimisation de la robustesse de grilles horaires grâce à l'évaluation de la résistance aux retards par simulation de plusieurs scenarii. Un graphe d'activité est utilisé, et différentes sources de retards et règles de régulation sont modélisées. Cette approche permet à la fois de prendre en compte le caractère stochastique des données d'entrée (durée de parcours, stationnement, etc.) et la diversité de règles de réaction à ces retards.

Larsen et al [118] introduisent également une méthodologie d'évaluation de la robustesse par simulation. En particulier, ils suggèrent quatre indicateurs basés sur les retards propagés : le retard secondaire moyen, le retard secondaire maximal, le retard total (primaire et secondaire) maximal, et le retard total moyen. Le métriques évaluant les retards secondaires sont cependant à privilégier, étant donné que les retards primaires sont considérés comme inévitables et ne dépendent pas de la qualité de la grille.

2.1.3.2 Optimisation

Les approches d'optimisation robuste classiques sont rapidement trop conservatrices quand elles sont appliquées au problème de conception des grilles horaires. Étant donné le niveau de service requis en terme de fréquence et de temps de trajet et les fortes contraintes d'exploitation, une solution est très vite non réalisable quand elle est perturbée, même pour un retard faible. Les aléas possibles sont par ailleurs nombreux et d'amplitude asymétrique et variable, ce qui complique la description d'un ensemble d'incertitude réaliste.

Pour pallier cette difficulté, de nombreuses études construisent des fonctions objectifs qui optimisent indirectement la robustesse sans utiliser ce cadre classique. Les deux principaux leviers d'amélioration de la robustesse qui sont exploités sont l'ajout de marges dans les durées prévues pour absorber d'éventuelles perturbations et l'ajustement des espacements (*buffer times*) entre des trains consécutifs afin de diminuer la propagation. Yuan et Hansen [184] proposent un modèle d'allocation d'espacements dont la somme est bornée en fonction du risque de création de retards secondaires en considérant une loi exponentielle comme distribution des retards primaires. Jovanović et al [105] utilisent une modélisation basée sur un graphe d'évènement pour construire une grille horaire robuste en n'ajustant que les espacements. La quantité de marge disponible est évaluée par compression puis optimisée de manière à réduire le nombre d'évènements impactés par des retards. Cette quantité est estimée par la taille des voisinages de sensibilités, c'est-à-dire l'ensemble des trains qui seraient impactés par les retards supérieurs à un seuil fixé pour un évènement donné.

D'autres études intègrent les variations possibles des horaires dans leurs modélisations. Liebchen et al [120] introduisent la notion de robustesse de récupération. On définit un problème d'optimisation, un ensemble de scenarii de perturbations $\mathcal{S}$ et les algorithmes de récupération $\mathcal{A}$. Une solution est alors robuste si pour tout $s \in \mathcal{S}$ on peut réparer la solution. Ce cadre de travail est ensuite appliqué au *Timetabling problem* pour lequel les perturbations de $\mathcal{S}$ sont des petits retards bornés et les algorithmes

de récupération sont la propagation des retards et l'annulation des correspondances. Le cadre de travail de la robustesse légère énoncée par Fischetti et Monaci [71] utilise le principe de robustesse de Bertsimas et Sim [21] où les contraintes doivent rester faisables malgré des variations bornées des coefficients. Des variables de recours sont ajoutées au modèle pour permettre de relâcher ces contraintes quand cela est nécessaire.

Un modèle d'optimisation stochastique à deux étapes a été proposé par Kroon et al [112] afin de modifier les suppléments de temps de trajet et les espacements de manière optimale en terme de retard total créé en considérant un ensemble de retards primaires modélisés par une loi exponentielle tronquée à 10 minutes.

2.2 Problème d'occupation des voies en gare

Les gares ont un rôle primordial dans l'organisation de la production ferroviaire. De nombreuses circulations s'y croisent et elles centralisent des activités variées (interface avec les voyageurs, roulement agents, roulement matériel, préparation des rames, etc.). La gestion des circulations en gares nécessite cependant d'être conçue avec soin car la congestion y est importante. Malgré le caractère local des planifications en gare, des fragilités dans les affectations de voies peuvent avoir des conséquences fortes le long des lignes ferroviaires en raison des phénomènes de propagation. On étudie ici les caractéristiques de ce problème, ainsi que les pistes identifiées pour améliorer la résistance aux retards.

2.2.1 Le Train Platforming Problem

Le problème d'affectation de voies en gare consiste à affecter à chaque circulation une voie à quai ainsi qu'un itinéraire dans le périmètre de la gare pour s'y rendre ou pour en repartir. Cette planification est localement construite une fois que la grille horaire est proposée quelques mois en avance et peut être ajustée jusqu'à la veille.

Trouver une planification faisable des occupations de voies en gare est très difficile en raison de la forte combinatoire liée à la complexité de l'infrastructure et des contraintes associées, et même en pratique souvent impossible en raison de la congestion du réseau. Au delà de la faisabilité, des questions de robustesse et de résilience des planifications se posent : il ne s'agit pas juste de trouver un ordonnancement faisable, ou proche de la faisabilité, mais également de l'optimiser pour réduire l'apparition des retards.

Le problème de recherche opérationnelle associé s'appelle le *Train Platforming Problem* (TPP). Le cas des gares est moins traité que celui des grilles horaires, mais un certains nombres d'études sont tout de même consacrées au TPP. En particulier, les articles de Zwaneveld et al [187] et de Kroon et al [113] sont parmi les premiers à l'avoir théorisé.

Dans sa version la plus simple et la plus générale, le TPP se ramène au problème de faisabilité consistant à trouver des affectations trains/itinéraires, où un itinéraire est constitué d'un chemin d'arrivée, un quai et un chemin de départ, en respectant les horaires des trains qui sont imposés et plusieurs contraintes. Il peut s'agir de contraintes de sécurité, de disponibilité de la ressource ou encore de contraintes commerciales. Les contraintes sont de deux ordres : celles interdisant l'utilisation d'une route par un train, et celles restreignant l'affectation simultanée d'une paire d'itinéraires par deux trains (par exemple s'ils veulent utiliser la même ressource en même temps, si les itinéraires se croisent, etc.).

Zwaneveld et al [187] établissent la NP-completude de ce problème en montrant son équivalence au problème de planification de n taches sur k machines non-identiques. Kroon et al [113] montrent également que le TPP est NP-complet à partir du moment où il peut exister au moins 3 routes

possibles par train en considérant une transformation polynomiale du problème 3-SAT, mais qu'il est polynomial quand tous les trains ont au maximum deux routes possibles.

Étant donné la congestion des gares, il est généralement impossible de trouver une solution réalisable au problème d'affectation des voies en gare, on va donc souvent optimiser le nombre de contraintes respectées et minimiser le nombre de trains non affectés. En pratique, tous les trains seront routés quitte à violer certaines contraintes, ce qui risque de générer des retards en opérationnel.

Plusieurs variantes du problème peuvent être rencontrées dans la littérature :

— la plus classique considère qu'il n'existe qu'un itinéraire possible entre un quai et les lignes d'arrivée et de sortie [24, 48, 161]

— les horaires des trains sont le plus souvent considérés comme fixes, mais certaines études permettent des changements limités dans la grille horaire [48, 68]. Dans certains cas, les quais sont supposés fixes. Par exemple Dewilde et al [68] cherchent à optimiser les horaires pour assurer un routage robuste dans les différentes gares du réseau.

— Sels et al [161] proposent une modélisation intégrant une incertitude sur la quantité de trains à placer. Une affectation de voies à quai est donnée à partir d'un ensemble de trains connus et d'un ensemble de nouveaux trains à placer pour assister le gestionnaire d'infrastructure dans son estimation de la capacité à augmenter le trafic prévu.

— Les gares sont parfois vues comme un cas particulier des jonctions où un arrêt est planifié. Une jonction est un point du réseau où de nombreuses lignes se rencontrent, avec comme les gares des points d'entrée et de sortie, et au sein duquel on cherche un routage pour une grille horaire fixée [65, 123].

— contraintes spécifiques : par exemple Carey et Carville [48] intègrent le cas où plusieurs trains différents sont réceptionnés sur la même voie à quai.

— la modélisation des itinéraires et contraintes en gare est très généralement macroscopique, cependant quelques études utilisent une modélisation microscopique des ressources. C'est par exemple ce que proposent Bešinović et Goverde [22] en découpant l'infrastructure en sections élémentaires homogènes (vitesse, courbure, etc.).

— la faisabilité de la solution varie : certains auteurs imposent le respect de toutes les contraintes mais introduisent des quais et itinéraires fictifs [159], d'autres modèles autorisent les violations de contraintes quand le conflit est réduit, et pénalisent en objectif ces violations [43].

On peut ajouter à cette liste les approches robustes qui sont étudiées dans la partie 2.2.3.

2.2.2 Résolution

Plusieurs méthodes de résolution exacte ou approchée du problème d'affectation des voies ont été proposées. On détaille ici la structure classique de ces méthodes, notamment le travail sur les données d'entrées et les différentes modélisations adoptées.

2.2.2.1 Prétraitement des données

La planification des occupations en gare repose sur trois types de données d'entrée : les horaires d'arrivée et de départ de chaque train, la description de l'infrastructure et les contraintes à respecter. La résolution mathématique de ce problème nécessite un prétraitement important de ces données [103]. En particulier, les différentes approches de résolution reposent sur l'énumération des itinéraires traversant la gare, la construction de sous-ensembles d'itinéraires admissibles pour chaque train en fonction

de ses caractéristiques (longueur de train, voies interdites, etc.) et la liste des paires d'affectations train/itinéraire qui occasionnent un conflit.

La liste des itinéraires et des combinaisons trains/routes est en générale très longue et il est recommandé de la réduire avant l'optimisation [68]. Il faut pour cela supprimer les itinéraires *dominés*. Un itinéraire est considéré dominé par un autre s'il immobilise au moins autant les mêmes aiguilles. Cela implique que remplacer cet itinéraire dominé par l'itinéraire dominant dans une solution est toujours équivalent ou meilleur. D'un point de vue théorique, cela correspond à des paires de variables où les variables saturées par la variable dominante sont incluses dans la liste des variables saturées par celle qui est dominée [65]. La liste des routes admissibles pour chaque train peut également être réduite si certaines routes sont d'emblée incompatibles avec les autres circulations les plus contraintes.

2.2.2.2 Formulations pour la résolution exacte

Modélisation par stable : le problème de routage des trains en gare et d'affectation de voies peut se voir comme un problème de stable maximum (*node packing*, c'est-à-dire un ensemble de noeuds non connectés entre eux de cardinalité maximum) en construisant un graphe d'incompatibilité des circulations. On note T l'ensemble des trains formés par un mouvement d'arrivée et un mouvement de départ, et R l'ensemble des routes complètes, c'est-à-dire avec un chemin d'entrée en gare, une voie à quai et un chemin de départ, et pour $t \in T$ on note R_t l'ensemble des itinéraires autorisés à t. On construit un graphe non orienté $G = (V, E)$, avec dans V l'ensemble des sommets formés par une association train-itinéraire admissible, c'est-à-dire un sommet $v = (t, r) \in V$ avec $t \in T$ et $r \in R_t$. Les arêtes du graphe représentent les paires d'affectations interdites dans la même solution. Les sommets correspondant à des trains utilisant des itinéraires incompatibles (s'ils se croisent, ont une voie ou un quai en commun, etc.) dans un intervalle de temps trop court et les sommets associés au même train sont donc reliés.

Le problème d'affectation de voie est bien équivalent à un problème de stable maximal : on cherche à sélectionner un ensemble de sommets du graphe d'incompatibilité de taille maximale et disjoints entre eux. Les sommets sélectionnés donneront des affectations réalisables d'itinéraires aux trains pour une journée donnée.

$$\begin{aligned} Max & \sum_{t \in T, r \in R_t} x_{t,r} \\ & x_{t,r} + x_{t',r'} \leq 1, \quad \forall ((t,r),(t',r')) \in E \\ & x_{t,r} \in \{0,1\}, \quad \forall (t,r) \in V \end{aligned} \tag{2.1}$$

Cette modélisation par stable a souvent été utilisée dans la littérature. Un des premiers exemples est donné par Zwaneveld et al [187] qui a appliqué un algorithme de Branch & cut pour résoudre un tel programme en 1996. Ils montrent que cette approche permet d'utiliser des résultats classiques du problème de stable maximal, notamment des inégalités valides. D'autres variantes ont été proposées. Caprara et al [43] imposent que tous les trains soient placés, optimisent par rapport à des préférences d'affectation et contraignent les affectations incompatibles sur des cliques de E et non sur des arêtes.

Set packing : Velasquez et al [170], Delorme et al [65] Lusby et al [123] ont proposé d'utiliser une formulation alternative en voyant le routage dans une jonction ou une gare comme un problème de *set packing*. Ce problème classique consiste à trouver dans la liste L_S de sous-ensemble d'éléments d'un ensemble fini S un nombre maximal de sous-ensembles disjoints entre eux. Ici, l'ensemble S est formé par les associations heure/section de voie et les sous-ensembles de S étudiés dans L_S sont des successions de voies formant des itinéraires admissibles pour chaque train. Ainsi, un élément de L_S est

sélectionné dans la solution optimale si le train correspondant peut être affecté au sillon formé sans générer de conflit avec les autres éléments sélectionnés.

On peut considérer la modélisation par le PLNE suivant. Soit T l'ensemble des trains de la journée, $M = V \times P$ l'ensemble des couples section de voie / période, et R_t la liste des sous-ensembles de M formant une route admissible pour $t \in T$. Il y a n variables de décisions binaires x_r avec $n = \sum_{t \in T} |R_t|$, et chaque variable représente le sous-ensemble de ressources choisies pour router le train t parmi les routes de R_t.

$$\begin{aligned} Max & \sum_{t \in T, r \in R_t} x_{t,r} \\ & \sum_{r \in R_t} x_r \leq 1, \quad \forall t \in T \\ & \sum_{t \in T} \sum_{r \in R_t : m \in r} x_r \leq 1, \quad \forall m \in M \\ & x_r \in \{0,1\}, \quad \forall r \in R_t, t \in T \end{aligned} \tag{2.2}$$

L'avantage de cette formulation selon les auteurs est qu'elle dispense d'énumérer tous les conflits entre paires de trains mais les intègre directement en vérifiant que les mêmes ressources ne sont pas utilisées en même temps par deux trains, cependant le nombre de contraintes dépend aussi de la taille de M, c'est-à-dire de la finesse du découpage en périodes de temps. Lusby et al [123] et Velasquez et al [170] utilisent un algorithme de génération de colonnes.

Problème de coloriage : Billionnet [24] propose de résoudre le problème d'affectation des quais sans routage en le formulant comme un problème de k coloration de graphes où chacune des k couleurs correspond à un quai, et deux trains sont reliés dans le graphe s'ils sont en gare au même moment. Les contraintes d'exploitation sur les quais inaccessibles sont prises en compte par des listes de couleurs admissibles pour chaque train. Les croisements des itinéraires menant aux quais sont contraints par un ensemble d'affectations interdites (paire de train et paire de couleurs incompatibles ensembles).

2.2.2.3 Heuristiques et métaheuristiques

Zwaneveld et al [187] proposent deux heuristiques pouvant notamment être utilisées comme initialisation des algorithmes de branch & cut. La première étudie chaque train t et lui affecte l'itinéraire r maximisant le nombre de trains pour lesquels il reste au moins une route compatible après affectation de r à t. La seconde fonctionne sur la même idée mais choisit l'itinéraire qui maximise le nombre de routes disponibles après routage de t pour les trains non placés.

Carey et Carville [48] utilisent une heuristique basée sur les mêmes mécanismes que les agents en charge de la construction manuelle des plans d'occupation des voies. Les trains sont étudiés individuellement par ordre chronologique, et chacune des voies à quai est proposée. En cas de conflit, les horaires du train peuvent être ajustés afin de lui affecter une voie. La qualité de la solution obtenue est jugée sur la quantité de modifications apportées aux horaires, le respect des préférences dans les choix des quais et enfin les temps de réoccupation.

Delorme et al [65] appliquent la métaheuristique GRASP sur le problème de routage dans une jonction vu comme un problème de set packing. Cet algorithme consiste à construire des solutions initiales à l'aide d'algorithmes gloutons randomisés puis à les améliorer par recherche locale.

Clarke et al [52] appliquent un algorithme génétique pour optimiser le placement des trains à partir d'une population de solutions initiales générées aléatoirement. Les solutions privilégiées dans les étapes de sélection sont celles avec le moins de conflits entre les trains.

Dewilde et al [68] optimisent conjointement le routage des trains et leurs horaires à l'aide d'une méthode tabou afin de maximiser les espacements entre les trains lors des routages en gare. La recherche tabou explore des voisinages reposant sur des changements d'horaires (bornés) pour un routage donné, puis le routage en gare est optimisé pour les nouveaux horaires proposés.

Bešinović et Goverde [22] proposent une approche originale basée sur une algèbre max-plus et un algorithme de propagation des retards afin d'améliorer la qualité des affectations (stabilité et robustesse). Leur algorithme fonctionne par permutations de routes, en particulier ils proposent une série de règles d'exclusions d'itinéraires, par exemple en fonction de la criticité des ressources qui le composent, ainsi que des règles d'inclusions qui sélectionnent des itinéraires alternatifs.

2.2.3 Robustesse en gare

Les mesures de robustesse, comme celles de capacité, ont eu un écho moins important en gare qu'en ligne. Plusieurs métriques et définitions peuvent s'adapter, à la différence que le problème de routage des trains en gare est un problème de faisabilité et non d'optimisation : les horaires sont fixés et l'objectif est d'affecter des quais et itinéraires aux trains. La robustesse s'étudie alors indépendamment du temps de trajet. On considère en général qu'une solution est robuste si elle minimise la quantité de retards créée sous de petites perturbations. Dans le cas des gares, elle est étudiée soit du point de vue des espacements entre les trains, soit du point de vue de la gestion de la capacité en gare [41]. Tout comme le problème des grilles horaires, on peut séparer les mesures de robustesse par indicateurs et celles par simulation.

Indicateurs de robustesse : elle est en général mesurée à partir des espacements entre les trains sur une même ressource. Cette absence de marge est responsable de la propagation des retards.

Carey [46] mesure la robustesse avec le risque de propagation en fonction des espacements. Il présente une famille de métriques probabilistes pour estimer le risque de retard secondaire de chaque train. Il s'agit de mesures locales où le risque de retard d'un train est calculé explicitement en fonction de sa probabilité de retard exogène et de la séquence de trains qui le précèdent. Une autre alternative est proposée avec des mesures déterministes visant à quantifier la dispersion des espacements. L'ordonnancement le plus robuste est celui où les plus petits espacements sont maximaux. Landex et al [116] construisent également un indice égal aux taux de cas où l'espacement entre deux trains est inférieur à un seuil donné parmi l'ensemble des paires de trains occupant des voies en conflits directement l'un après l'autre. Ce critère est également utilisé pour optimiser la robustesse dans certaines études, avec l'idée qu'un routage où les espacements sont maximaux va limiter la propagation des retards [68].

Simulation : Carey et Carville [47] considèrent qu'une planification robuste engendre un minimum de retards secondaires si elles est soumise à de petites perturbations. Ils proposent donc une méthodologie où des petits retards sont générés à l'aide de lois de probabilité de référence, puis injectés dans une solution du TPP. La solution n'est en général plus réalisable et est réparée en propageant les retards. Une alternative considère de changer l'affectation de quai quand cela est possible pour éviter la propagation, ce qui limite fortement le nombre de minutes créées. Le simulateur peut alors permettre de quantifier l'impact d'une décision sur les performances.

Dewilde et al [68] évaluent le gain de robustesse apporté par leur approche en simulant des retards par une loi exponentielle calibrée sur le retard moyen et en mesurant en sortie le pourcentage de trains retardés et la quantité de retards secondaires générés.

Plus récemment, Bešinović et Goverde [22] ont construit un modèle d'évaluation de la robustesse basé sur de la simulation de petites perturbations par une loi exponentielle dans un graphe d'évènements. Cette estimation sert ensuite à alimenter une heuristique pour gagner en robustesse.

Optimisation de la robustesse : elle est de plus en plus souvent intégrée à la résolution du TPP. On recense les stratégies suivantes.

Le cadre de travail de la robustesse de récupération de Liebchen et al [120] a été appliqué également au problème d'affectation des voies par Caprara et al [42]. L'idée est de se protéger contre la propagation des retards en assurant que la solution pourra revenir en situation nominale grâce à un algorithme de récupération si le scénario fait partie d'un ensemble de perturbations défini lors de l'optimisation. Ici, les perturbations couvertes sont représentées par un ensemble de scénarios de retards des trains budgetisés par Δ, c'est-à-dire tel que la somme des retards individuels dans le scénario soit inférieure à Δ. La récupération est permise ensuite par la propagation des retards entre les trains, sans autre opération de replanification. Une formulation est proposée utilisant la formulation classique de Caprara et al [43] à laquelle est ajouté un ensemble de variables représentant le retard de chaque train pour chaque scénario de perturbation. La détérioration dans le pire des cas, qui est égale à la plus haute somme de ces variables pour le pire cas de perturbation, est minimisée en objectif.

Caprara et al [41] soumettent deux autres stratégies de robustesse des routages en gare : l'optimisation de la capacité d'absorption des retards et les quais de secours. La première vise à assurer des espacements suffisants entre deux occupations d'une même ressource par deux trains en les pénalisant. Cette méthode va consommer plus de capacité en avant gare mais permet d'absorber les petites perturbations sans replanifier. Utiliser des quais de secours donne des stratégies en cas de perturbations. Dans ce cas là, il est préférable de donner un quai de secours proche du quai d'origine.

Burggraeve et al [36] proposent d'intégrer les retards récurrents des trains dans l'optimisation des espacements pour limiter la création des retards secondaires. Pour cela, les trains sont étudiés par paires, et en fonction des retards moyens des deux trains de la paire, une pondération utilisée en objectif est assignée pour signifier si l'espacement est insuffisant ou non.

Bešinović et Goverde [22] optimisent conjointement la stabilité, mesurée par la marge disponible sur les ressources les plus critiques, la robustesse, vue comme la quantité cumulée de retards secondaires créés par simulation, et l'utilisation équilibrée des voies, ce qui permet une robustesse plus générale en évitant une trop grande sollicitation de ces ressources. Cette combinaison de ces trois objectifs donne de bonne performances.

2.3 Données dans le transport ferroviaire

Les progrès technologique des dernières décennies ont accompagné et motivé la modernisation de l'offre de transport. Outre la mise en place d'outils d'aide à la décision pour l'optimisation de la production, la collecte et l'analyse de données s'est de plus en plus démocratisée. Le spectre des données recueillies s'agrandit, et elles deviennent de plus en plus accessibles. L'utilisation principale de ces données concerne encore majoritairement le suivi des opérations et de la performance, ainsi que la calibration des paramètres d'entrée (demande, temps de parcours, etc.), cependant de nouvelles méthodologie émergent. En particulier, les approches prédictives sont prometteuses pour adapter au mieux l'offre à la demande ou aux contraintes opérationnelles. Cette section énumère tout d'abord les bases de données principales collectées lors des opérations ferroviaires. Des exemples de valorisation de ces bases sont ensuite brièvement présentés.

2.3.1 Les bases de données classiques

Il existe deux types de données ferroviaires : les données automatisées, dont la collecte se fait à l'aide de capteurs, et les données dites manuelles dont l'enregistrement dépend de l'intervention d'un agent [60]. L'utilisation de ces premières est plus courante dans la littérature. En effet, les données collectées automatiquement sont plus fiables et plus structurées (le format des données reste

en général homogène et est lié au système d'enregistrement), mais elles nécessitent l'installation de capteurs, de maintenance et le déploiement d'un système d'information. Au contraire, les données non automatisées seront moins structurées et plus subjectives. C'est par exemple le cas des données textuelles des rapports d'incidents.

Collecte automatique : les systèmes de collecte automatique de données se sont fortement développés dans le transport public et en particulier dans le transport ferroviaire. Ils sont souvent référés en tant que ADCS dans la littérature (*Automated Data Collection Systems)*. Il en existe trois familles principales [78, 168, 176] :

— Automatic Vehicle Location (AVL) : ce sont des relevés horodatés de la position des trains au cours du temps. Ces données peuvent être obtenues par des relevés GPS ou à l'aide de balise fixes qui enregistrent les passages et occupations des voies des trains. Ces données sont en général remontées en temps réel pour permettre un suivi des circulations, mais sont également archivées pour pouvoir analyser les performances. Le niveau de détail de l'information peut être variable, tant par le pas de temps utilisé que par la précision de la localisation. Les données les plus précises sont liées à la signalisation et enregistrent ainsi chaque passage, occupation et libération d'une section de voie. Les retards sont obtenus en évaluant les déviations entre ces observations et la grille horaire prévue.

— Automatic Passenger Counting (APC) : de nombreux systèmes ont été développés pour estimer automatiquement l'affluence dans les trains et dans les stations, ainsi que le nombre de passagers montants et descendants des trains à chaque arrêt. Les systèmes APC utilisent par exemple des technologies comme le wifi, le bluetooth, des capteurs infrarouge ou de l'analyse des vidéos.

— Automatic Fare Collection (AFC) : ces données contiennent les validations des titres de transport, avec le lieu et l'heure de passage. Selon le réseau, certains systèmes enregistrent le point d'entrée et le point de sortie du réseau. La constitution de ces bases de données est supportée par l'utilisation de plus en plus fréquente de *smart cards* (cartes d'abonnement et cartes prépayées) qui permettent un suivi plus long et individualisé des trajets empruntés par les usagers.

Il existe d'autres données collectées automatiquement mais dont l'usage est moins généralisé et la collecte est moins organisée. Par exemple les données liées aux requêtes effectuées sur les applications ou sites internet des transporteurs peuvent être utilisées pour contribuer à estimer a demande et les messages postés sur les réseaux sociaux, en particulier Twitter, peuvent aider à mieux comprendre l'expérience utilisateur. Dans un autre registre, la consommation énergétique peut également être étudiée, par exemple pour optimiser la consommation en ligne, ou encore pour calibrer la facturation. Quelques sujets d'analyse d'images émergent également, notamment sur les bases de données constituées par des caméras embarquées sur les trains qui permettent de détecter des anomalies avant qu'un incident ne survienne.

Les données complémentaires : d'autres données, dont la collecte n'est pas complètement automatisée ont un rôle prometteur [60, 78]. Quelques exemples sont recensés ici :

— Les données d'infrastructure : elles peuvent être fixes, comme par exemple les données d'architecture des voies, de positionnement des aiguilles, etc. mais également variables si on inclut les données des capteurs utilisés pour surveiller l'état de l'infrastructure, comme la température des rails.

— Relevés d'incidents : les données AVL permettent uniquement d'enregistrer les passages des trains et leurs retards. Elles peuvent ensuite être complétées manuellement par les agents pour former des archives des opérations. En particulier, les causes des retards supérieurs à un seuil donné, les causes des annulations, ou descriptions de la propagation des retards etc. peuvent être

manuellement renseignées [4]. L'identification des causes des retards et de leurs conséquences est capitale pour retracer le cours des évènements et identifier la responsabilité des différents acteurs.

— Planifications : le plan de transport est encore conçu et enregistré à la main, mais constitue une base de données riche, surtout si elle est couplée aux réalisations en opérationnel ou aux données d'infrastructure. Les données de planification contiennent à la fois les horaires, les plages travaux, les limitations de vitesse, etc.

— Historiques de maintenance : cette catégorie peut concerner aussi bien les archives de maintenances réalisées sur voies que sur le matériel roulant, les données d'incidents ou d'inspection. Dans tous les cas, ces données sont valorisables pour le suivi de la production ou pour des projets de maintenance prédictive.

— Comptages manuels et sondages : ils sont encore largement entrepris pour estimer l'affluence en gare et comprendre les habitudes des voyageurs. Les données APC et AFC tendent à faire disparaître ces pratiques : les relevés manuels sont coûteux à réaliser, ce qui limite leur fréquence.

2.3.2 Les différents cas d'usages

Wilson et al [176] ont proposé une classification des fonctions des données en 4 catégories : support pour la planification, la gestion opérationnelle du service, l'information voyageur ou encore l'évaluation de la performance. On présentera plus en détail les différents types de travaux et techniques utilisant des données pour améliorer les planifications dont la problématique est au coeur de ce sujet. Les autres fonctions seront introduites plus succinctement.

2.3.2.1 Construction des planifications ferroviaires

Les analyses hors-ligne des données ferroviaires permettent une meilleure description du fonctionnement opérationnel, l'identification de problèmes récurrents ou encore la calibration des durée de fonctionnement (temps de marche, temps d'occupation à quai,...). Ces analyses peuvent ensuite être utilisées pour améliorer les planifications. Que ce soit dans les processus industriels ou dans la démarche académique, l'analyse des données est en général disjointe de la prise en décision : les données sont étudiées à part afin d'en comprendre les caractéristiques, et les résultats pertinents sont exploités pour assister la production.

Robustesse aux retards : plusieurs exemples de planifications intégrant des résultats d'analyses de données réelles ou modélisées de retards ont été développés ces dernières années. Modéliser les retards permet de mieux identifier et pallier les faiblesses des solutions, et à terme de réduire la propagation des retards.

La plupart des études de ce type n'utilisent pas de données réelles mais des modèles simplifiés pour mesurer le risque de retard. Vromans estime des suppléments de temps optimaux en modélisant les retards par une loi exponentielle afin d'évaluer le retard moyen ou la ponctualité [171]. De même, Vansteenwegen et Van Oudheusden [169] ainsi que Sels et al [160] proposent de modéliser les probabilités des retards des trains pour inclure leurs conséquences dans la fonction objectif lors de l'optimisation des grilles horaires. Pour leur part, Dewilde et al [68] construisent un routage robuste itérativement en simulant des retards.

Simulation : les données peuvent servir à la construction de simulateurs stochastiques calibrés. De tels simulateurs sont par exemple utiles pour évaluer en amont la qualité d'un planning en générant des perturbations aléatoires. Il est important de calibrer un tel système pour assurer le réalisme et la fiabilité des résultats obtenus.

L'approche la plus simple est de modéliser la variable à simuler par une loi de distribution dont on estime les paramètres à partir des données réelles. C'est par exemple ce que font Larsen et al [118] en modélisant les temps d'occupation par une loi de Weibull paramétrée par maximum de vraisemblance, pour les heures de pointes et les heures creuses.

La seconde approche consiste à initialiser le modèle du générateur de perturbation, puis à itérativement comparer les résultats simulés avec les données observées. Koutsopoulos et al [111] proposent une méthodologie de calibration automatique d'un simulateur de retards basée sur la minimisation d'une fonction d'erreur entre les mesures observées et simulées. Büker [34] s'intéresse à la distribution des retards primaires puisqu'ils ne sont pas séparés des retards secondaires dans les données. La distribution est obtenue par de la simulation en comparant à la réalité. Cui et al [57] appliquent des techniques d'apprentissage par renforcement pour modéliser la probabilité de retards primaires aux endroits du réseau où il n'y a pas de relevés. Les paramètres sont mis à jour jusqu'à convergence

Estimation de la demande : des modèles se basent sur les données AFC et APC (systèmes de comptages des voyageurs et bornes de validation) et sur les données du réseau pour fournir une estimation de la demande. L'objectif principal est d'établir la *matrice OD* (matrice Origine/Destination) qui donne pour chaque paire de stations et par tranche horaire une estimation du volume de passagers attendus [6,176]. Cette matrice concerne principalement les trajets journaliers (réseaux denses comme les trains de banlieue ou des trains régionaux). D'autres analyses peuvent également être menées :

- comprendre la nature des trajets (professionnels, occasionnels, etc.)
- connaissance de l'origine et destination réelles (le point d'entrée et de sortie sur le réseau sont connus, mais l'usager utilise peut-être un autre mode de transport sur une partie du chemin, ce qui gagne à être connu)
- trouver le chemin parcouru entre l'origine et la destination, en particulier dans un réseau dense et à haute fréquence où plusieurs lignes sont accessibles et plusieurs trains d'une même ligne ont pu être empruntés.

Ces estimations de matrices OD sont utilisées pour évaluer les fréquences de trains nécessaires et concevoir les lignes desservies mais aussi pour mesurer les gains de différentes politiques pour les usagers. Elles peuvent également être utilisées pour pondérer les trains prioritaires d'un point de vue passager (par exemple un retard sur un train à forte affluence pénalise plus et peut donc être priorisé).

2.3.2.2 Autres cas d'usage

Gestion opérationnelle du service : dans ce cas, les données sont principalement exploitées en temps réel afin de délivrer toutes les informations nécessaires pour permettre le bon fonctionnement des opérations. De plus en plus de travaux se penchent sur l'intégration des flux de données en temps réel pour la replanification des opérations grâce à des outils d'aide à la décision.

Information voyageur : deux pistes principales de modèles prédictifs utilisant des données du transport ont été identifiées pour soutenir l'information voyageur :

- Prédiction de retards : l'objectif principal est d'actualiser en temps réel les retards estimés des trains pour un futur proche afin de permettre au mieux aux usagers de s'adapter. Des exemples hors ligne où les retards sont estimés en avance sont également envisageables, par exemple pour déconseiller des correspondances risquées. La prédiction de retards peut concerner à la fois des circulations sur grande ligne ou du trafic régional ou de banlieue.
- Prédiction d'affluence : l'idée est de donner en avance aux usagers des informations sur la congestion des moyens de transports qui les intéressent, en particulier en zone dense.

La précision et la fiabilité de l'information sont cruciales pour l'expérience des voyageurs qui peuvent alors mieux s'adapter et s'organiser : report du trajet sur un train à moins fort affluence, anticipation des conflits (correspondance ratée, etc.) et utilisation d'un trajet alternatif, etc [168].

Au delà de l'amélioration de l'expérience utilisateur, les acteurs ferroviaires gagnent également à transmettre plus d'informations. Par exemple, une forte affluence présente des risques de sécurité (quais surchargés, malaises voyageurs, etc.) et occasionne des temps excessifs d'occupation des quais qui créent et entretiennent les retards. Communiquer sur l'affluence peut permettre de partiellement lisser la charge, ce qui va limiter ces phénomènes.

Évaluation de la performance L'évaluation de la performance peut être menée en temps réel pour évaluer l'adéquation des mesures de gestion opérationnelle prises, mais elle est surtout utile après les opérations pour analyser et comprendre le déroulé des opérations.

— Reconnaître la responsabilité des acteurs en cas de perturbation
— Identifier les causes des perturbations [168]
— Mettre en place des indicateurs et critères adaptés qui peuvent aiguiller l'amélioration de l'offre

2.4 Données de retards : analyses et prédictions

Les données de retards sont une des bases les plus étudiées dans la littérature. Comme cela a été évoqué dans la partie précédente, l'analyse des performances opérationnelles a de nombreux intérêts, autant en horizon opérationnel que tactique et stratégique. Leur étude a cependant été assez désorganisée. Les cas d'études sont très variés : les retards peuvent différer selon le pays, le type de circulation ou l'infrastructure où ils sont observés, mais ce qui distingue le plus les études de retards est l'horizon temporel d'analyse. C'est surtout le cas pour des analyses prédictives, où les retards peuvent être modélisés en temps réel ou quelques mois en avance.

On propose de classer les différents travaux rencontrés dans ce domaine selon 5 catégories. Les 4 premières reprennent une séparation classique des analyses statistiques de données selon le type d'approche utilisée (paramétrique ou algorithmique) et selon l'objectif de l'étude (information ou prédiction) [31,162]. Les méthodologies informatives vont chercher à expliquer ou décrire les données, par exemple en identifiant leur structure sous-jacente ou des relations de causalité. Les études prédictives visent à construire un modèle qui pourra être appliqué à de nouvelles données pour en estimer la valeur [162]. Certaines méthodes sont communes, un même modèle peut à la fois être utilisé pour décrire une variable ou la prédire. Les approches paramétriques modélisent la variable d'intérêt par une fonction connue dont les paramètres sont estimés. Des régressions linéaires, estimation de densité par maximum de vraisemblance ou certains tests statistiques sont des approches paramétriques. Les méthodes non paramétriques, ou algorithmiques, ne font pas d'hypothèses sur la forme des données. Elles fonctionnent en boîte noire, ce qui les rend moins interprétable mais également plus flexibles [31]. Plusieurs algorithmes d'apprentissage supervisé appartiennent à cette catégorie, comme les forêts aléatoires ou les réseaux de neurones.

La dernière, qui ne suit pas la même logique, contient les estimations par modèles de propagation de retards explicites : les retards sont propagés de proche en proche dans le réseau.

2.4.1 Analyse informative

L'analyse informative des retards vise à identifier les propriétés des données grâce à des outils statistiques. Trois familles de problèmes sont étudiées dans ce contexte : l'étude de la distribution des retards, l'évaluation des facteurs impactant la ponctualité et l'identification de motifs récurrents entre les circulations retardées.

2.4.1.1 Distribution de probabilité des retards

Approche classique : l'approche la plus largement utilisée pour modéliser la distribution d'une variable aléatoire est d'utiliser une loi connue dont on estime les paramètres en fonction des données. On atteste ensuite de l'adéquation du modèle aux données, grâce à un test statistique tel que le test de Kolmogorov-Smirnov (KS). Cette méthodologie a été beaucoup utilisée, en particulier dans les premières études consacrées à l'étude des retards.

Goverde [81] et Yuan [183] ont chacun appliqué cette méthodologie respectivement sur les données de Eindhoven et de la Hague. Les deux études portent sur les retards par ligne à l'arrivée, au départ et au temps d'arrêt. Les lois candidates sont sélectionnées si elles coïncident avec les estimateurs par noyaux et les histogrammes des données puis testées avec le test KS. Goverde conclut en faveur de la loi exponentielle pour modéliser les retards positifs à l'arrivée et Yuan la loi Weibull.

Plus récemment, Wen et al [175] se sont intéressés aux distributions de retards primaires sur la ligne Wuhan-Guangzhou grâce à une base de 1249 enregistrements de retards primaires qui est étudiée selon la gare d'observation ou selon la plage horaire. L'adhérence des lois Weibull et lognormales est évaluée avec le test KS. Ces deux lois sont rejetées sur l'ensemble des données mais acceptées sur des sous-ensembles de données par heure ou par zone géographique.

Yang et al [181] étudient la distribution des retards conditionnellement à leur cause. Ils identifient 11 familles d'incidents dont ils étudient les caractéristiques et estiment la loi. Plusieurs distributions connues sont testées et comparées en fonction de leur vraisemblance. Pour chaque famille d'incidents, un test de Kolmogorov-Smirnov est appliqué pour vérifier l'adhérence de la meilleure loi candidate.

Cette méthodologie est cependant remise en cause par Harrod et al [94] car la statistique du test dépend directement de la taille de l'échantillon étudié. Cet article souligne que les études utilisant cette approche étaient réalisées sur des bases de données de tailles variables, mais toujours relativement réduites (ce qui est effectivement le cas dans les études citées précédemment), ce qui peut avoir tendance à fausser les résultats du test. Les bases de données massives formées par les systèmes actuels ne permettent plus d'utiliser le test de KS. Dans le cas de cet article, les lois normale, lognormale et exponentielle sont appliquées à large base de plus de 50000 retards au départs observés sur le réseau danois et sont toutes rejetées par le test, alors qu'elles sont acceptables sur des échantillons de taille réduite de ce même set. Ces expériences ne soutiennent pas les résultats de la littérature. En particulier, les auteurs contredisent la bonne adhérence de la loi exponentielle aux données de retards danoises en montrant que la propriété de perte de mémoire n'est pas observée sur les données réelles. Ils recommandent pour leur part l'utilisation de distributions mixtes.

Autre méthode : peu d'études ont cherché à connaître la distribution des retards par d'autres approches. Cui et al [57] estiment la probabilité de retard primaire non nul et sa valeur moyenne pour un simulateur à l'aide d'apprentissage par renforcement : à chaque étape, des retards sont simulés, propagés, puis comparés à la réalité pour ajuster les distributions. Cette approche permet donc l'estimation de la probabilité de retards primaires conditionnellement au point du réseau étudié, le type de train et le type d'évènement (arrivée, départ, etc.) en se basant sur les données de retards comprenant les retards secondaires et primaires relevés à d'autres points du réseau.

2.4.1.2 Description et explication des retards

De nombreux papiers traitent de la ponctualité et de sa relation aux différents éléments internes ou externes à la production ferroviaire. L'explication des retards, c'est-à-dire leur cause, est un problème partiellement résolu par les relevés d'opérations. Les systèmes de collecte automatique de données de circulations sont en général renseignés à la main dans le cas de retard important (cf 2.3.1). La cause du retard primaire est connue pour les grands retards, mais les créations de petits retards ne

sont pas claires quand elles sont liées à l'instabilité du système (temps de stationnement trop long, porte bloquée,...). Pour les retards secondaires, le mécanisme de propagation est complexe et n'est pas toujours connu, voire partiellement aléatoire, comme par exemple en raison de l'affluence en gare imprévisible face à une situation perturbée. Les études présentées dans cette sous-section tendent à décrire les retards à la fois en identifiant les causes et facteurs influençant l'apparition de retards, mais également en détectant des motifs cachés.

Facteurs corrélés : ce paragraphe recense quelques exemples d'études utilisant des approches statistiques pour décrire la relation entre les retards et d'autres variables identifiées comme potentiellement importantes. La méthodologie utilisée est très classique et repose en général sur des études de la corrélation entre les variables, des tests statistiques pour évaluer la significativité des différences observées et la modélisation des retards comme des régressions linéaires des autres paramètres. Un état de l'art plus complet peut être trouvé dans l'article de Palmqvist et al [145].

Goverde et al [83] proposent plusieurs analyses d'une base de retards mesurés à la station de Eindhoven, permettant de mettre en évidence la corrélation de certains éléments de contexte avec la ponctualité, comme par exemple la ligne et la direction, ainsi que le type de jours et le moment de la journée. La relation entre ces éléments et les retards n'est pas quantifiée, mais les auteurs montrent que la distribution s'en retrouve affectée, soit en comparant la valeur moyenne et le taux de retards entre les différents cas, soit en visualisant directement les différentes distributions conditionnellement à la valeur d'une variable.

Brazil et al [29] étudient l'impact des évènements météorologiques sur la ponctualité pour le métro de Dublin. Ils représentent les retards par une régression linéaire, dont ils profitent de l'interprétabilité des coefficients estimés pour quantifier la relation à la ponctualité de chaque variable.

Palmqvist et al [144] cherchent à identifier les variables ayant un impact sur la ponctualité de chaque train en étudiant les relevés d'opérations sur une voie combinés à la grille horaire, à la capacité et à des données météorologiques. La variable cible correspond au pourcentage du trajet d'un train où son retard est inférieur à 5 minutes. Un t-test de Welsh est utilisé pour évaluer la significativité de la relation, et les coefficients de la régression linéaire sont exploités pour comprendre la tendance de cette relation. Ils identifient notamment un impact important des températures extrêmes sur la ponctualité, mais également le nombre d'aiguillages traversés, les marges sur les temps de trajets et le nombre d'interactions entre les circulations en ligne et en gare. Ce genre d'étude permet à terme de proposer des stratégies pour s'adapter au retards.

Qin, Ma and Jiang [153] modélisent les retards en temps réel par des processus de Wiener pour étudier les facteurs impactant l'évolution des retards sur une ligne mixte. Ils montrent notamment que les relevés météorologiques, le type de circulation, le retard au départ et les informations sur les incidents ont un impact significatif sur les retards, impact qui peut être également quantifié en étudiant les coefficients dans les régressions estimant les paramètres du processus.

Zakeri et Olsson [185] se concentrent sur l'effet qu'ont les facteurs météorologiques sur le nombre moyen de trains ponctuels au départ par jour et par semaine (ponctualité) sur une ligne en Norvège. En plus d'étudier la corrélation à la variable cible, ils modélisent la ponctualité par une régression linéaire de ces variables, mettant en évidence un fort impact de la neige et des températures basses sur les retards.

Ces travaux confirment la présence de divers profils de retards. On peut séparer les facteurs identifiés par ces recherches en plusieurs familles : la route empruntée (ligne, arrêts, infrastructure, etc.), la grille horaire (marges allouées, temps d'arrêts, interaction avec d'autres trains), les conditions météorologiques et le contexte temporel.

Identification de motifs récurrents : l'objectif des études suivantes est d'identifier les mécanismes de propagation et d'occurrence de retards non détectés. En effet, si une partie des propagations de retards sont mécaniques, certains impacts n'apparaissent pas facilement, comme par exemple en cas de conflit de ressource humaine ou matérielle, de correspondance, etc.

Cule et al [58] adaptent le principe de règles d'association afin d'identifier des motifs dans la propagation des retards. L'idée est de chercher les retards apparaissant souvent simultanément. Les associations sont identifiées par plage horaire et les items sont les trains retardés.

Yabuki et al [178] appliquent également des règles d'association qui permet de moyenner les co-occurences de retard afin de vérifier si les mesures mises en oeuvre pour réduire les petits retards en zone dense ont été bénéfiques.

Cerreto et al [49] utilisent un algorithme de *k-means* pour identifier des profils de retards récurrents non détectés. La base de données étudiée concerne une ligne danoise où circulent trois types de trains, et chaque observation contient l'opération concernée, le retard observé, l'heure prévue, la desserte et le type de train. Seuls les retards de plus de trois minutes sont utilisés, les autres formant une classe isolée d'emblée par les auteurs. Les clusters formés par l'algorithme correspondent après à des profils de retards interprétables (où le train est ponctuel, où il acquiert son retard, comment le retard évolue).

2.4.2 Analyse prédictive

La mise en place de prédictions basées sur les données de retard a eu un succès important ces dernières années dans la mesure où les résultats permettent de répondre à de nombreux problèmes opérationnels. On différencie deux horizons temporels de prédiction distincts :

— L'horizon long-terme, correspondant aux prédictions stratégiques et tactiques. L'enjeu est d'étudier les retards du passé avec des données souvent macroscopiques pour ressortir des tendances pour un futur éloigné (quelques années à quelques jours). Les données disponibles sont susceptibles d'évoluer (les planifications, mouvements techniques, maintenance) et beaucoup d'informations pertinentes ne sont pas disponibles en avance, comme la météorologie ou les incidents sur le réseau. De telles approches sont intéressantes pour améliorer la robustesse des planifications ou mettre en place de nouvelles stratégies.

— L'horizon court-terme, l'objectif est d'étudier les comportements du passé pour prédire le jour même le retard aux prochains arrêts d'un train. Les données sont plus riches, étant donné que leur évolution est connue (suppressions de trains, incidents, retards courants, etc.). Les principaux cas d'usage des prédictions court-terme sont l'information voyageur et la replanification en temps réel.

2.4.2.1 Prédictions stratégiques et tactiques

Yaghini et al [179] utilisent un réseau de neurones pour prédire les retards sur le réseau Iranien. Ils utilisent comme variables explicatives les couples Origine-Destination, des variables temporelles (heure, jour, mois) et la ligne d'étude. Les retards sont séparés en 10 classes correspondant à la sortie du réseau. L'entraînement du modèle se fait sur une à plusieurs années consécutives afin de prédire l'année suivante.

Markovic et al [130] étudient la relation entre les retards à l'arrivée et l'infrastructure à long-terme, notamment pour pouvoir assister les prises de décisions concernant les changements sur le réseau. Les données prises en compte sont la catégorie du train, l'heure prévue d'arrivée, la description de l'infrastructure, le temps et la distance parcourues et les espacements. L'article met en évidence des corrélations fortes entre les retards et les variables étudiées, cependant l'analyse est portée sur une base de données de taille réduite (environ 700 trains) pour une infrastructure complexe et sans séparation temporelle entre l'entraînement et la validation.

Wang and Work [173] expérimentent des vecteurs à processus auto-régressifs, c'est-à-dire en exprimant le retard d'un train comme une régression linéaire des p observations précédentes de ce train. Les modèles sont entraînés sur deux ans, et testés pour chaque train sur les 30 prochains trajets de l'année suivante.

Wang and Zhang [172] prédisent des retards long-termes à l'aide d'un algorithme de gradient boosting en fonction des archives de retards, des horaires planifiés et de la météo. 75 jours sont utilisés pour construire les arbres et les 15 jours suivants pour la prédiction.

2.4.2.2 Prédictions opérationnelles

Peters et al [148] construisent deux systèmes de prédiction du retard pour du monitoring en temps réel des circulations sur le réseau allemand. Le premier est basé sur des règles explicites de transmission de retards entre les trains et le second utilise un réseau de neurones. Les retards sont catégorisés par type de train et par gare, et les prédictions sont comparées basées sur la MSE. Les auteurs concluent que les réseaux de neurones surpassent les modèles experts en identifiant des dépendances cachées.

Pongnumkul et al [151] étudient la moyenne glissante et l'algorithme de k plus proches voisins comme alternatives à la translation pour estimer le retard d'un train à ses prochains arrêts. Les modèles sont entraînés sur 6 mois et testés sur le mois suivant, puis comparés par MAE.

Wang and Work [173] utilisent des vecteurs autoregressifs pour estimer le retard d'un train en temps réel (ainsi qu'à long terme comme décrit précédemment). Les données utilisées ne sont plus des observations d'autres jours mais les retards réalisés aux arrêts précédents. Un terme est ajouté pour prendre en compte les retards des trains circulant en même temps.

Kecman et Goverde [109] estiment en temps réel les temps de marche et d'occupation à quai à l'aide d'historiques de données d'occupation des voies. Ils construisent trois modèles globaux, c'est-à-dire définis sur une partie du réseau néerlandais, utilisant des arbres de décisions, des forêts aléatoires et des régressions robustes. En complément, des modèles locaux sont également construits pour prédire les temps de marche sur des sections spécifiques et les temps d'occupation pour une gare spécifique. Ils montrent que les méthodes locales permettent d'obtenir de meilleurs résultats.

Oneto et al [143] utilisent des réseaux de Machine Learning Extreme pour résoudre ce problème sur le réseau Italien. L'algorithme est entraîné sur 6 mois sur les données de retards, avec comme features la description du point du réseau, du moment, du train et de ce qu'il s'est passé avant l'arrêt considéré. Un papier précédent des mêmes auteurs [142] a également appliqué des méthodes de Forêts aléatoires et de moindre carrés régularisés par noyaux, avec de bonnes performances des forêts aléatoires. Par ailleurs ils remarquent que l'utilisation de données météorologiques en complément des autres variables améliore les performances.

Deux travaux de prédiction en temps réel portent sur des cas d'études français. Chapuis [51] a appliqué des réseaux de neurones pour prédire en temps réel les retards sur le RER D en fonction des retards passés. Les réseaux sont entraînés sur des périodes de 10 jours glissantes et testés sur une journée après. Chandesris et Chapuis [50] prédisent les retards et le niveau d'affluence dans les trains à l'aide de méthodes non paramétriques établissant une prédiction à partir d'observations proches de la situation courante. L'expérience est concluante à la fois pour les TGV et pour les Transiliens.

Nair et al [138] proposent de combiner plusieurs modèles différents pour estimer les retards aux prochains arrêts. Ils utilisent notamment des régressions par noyaux, des forêts aléatoires et des méthodes de simulation du trafic. Combiner les méthodes donne de meilleures performance que les algorithmes pris seuls et permet de réduire la variance des prédictions. Leurs méthodes ne s'appliquent cependant pas aux trains au delà de 24h avant les opérations : leurs expérimentations ont montré qu'ils obtenaient des prédictions à peine meilleures que la grille horaire.

2.4.3 Approches par propagation

Une approche alternative calculant explicitement la propagation des retards a souvent été utilisée, mais ne fait pas l'objet de ces travaux. L'idée générale est de représenter les opérations ferroviaires par des graphes d'évènements (*time-event graph*) [81, 90] : les opérations consistent en une liste de processus associés à une durée, et dont le début et la fin sont marqués par des évènements (arrivées, départs, passages en un point). Le graphe est constitué de sommets représentant les évènements et d'arcs symbolisant les interdépendances entre ceux-ci. Ces arcs peuvent représenter un processus, comme une occupation de voie à quai ou la marche entre deux points, ou une relation de précédence entre des évènements liés à deux trains différents, par exemple via des contraintes d'espacements ou des correspondances. Les durées prévues et les marges intégrées doivent être paramétrées en amont. Hansen et al [90] utilisent par exemple des archives de données d'occupation des voies pour estimer avec précision les durées nécessaires pour réaliser les processus. Ces modèles sont utilisés à la fois en temps réel pour estimer les effets des retards primaires au fur et à mesure qu'ils apparaissent, et ainsi permettre une replanification efficace, mais aussi pour évaluer la qualité des grilles proposées.

La propagation des retards au sein du graphe peut ensuite se faire de manière déterministe en injectant des retards primaires sur certains sommets et en étudiant les conséquences sur les noeuds suivants en fonction des durées de chaque processus et des marges disponibles qui permettent d'absorber le retard. Les algèbres max-plus sont très adaptées pour calculer la propagation sur ces graphes, comme l'a montré Goverde dans plusieurs de ses recherches [22, 81, 82]. En effet, les contraintes de propagation de retards s'écrivent de manière linéaire dans de telles algèbres.

Corman et Kecman ont également proposé une version stochastique de ce graphe, vu comme un réseau bayésien [54], permettant de modéliser la distribution de valeurs autour des retards prédits. Les retards sur les évènements sont actualisés en temps réels, permettant au modèle d'estimer la probabilité de retard des noeuds successeurs dans le réseau.

De tels modèles représentent efficacement les interdépendances entre les circulations mais nécessitent de décrire précisément les relations entre les trains, notamment l'ordre des circulations et des évènements. Ces relations sont cependant susceptibles d'évoluer en fonction des décisions prises pour réguler le trafic.

2.5 Positionnement

L'objectif de ces travaux est d'aider à augmenter la résistance aux perturbations des affectations de voies grâce à une analyse approfondie des historiques de retards de trains. Le positionnement de cette thèse par rapport aux axes de littérature ferroviaire est explicité ci-dessous.

La recherche opérationnelle s'est imposée comme une collection d'outils efficaces pour permettre d'appréhender la complexité de la production ferroviaire et permettre de concevoir automatiquement des planifications en prenant en compte toutes sortes de contraintes et objectifs. Les problèmes d'affectations de voies en gare n'ont pas fait exception et ont également été bien couverts dans la littérature avec plusieurs modélisations proposées ainsi que des algorithmes de résolution exacts et approchés.

Quels que soient les problèmes étudiés, les solutions proposées sont mises à mal en conditions réelles en raison des perturbations, et en particulier des retards. Il existe plusieurs types de retards, et leur fréquence et leur amplitude varient beaucoup, allant de quelques secondes à plusieurs heures. Les petits retards sont courants et doivent pouvoir être absorbés au mieux par le plan de transport à l'aide de marges sur les temps de trajet ou durée d'occupation des voies, et la propagation doit être freinée en agençant les circulations stratégiquement afin d'éviter d'impacter des trains ponctuels. Les retards plus importants (de la dizaine de minutes à plusieurs heures) sont bien plus rares et ne peuvent pas être absorbés par les marges sans occasionner de baisse de niveau de service, mais ils sont

pris en charge par des modifications du plan de transport en opérationnel afin de permettre un retour à la normale avec un impact minimal.

Plusieurs stratégies d'optimisation de la robustesse et de consommation de capacité ont donc été développées pour anticiper ces perturbations et permettre un déroulement fluide des opérations. On retiendra notamment comme définition de la robustesse la capacité d'une solution à limiter la propagation des retards en raison de conflits de ressources (croisements d'itinéraires, occupations de voies, etc.). Cette définition manque cependant de clarté et sa mesure effective peut prendre différentes formes. Les mesures les plus réalistes quantifient la robustesse collective en évaluant l'impact des retards sur le trafic en général, mais sont bien plus complexes et lourdes à modéliser. À l'inverse des mesures individuelles se concentrent sur l'effet qu'a chaque train sur les autres en cas de perturbation, indépendamment des aléas sur les autres circulations. Elles sont plus simples à mettre en place mais ne représentent que partiellement les effets de trafic. On choisira ici de se concentrer en premier lieu sur les mesures individuelles qui s'intègrent mieux dans des modules d'optimisation pour l'affectation de voies.

Quel que soit le problème d'étude, la recherche de robustesse est encore largement déterministe, avec une optimisation basée sur les caractéristiques de la solution (espacements, pourcentage de capacité consommée, etc.). De plus en plus de travaux modélisent cependant les retards pour apporter une information stochastique sur les perturbations, par exemple par de la simulation stochastique pour évaluer la robustesse ou par la construction de fonctions objectif intégrant le risque de retard. Cependant ces modélisations utilisées pour de la prise de décision sont encore très souvent trop simplificatrices, par exemple en utilisant des distributions exponentielles pour représenter les retards, et souvent sans prise en compte de certains indicateurs affectant la distribution, comme la ligne ou l'heure. Les historiques de données sont très rarement utilisés, en général uniquement pour paramétrer les lois utilisées. La motivation de ces travaux est d'apporter une perspective plus réaliste et complète à la recherche de robustesse en intégrant directement dans le mécanisme d'optimisation des considérations basées sur des perturbations effectivement rencontrées dans le passé. En particulier, on cherche à être en mesure d'évaluer le risque associé aux partages de ressources, et ainsi pouvoir identifier correctement les conflits potentiels en fonction des trains impliqués et de la marge disponible.

En se concentrant sur les articles dédiés à l'étude des retards par des méthodes d'analyse et d'apprentissage statistique, on constate que les données de retards contiennent des motifs redondants qui peuvent être exploités pour mieux comprendre le déroulement des opérations, voire anticiper des conflits prévisibles. En particulier, de nombreuses recherches ont identifié une relation entre les retards et les conditions météorologiques, la mission planifiée, les éléments d'infrastructure empruntés ou encore les circulations en cours au même moment. De plus en plus de chercheurs et industriels utilisent ces observations pour construire des modèles de prédiction de retards conditionnellement à ces éléments de contexte. La revue de littérature donnée ici montre que les cas d'études sont très variés, que ce soit par rapport au périmètre d'étude (gare, ligne, réseau), l'horizon temporel d'apprentissage et de prédiction ou encore les données utilisées pour les variables cibles et explicatives. Il y a tout de même une forte prévalence de modèles de prédiction à court-terme pour lesquels la prédiction porte sur le retard aux prochains arrêts connaissant le retard actuel et l'état du trafic.

Dans le cadre de cette thèse, on cherche à anticiper les retards les plus probables pour proposer une adaptation robuste des affectations de voies. Pour que cette adaptation ait lieu dans de bonnes conditions, elle doit se situer avant la veille des opérations, ce qui place l'analyse des retards dans un cadre long-terme. Contrairement à ce qui est fait dans la plupart des études de référence, on ne cherche pas à estimer la valeur précise du retard, ce qui ne paraît par ailleurs pas réaliste compte tenu de la forte variance des retards, de l'importante ponctualité des trains et du peu d'information disponible au moment de la prédiction. La direction adoptée dans ces travaux est plutôt de construire des modèles d'estimation de la probabilité associée à chaque valeur de retard, afin de pourvoir quantifier les risques

de conflit dans certains enchaînements de trains sur des ressources en gare.

La représentation par des distributions de probabilités des trains n'est pas nouvelle puisque de nombreuses études ont cherché une loi adéquate à l'aide de tests statistiques. Ces distributions sont cependant encore très largement paramétriques, construites sur des bases de données de taille souvent réduite et elles n'exploitent pas la variété de profils de retards existants. Les retards sont encore très largement modélisés par une loi négative exponentielle bien qu'ils n'en respectent pas certaines propriétés fondamentales comme la relation entre l'espérance et la variance, ou les propriétés d'absence de mémoire.

On montrera par ces travaux qu'on peut gagner en précision et en réalisme en estimant la distribution de retards des trains conditionnellement à leur profil (mission, temporalité, etc.) et avec des distributions plus complexes. Deux approches sont étudiées, une avec des modèles paramétriques généralisant les premières études sur les distributions de retards, et une autre non paramétrique basée sur l'algorithme de forêt aléatoire dont les bonnes performances ont été observées sur les problèmes de prédiction de retards en temps réel.

Les distributions de retards estimées sont ensuite explicitement utilisées dans un algorithme de recherche locale dont le but est de diminuer les conflits en gare en fonction du risque évalué pour chaque paire de trains occupant successivement des ressources incompatibles. La résolution utilise plusieurs heuristiques et méta-heuristiques qui optimisent à la fois la faisabilité de la solution en minimisant le nombre de minutes de retard créées par des situations non réalisables, mais également la robustesse en réduisant le risque de propagation des retards grâce à une fonction d'évaluation des risques de conflit entre chaque paire de circulations. Plusieurs contraintes industrielles peu traitées dans la littérature sont abordées ici, comme les niveaux de priorité variables entre les trains, l'intégration de situations non faisables sans variables fictives ou les opérations de coupe et accroche de rames à quai.

C'est à notre connaissance le premier exemple de méthodologie *data-driven* appliquée au problème d'affectation de voies en gare, et de manière général un des rares cas d'analyse prescriptive utilisant un système de prédictions sur données réelles pour l'aide à la décision ferroviaire.

Première partie

Estimation de risque de retards

Chapitre 3

Approches paramétriques et non-paramétriques pour l'estimation de probabilités conditionnelles

L'objectif général de ces travaux est d'intégrer le risque de retard des trains dans les affectations de voies en gare pour en améliorer la robustesse. L'approche adoptée pour cela, et qui sera explicitée sur les données réelles dans le chapitre 4, consiste à analyser les observations de retards de trains et leur relation à plusieurs facteurs (date, origine, densité du trafic, etc.) puis à estimer la distribution de probabilité de retards des trains conditionnellement à ces facteurs. Ces distributions estimées peuvent ensuite être utilisées pour adapter les planifications afin de minimiser les risques de conflits de circulation liés aux retards. Les observations de retards de trains utilisées dans ces travaux sont mesurées en minutes et tronquées. La variable aléatoire correspondante est donc discrète, numérique et bornée.

Ce chapitre expose différentes méthodes statistiques et d'apprentissage qui ont été identifiées pour permettre de répondre à ce problème. Le type de données supporté par chacune de ces méthodes et les généralisations possibles sont spécifiés à chaque étape.

Sauf spécifications contraires, on utilisera les conventions suivantes. Des lettres majuscules sont utilisées pour désigner les variables du modèle : les variables explicatives (ou *features*) sont notées X, avec p composantes $X_1, ..., X_p$ et la variable aléatoire cible est notée Y. Les éventuelles données ou valeurs observées sont notées par des lettres minuscules, comme pour un échantillon d'observations $\mathbf{y} = (y_1, ..., y_n)$. Les vecteurs sont symbolisés par des lettres en gras minuscules et les matrices par des lettres majuscules en gras, par exemple $\mathbf{X} = (\mathbf{x_1}, ..., \mathbf{x_p})$.

3.0.1 Distribution d'une variable cible

3.0.1.1 Définition

Les algorithmes classiques d'apprentissage supervisé reposent en général sur la prédiction d'une valeur ponctuelle, catégorielle pour une classification supervisée, ou réelle dans le cas d'une régression. Ce chapitre traite du cas particulier des estimations de distributions de probabilités individuelles. Elles consistent à décrire la distribution de probabilité de la variable de réponse conditionnellement aux valeurs d'un ensemble de variables explicatives ou *features*.

Deux approches sont considérées pour la prédiction probabiliste, une paramétrique et une autre non-paramétrique. Les méthodes paramétriques font l'hypothèse de la forme de la fonction f qui décrit les données, et en estiment les paramètres. Les approches non-paramétriques ne font pas d'hypothèse de départ, elles peuvent donner une meilleure adhérence aux données, mais fournissent des estimations moins lisses et qui fonctionnent le plus souvent en boîte noire, ce qui les rend moins interprétables [31].

Données continues : soit Y une variable aléatoire continue à valeurs dans $\mathbb{R}$ et $X_1, ..., X_p$ des variables explicatives. L'estimation de la fonction de densité de probabilité de Y conditionnellement à $X_1, ..., X_p$ consiste en l'estimation de la fonction $f : u, \boldsymbol{X} \to \mathbb{R}$ telle que :

$$\mathbb{P}\left[a < Y < b | X_1, ..X_p\right] = \int_a^b f(u, X_1, .., X_p)du \quad \text{pour tout} \quad a < b$$

Données de comptage : le cas particulier des données de comptage sera utilisé ici à plusieurs reprises. Ces données quantitatives discrètes sont à valeurs dans $\mathbb{N}$ et caractérisent des variables aléatoires issues du comptage. Les retards de trains peuvent par exemple être vus comme des comptes de minutes. L'estimation de probabilités consiste alors à déterminer la fonction de masse f_Y en chaque valeur $y \in \mathbb{N}$: $f_Y(y, X_1, ..., X_p) = \mathbb{P}\left[Y = y | X_1, ..X_p\right]$.

Données qualitatives : Soit $Y \in C_1, ...C_K$ une variable à K modalités, on cherche à déterminer la fonction de masse f pour tout k : $f_Y(C_k, X_1, ..., X_p) = \mathbb{P}\left[Y = C_k | X_1, ..X_p\right]$.

3.0.1.2 Applications

La plupart des travaux existants utilisent des méthodes paramétriques pour le domaine du biomédical, avec en particulier des régressions logistiques et des modèles de survie (Modèles de Cox). Un autre domaine où l'estimation de fonction de densité est très présent est celui de la météorologie, où des variables comme la température ou les précipitations suivent des distributions continues ou mixtes qu'on cherche à représenter. Les modèles linéaires généralisés sont également utilisés pour modéliser des phénomènes ayant des distributions spécifiques, comme avec des régressions de Poisson ou des régressions exponentielles.

Le cas des variables binaires est de loin le plus traité, et a servi de cadre de travail pour le développement des méthodes utilisant des algorithmes non paramétriques de Machine Learning pour l'estimation de probabilités, appelées parfois *probability machines* [127]. Là encore, les premiers travaux qui formalisent ces approches proviennent principalement de la communauté de statistiques médicales.

La prédiction de probabilités permet de quantifier l'incertitude liée à la valeur prédite. Cet aspect est particulièrement pertinent quand l'estimateur de probabilité sert de support pour prendre des décisions, comme par exemple pour informer les patients de leur pronostic ou pour répartir des ressources limitées aux sujets avec les meilleures chances. L'estimateur de probabilité doit à la fois renvoyer des distributions de probabilité fiables, qui rendent bien compte des risques réels, on parle alors de calibration, mais il doit également être discriminant, c'est-à-dire être en mesure de classer les sujets selon leur risque [164].

Une revue des différentes stratégies pour l'estimation de distributions de probabilité individuelles est proposée en section 3.1. La section 3.2 énonce ensuite les difficultés liées à l'évaluation de ces modèles et détaille les différentes métriques et tests de validation, avec des scores de performance globale, des tests de calibration et des mesures de discrimination.

3.1 Estimation de probabilités individuelles

Il existe deux grandes familles de modèles statistiques : les modèles paramétriques, qui représentent la variable de réponse comme une fonction connue des variables explicatives et dont on cherche à estimer les paramètres, et les modèles algorithmiques ou non-paramétriques, qui fonctionnent en boîte noire et ne font pas d'hypothèse sur la forme des données [31].

Cette section présente une approche de chaque type pour estimer la distribution de probabilité d'une variable conditionnellement à des variables explicatives. Le cas des modèles paramétriques est introduit en premier, avec des modèles linéaires généralisés et leurs différentes extensions, puis les forêts aléatoires sont étudiés dans un second temps.

3.1.1 Cas paramétrique : les modèles linéaires généralisés

Une régression linéaire modélise une variable aléatoire Y comme une combinaison linéaire de une ou plusieurs autres variables dites explicatives et d'une variable aléatoire normalement distribuée :

$$y_i = \beta_0 + \beta_1 x_{i1} + ...\beta_p x_{ip} + \epsilon_i \text{ avec } \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{3.1}$$

pour $i = 1..n, k = 1..p$, avec y_i une observation de la variable Y et les x_{ik} variables explicatives. Les termes d'erreur ϵ_i sont supposés indépendants et distribués selon une loi normale centrée de variance σ^2. De manière équivalente :

$$\begin{array}{l} Y \sim \mathcal{N}(\mu, \sigma^2) \\ \mu = \beta_0 + \sum_{i=1}^{p} \beta_i X_i \end{array} \tag{3.2}$$

Les hypothèses faites sur les données sont fortes, mais peuvent être partiellement contournées en utilisant des modèles linéaires généralisés (GLM). Ces modèles, introduits en 1972 par Nelder et Wedderburn [139] permettent de représenter une variable de réponse ne suivant pas une loi normale, étant définie sur un domaine de valeurs restreint ou encore ayant une variance variable. Ils s'articulent autour de trois parties :

- une composante aléatoire : les observations $y_1, ..., y_n$ dont on suppose qu'elles sont issues d'une variable aléatoire Y de densité f appartenant à la famille exponentielle.
- un modèle linéaire : $\boldsymbol{\eta} = \boldsymbol{X\beta}$ avec $\boldsymbol{X}$ la matrice des variables explicatives et $\boldsymbol{\beta}$ le vecteur des paramètres
- une fonction de lien g qui relie le modèle linéaire aux paramètres de la loi de Y : $g(\mu) = \eta$ avec μ le paramètre de la fonction f.

Un modèle linéaire généralisé peut donc s'écrire :

$$\left\{ \begin{array}{ll} Y & \sim \text{Dist}(\mu) \\ g(\mu) & = \boldsymbol{X\beta} \end{array} \right. \tag{3.3}$$

Ces modèles permettent donc d'estimer grâce à une régression linéaire le paramètre de loi de chaque observation en fonction de ses variables explicatives. Le vecteur des paramètres de régression $\boldsymbol{\beta}$ est déterminé par maximum de vraisemblance.

Un modèle linéaire généralisé peut s'adapter à des variables de réponses binaires, réelles positives ou encore discrètes. Un résumé des modèles les plus courants est présenté dans le tableau 3.1. Plus d'informations concernant ces modèles, le choix de la fonction de lien ou l'expression des paramètres peuvent être trouvés dans l'article de Nelder et Wedderburn [139]. En particulier, Nelder et Wedderburn donnent un cadre théorique générique à ces modèles, quel que soit le support de la variable de réponse, avec une forme prédéfinie de la fonction de densité.

Les GLM représentent la distribution globale de la variable d'intérêt, ce qui permet à la fois de prédire sa valeur de manière ponctuelle, généralement en renvoyant sa moyenne μ, mais également d'avoir accès à l'ensemble des statistiques de la loi, comme l'expression des quantiles ou de la mode.

Distribution	Domaine	Fonction de lien	Espérance	Variance
Normale	$\mathbb{R}$	$g(\mu_i) = \mu_i$	μ_i	σ constante
Poisson	$\mathbb{N}$	$g(\mu_i) = log(\mu_i)$	μ_i	$\sigma_i = \mu_i$
Binomiale	$\{0, 1\}$	$g(\mu_i) = log(\frac{\mu_i}{1-\mu_i})$	μ_i	$\sigma_i = \mu_i(1 - \mu_i)$
Exponentielle	$\mathbb{R}^+$	$g(\mu_i) = -1/\mu_i$	μ_i	$\sigma_i = \mu_i^2$

TABLEAU 3.1 – Cadre classique des GLM

3.1.1.1 Extension GAMLSS : modélisation des paramètres de localisation, forme et échelle

Une extension des modèles linéaires généralisés a été proposée par Rigby et Stasinopoulos [156,163] avec la formulation GAMLSS (*Generalized Additive Models for Location, Scale and Shape*) et a été implémentée via la librairie R *GAMLSS*. Cette extension est moins restrictive que les GLM de Nelder et Wedderburn : la forme générique de la loi de probabilité n'est plus contrainte, permettant d'utiliser une grande variété de distributions de référence. Par ailleurs, les modèles GAMLSS permettent de modéliser simultanément plusieurs paramètres des distributions comme fonctions des variables explicatives.

Soit Y la variable de réponse suivant la loi de probabilité $\mathcal{P}$ de paramètres μ, σ, ν, τ, un modèle GAMLSS peut s'écrire :

$$\begin{aligned} Y &\backsim \mathcal{P}(\mu, \sigma, \nu, \tau) \\ g_\mu(\mu) &= \eta_\mu = X_\mu \beta_\mu \\ g_\sigma(\sigma) &= \eta_\sigma = X_\sigma \beta_\sigma \\ g_\nu(\nu) &= \eta_\nu = X_\nu \beta_\nu \\ g_\tau(\tau) &= \eta_\tau = X_\tau \beta_\tau \end{aligned} \tag{3.4}$$

Les coefficients $\beta_i, i \in \{\mu, \sigma, \nu, \tau\}$ sont estimés par maximum de vraisemblance. Ils sont optimisés un à un en considérant les autres fixés jusqu'à convergence de la vraisemblance.

Les paramètres μ et σ représentent en général les paramètres de localisation et d'échelle. Ils ne correspondent pas nécessairement à l'espérance et la variance comme dans le cadre classique des GLM où seule la moyenne de la variable est modélisée. ν et τ sont des paramètres de forme, et mesurent l'asymétrie (*skewness*) et l'allure des queues de distribution (*kurtosis*).

Dans l'approche GAMLSS, aucune forme générique de la distribution n'est imposée et la fonction de lien est établie en fonction du domaine du paramètre : une fonction logarithme est par exemple utilisée pour un paramètre strictement positif, une fonction logit s'il est à valeurs dans $[0, 1]$ ou la fonction identité s'il n'a pas de domaine restreint. La liste des distributions prises en charge, ainsi que les propositions de modélisations associées sont décrites plus en détail dans la documentation de la librairie GAMLSS et autres références proposées par les auteurs [154, 155, 163].

3.1.1.2 Variantes classiques

Les données réelles n'obéissent cependant pas toujours aux hypothèses de modélisation liées aux distributions. En particulier, Hilbe [96] identifie plusieurs types de violation, dont les cas de réponse tronquée, de censure ou de zéro excessifs. Cette partie présente les différentes stratégies et modèles utilisés pour pallier ces difficultés.

Prise en charge des données nulles : dans de nombreuses situations les données réelles présentent une quantité excessive de données nulles. C'est en particulier le cas quand la valeur nulle signifie une

absence d'observation (quantité de pluie journalière, nombre de visites chez le médecin,...). Cela peut avoir plusieurs conséquences [135] :

— une présence excessive de zéros est souvent signe de surdispersion qui peut mener à des échecs d'adhérence des lois. Ce risque est récurrent pour des modèles classiques comme ceux présentés dans le tableau 3.1 où la variance s'exprime explicitement en fonction de l'espérance.

— un certain nombre de distributions de référence réelles positives ne sont pas définies en zéro (par exemple la loi Weibull).

— les distributions réelles ne sont ni discrètes ni continues en raison de cette masse en zéro. On parle de distributions mixtes ou semi-continues.

Une première approche utilisée notamment pour pallier l'impossibilité de modéliser des valeurs nulles par des distributions continues est de bruiter légèrement ces données pour obtenir des quantités infinitésimales mais strictement positives ou de remplacer les valeurs nulles par une quantité constante. La substitution de valeurs a été souvent utilisée, en particulier en chimie où les valeurs nulles correspondent aux observations inférieures au seuil de détection. Plusieurs recherches ont cependant montré que cette approche fausse les estimations des différentes statistiques en ajoutant des motifs qui n'existent pas dans les données, ou en camouflant les motifs existants [95].

Une alternative est d'utiliser des modèles à inflation de zéros (*zero-inflated models)* qui sont formés par le mélange d'un modèle linéaire généralisé et d'un modèle binomial [12]. Cette approche est dédiée dans la littérature aux problèmes de comptage. Soit D une loi de probabilité à valeurs dans $\mathbb{N}$ de paramètre μ, un modèle à inflation de zéro s'écrit :

$$Y \sim \begin{cases} 0 & \text{avec probabilité } \pi \\ \mathbb{P}_D[\mu] & \text{avec probabilité } 1-\pi \end{cases} \tag{3.5}$$

Les coefficients du modèle permettant d'estimer μ et π sont calculés conjointement, et la distribution de probabilité de Y prend la forme suivante :

$$\mathbb{P}[y_k \leq t|\boldsymbol{x_k}] = \pi_k + (1-\pi_k)\mathbb{P}_D[y_k \leq t|\boldsymbol{x_k}] \tag{3.6}$$

Une autre alternative est d'utiliser des modèles en deux parties, ou modèle de Hurdle dans le cas de données de comptage. Une première partie binaire sépare les données nulles des données non nulles, et une seconde modélise les valeurs strictement positives [12, 135]. Soit $\boldsymbol{y} = (y_1, .., y_n)$ le vecteur des observations et$\boldsymbol{X}$ la matrice des covariés. Alors pour $t \geq 0$:

$$\mathbb{P}[y_k \leq t|\boldsymbol{x_k}] = \pi_k + (1-\pi_k)\mathbb{P}_{D_{tr}}[y_k \leq t|y_k > 0, \boldsymbol{x_k}] \tag{3.7}$$

Le premier modèle, classiquement une régression logistique, estime $\pi_k = \mathbb{P}[y_k = 0|\boldsymbol{x_k}]$ et $\mathbb{P}_{D_{tr}}[y_k \leq t|y_k > 0, \boldsymbol{x_k}]$ est estimée par le second modèle pour chaque y_k et chaque $t > 0$ à l'aide d'une distribution D_{tr} tronquée en zéro. Cette technique est particulièrement recommandée quand il y a une différence structurelle entre les valeurs nulles et les autres ou dans le cas où les données non nulles suivent une distribution continue.

Dans le cadre de données de comptage, les modèles en deux parties sont proches des modèles à inflation en zéro, cependant l'expression de la vraisemblance diffère puisque dans un modèle à inflation, la masse en zéro est à la fois due au modèle de comptage et au modèle binaire, alors que dans un modèle en deux parties, la masse nulle n'est liée qu'au premier modèle, le second étant tronqué en zéro. Cette nuance rend les modèles à inflation en zéro pertinents pour modéliser des réponses où une partie de la population est toujours nulle, et une autre l'est parfois, comme par exemple le nombre de participations à une activité [135]. Il y a également une différence dans l'optimisation des deux approches puisqu'un

modèle en deux parties va entraîner indépendemment les deux modèles, dont le second sur un sous-ensemble des données alors que les modèles à inflation de zéros estiment simultanément les paramètres des deux sous-modèles.

L'analyse de survie : les données de survie représentent la durée attendue avant l'occurrence d'un évènement (*time to event*) [122]. La variable modélisée est donc une variable aléatoire continue non nulle. Ces modèles sont particulièrement utilisés en statistiques médicales pour étudier la mortalité d'une population, les facteurs de risques, etc. Un aspect important de ces données est la présence d'observations censurées, c'est-à-dire des observations qui ont survécu plus longtemps que le temps de l'étude, et dont la valeur exacte est de fait inconnue. Ces données contiennent une information partielle qui doit être prise en compte même si la valeur cible n'est pas connue.

Ces modèles sont intéressants ici car ils visent à modéliser la fonction de survie S qui est égale à $1 - F(t)$ avec F la fonction de répartition. La fonction S se modélise par :

$$S(t) = exp^{-\int_0^t h(u)du} \tag{3.8}$$

avec h le taux de survie dont l'expression est déterminée par les hypothèses du modèle. Dans le cas d'une modélisation paramétrique où les données sont modélisées par une fonction connue comme la loi exponentielle ou Weibull, le modèle de survie sans censure est équivalent à un modèle linéaire généralisé.

Troncature : l'échantillon $x_1, ..., x_n$ est tronqué si chaque observation est inférieure à une quantité connue T_d dans le cas d'une troncature à droite, ou supérieure à T_g pour une troncature à gauche. La troncature est fréquente, notamment à gauche dans le cas discret quand des valeurs nulles ne sont pas observées ou modélisées. La troncature de l'échantillon nécessite de normaliser les fonctions de densité (resp. de masse) utilisées pour modéliser les données afin de pouvoir intégrer à 1 (resp. sommer à 1) malgré une probabilité nulle au delà du seuil de troncature [53].

Un GLM entraîné avec une distribution tronquée perd en interprétabilité. Par exemple pour des GLM classiques, la moyenne est généralement égale au paramètre de localisation comme pour une régression exponentielle, cependant après normalisation la moyenne prendra une autre forme [14].

L'utilisation d'une loi normalisée est indispensable sans quoi le modèle est faux. En effet, le GLM prédira des probabilités non nulles aux données au delà du seuil de troncature malgré l'impossibilité de les observer. Il y aura donc nécessairement un défaut de calibration des probabilités prédites.

3.1.2 Cas non-paramétrique et Machine Learning

L'estimation de distributions de probabilités à l'aide de méthodes non paramétriques ou algorithmiques est plus récente et moins développée que les approches par modèles paramétriques présentées précédemment. Plusieurs adaptations d'algorithmes de Machine Learning ont tout de même été proposées pour permettre de renvoyer des distributions de probabilités et non plus des valeurs ponctuelles. Cette section décrit l'utilisation de forêts aléatoires. On rappelle d'abord le fonctionnement général des algorithmes d'arbres et de forêts, puis on détaillera leur utilisation pour estimer des probabilités dans le cas de variables binaires et nominales.

3.1.2.1 Arbres de probabilité et Forêts Aléatoires

Les arbres de décision : un arbre de décision binaire est un outil qui permet d'obtenir une partition de l'espace des observations en séparant récursivement la base d'apprentissage par rapport à une valeur seuil d'une des variables explicatives. Un exemple d'une telle partition pour deux variables

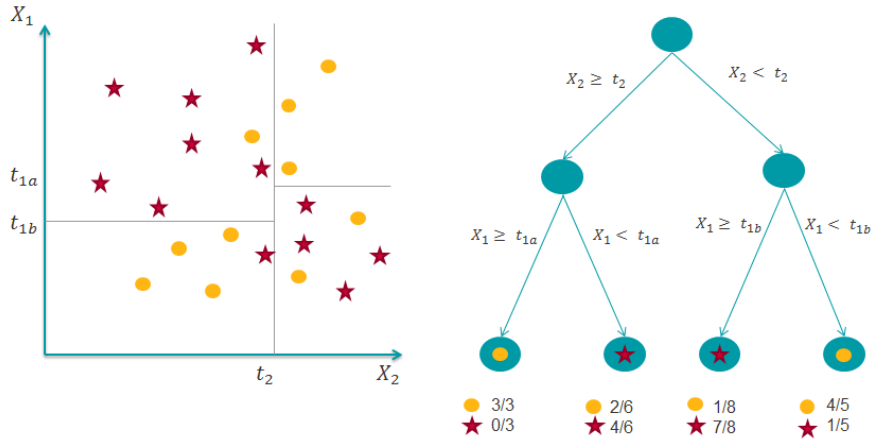

FIGURE 3.1 – Schéma de partition par un arbre de décision binaire

explicatives est présenté dans la figure 3.1 : l'espace des données est divisé en plusieurs régions en fonction des valeurs des features. La séparation va chercher à optimiser un critère défini, comme par exemple l'homogénéité des noeuds formés. Une valeur est associée à chaque région, et ces valeurs servent de prédictions pour de nouveaux individus.

Soit un ensemble de n observations $y_i, i = 1...n$ décrites par p variables explicatives $(x_{i1}, ..., x_{ip})$. À chaque étape, chacune des variables explicatives est étudiée afin de trouver le point de séparation qui permet la meilleure amélioration du critère de séparation. La variable et le seuil donnant la meilleure amélioration sont retenus et utilisés pour générer deux nouveaux noeuds, sur lesquels la même approche est appliquée. Un arbre maximal est obtenu quand les noeuds sont homogènes, c'est-à-dire que toutes les observations du noeuds ont la même valeur.

Ces arbres sont connus pour être assez peu robustes aux variations dans les données [74]. Il est recommandé d'en limiter la complexité pour éviter des phénomènes de sur-apprentissage, par exemple en imposant un critère d'arrêt dans la construction de l'arbre (profondeur maximale ou taille minimale des noeuds) ou en construisant un arbre de taille maximale, puis en élaguant des branches.

Ces arbres peuvent être utilisés pour des prédictions de variables qualitatives ou quantitatives. Les seules modifications dans l'algorithme concernent le critère de séparation utilisé et le choix de la valeur associée à chaque région [74]. Pour une régression, les séparations se feront habituellement en minimisant une fonction de coût comme la somme des moindres carrés, et la valeur contenue par chaque noeud est la moyenne des valeurs de ses individus. Pour de la classification, la valeur associée à chaque noeud terminal est la valeur du *vote majoritaire*, c'est-à-dire la valeur $k(m) = \text{argmax}_k \hat{p}_{m,k}$ avec $\hat{p}_{m,k}$ la proportion de représentants de la classe k dans le noeud m. Les critères de séparation classiques sont l'erreur de classification $e = 1 - \hat{p}_{m,k(m)}$, l'indice de Gini $e = \sum_{k=1}^{K} \hat{p}_{m,k}(1 - \hat{p}_{m,k})$ ou l'entropie $e = -\sum_{k=1}^{K} \hat{p}_{m,k} ln(\hat{p}_{m,k})$.

Forêts aléatoires : cet algorithme repose sur l'idée *bagging* ou *bootstrap aggregating* qui vise à réduire la variance d'un modèle en l'entraînant sur plusieurs échantillons tirés aléatoirement de la base de données d'origine et en moyennant les prédictions obtenues. Ce principe est utilisé pour les forêts aléatoires afin de moyenner les résultats de plusieurs arbres de décision. Ces arbres sont randomisés à la fois par bagging et par une sélection aléatoire des variables utilisées. Les forêts aléatoires ont été formellement introduites par Leo Breiman en 2001 pour résoudre des problèmes de classification et de

régression [30]. Pour B arbres, l'algorithme suit les étapes suivantes :

1. Pour $b = 1...B$:
   (a) Extraire avec remise un sous ensemble d'observations de taille N de la base d'origine
   (b) Construire un arbre t_b en répétant les étapes suivantes jusqu'à ce que le critère d'arrêt soit atteint :
      i. Sélectionner aléatoirement m variables explicatives parmi les p variables disponibles
      ii. Trouver la variable et le point de partition maximisant le critère de séparation
      iii. Générer deux noeuds fils
2. Renvoyer en sortie les T arbres

Pour prédire la valeur d'un nouvel individu, la forêt agrège les informations fournies par les feuilles des arbres. Pour une classification, la classe prédite sera classiquement la valeur prédominante parmi les noeuds terminaux (vote majoritaire), quand une forêt de régression renverra la valeur moyenne de ces noeuds.

La randomisation par sélection aléatoire de features et par échantillonnage des données d'apprentissage est très importante car elle permet de réduire fortement la variance de la forêt en réduisant la corrélation entre les arbres, sans impacter le biais [74]. En effet, les arbres obtenus par bagging sont identiquement distribués mais non indépendants. Pour B variables aléatoires indépendantes et identiquement distribuées de variance σ^2, la variance est égale à σ^2/B. Si ces variables ne sont pas indépendantes et ont entre elles un coefficient de corrélation ρ alors la variance de la moyenne est égale à $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$. Ainsi augmenter le nombre d'arbres fera baisser le second terme sans réduire le premier. L'erreur tend à converger, une valeur trop grande de B n'améliorera pas la qualité des prédictions mais générera du temps de calcul inutile.

Le nombre m de variables considérées pour séparer un noeud peut varier de 1 à p, mais en général on conseille d'utiliser $m = \sqrt{p}$ pour une classification et $m = p/3$ pour une régression [74, 77]. Le critère d'arrêt permet de limiter la complexité de l'arbre, mais n'est pas nécessaire. Divers paramètres peuvent être utilisés : profondeur maximale, nombre d'éléments minimal par feuille, taille minimale d'un noeud avant séparation, etc. Breiman recommande pour sa part de ne pas élaguer les branches comme pour des arbres CART, étant donné que l'utilisation de bagging et de sélection aléatoire de features pour construire la forêt permet déjà de compenser la forte variance et le manque de robustesse des arbres de décisions utilisés seuls [30].

3.1.2.2 Probabilités d'appartenance de classe

Les approches par arbres de décision et par forêts aléatoires permettent de travailler à la fois sur des données quantitatives continues et des données qualitatives nominales. L'utilisation des arbres pour prédire des probabilités se réduit cependant aux données qualitatives (binaires ou multicatégorielles) où on cherche à quantifier l'incertitude liée à la prédiction de classe.

Arbres de probabilité : les arbres d'estimation de probabilités (*PET, probability estimation tree*) ont été initialement introduits pour résoudre le problème d'évaluation des probabilités d'appartenance à des classes. Cette estimation apporte une information plus riche que la simple classification supervisée, par exemple pour ordonner les classes selon leur probabilité d'occurrence [152].

Les noeuds terminaux ne contiennent plus une valeur mais une distribution de probabilité. Classiquement, la probabilité d'appartenance à la classe C_k pour l'élément y de covariés $\mathbf{x}$ dans l'arbre T est estimée par la fréquence relative de C_k dans $T_{\boldsymbol{x}}$, le noeud terminal de y [129].

$$\mathbb{P}[y = C_k|\boldsymbol{x}] = \frac{N_k(T_{\boldsymbol{x}})}{N(T_{\boldsymbol{x}})} \tag{3.9}$$

avec $k \in 1, ...K$, $N_k(T_{\boldsymbol{x}})$ le nombre d'éléments de la base d'apprentissage appartenant à la classe C_k et étant dans la feuille $T_{\boldsymbol{x}}$, et $N(T_{\boldsymbol{x}})$ le nombre total d'éléments dans la feuille.

Cependant comme pour les arbres de décision, les prédictions obtenues sont souvent de mauvaise qualité. Plusieurs raisons à cela : les séparations successives forment des échantillons de plus en plus réduits, la plupart des critères de séparation se focalisent plus sur les erreurs de prédiction ponctuelle que sur la fiabilité des prédictions et la structure des arbres fait que les distributions produites sont les mêmes au sein d'une région donnée, sans prendre en compte la position des observations par rapport aux frontières de la région [129]. Ces estimations peuvent être améliorées de deux manières. La première approche est de corriger les probabilités prédites en utilisant une autre formule que 3.9, ce qui permet en particulier de lisser les distributions dans le cas de petits échantillons de données [27]. L'autre idée est d'utiliser du bagging.

Forêts : plusieurs études ont montré que l'utilisation de bagging ou de forêts permet d'améliorer la qualité des probabilités prédites [27, 152]. L'algorithme de forêt est le même que présenté plus haut : pour chacun B arbres, un échantillonage des données d'apprentissage est effectué, un arbre est construit dessus en utilisant un critère de classification supervisé (type cross-entropie) et pour chaque arbre T_b et chaque élément y de covariés $\boldsymbol{x}$ on mesure la fréquence relative de chaque classe k dans $T_{b,\boldsymbol{x}}$ le noeud terminal de y :

$$p_{b,k}(y, \boldsymbol{x}) = \frac{N_k(T_{b,\boldsymbol{x}})}{N(T_{b,\boldsymbol{x}})} \tag{3.10}$$

Les probabilités de chaque classe sont estimées en moyennant les fréquences d'occurrence obtenues dans chaque arbre [147] :

$$\mathbb{P}[y = C_k|\mathbf{x}] = \frac{1}{B}\sum_{b=1}^{B} p_{b,k}(y, \boldsymbol{x}) \tag{3.11}$$

Ce cadre d'utilisation des forêts aléatoires n'a été formalisé que plus récemment par Malley et al en 2012 [127] puis Kruppa et al en 2014 [114, 115] pour des prédictions binaires à l'aide d'arbres de classification, mais également d'arbres de régression car dans le cas binaire la moyenne et la probabilité d'observer 1 sont égales. La distinction entre les arbres de régression et de classification se fait uniquement sur le critère de séparation utilisé. Le terme de machine de probabilités (*probability machines*) a également été introduit dans ces travaux.

Comme dans le cas général, les forêts permettent de réduire la variance des prédictions et d'en augmenter la précision. En effet, on constate que les arguments en défaveurs des arbres PET sont palliés par l'utilisation de forêts : le bagging permet d'avoir à disposition un grand nombre d'échantillons différents, et l'utilisation d'observations et de variables différentes a un impact sur les frontières des régions qui varient selon les arbres.

3.1.2.3 Approches alternatives

Plusieurs autres algorithmes de Machine Learning ont été proposés pour répondre à ce problème d'estimation de probabilités conditionnelles mais ne seront pas traités plus en détail :

— l'algorithme de k-plus proches voisins permet également de s'adapter à la prédiction de probabilités en renvoyant pour chaque élément la distribution formée par les valeurs de ses voisins les plus proches. Dans la version ensembliste proposée par Malley et al [127] et reprise par Kruppa et al [114], les distributions sont moyennées sur plusieurs itérations de l'algorithme utilisant des bases de données tirées aléatoirement à chaque fois (bagging).

— une adaptation des machines à support de vecteurs (SVM) pour prédire des probabilités d'appartenance de classe a posteriori à été proposée par Platt [150] en appliquant une fonction sigmoïde aux prédictions : pour une prédiction binaire, $\mathbb{P}[y = 1|\boldsymbol{x}] = 1/(1 + exp(Af(\boldsymbol{x}) + B)$ où A et B sont estimés par maximum de vraisemblance.

— pour les méthodes de boosting, Niculescu-Mizil et Caruana [140] rappellent les formules classiques utilisées pour estimer les probabilités d'évènements dichotomiques en fonction de la sortie de l'algorithme Adaboost. Comme pour les SVM, la sortie de l'algorithme de Boosting peut être passée dans une fonction sigmoïde pour obtenir des probabilités

— les réseaux de neurones peut également prédire des probabilités en utilisant par exemple la fonction d'activation softmax [74, 147]. Pour K classes et $f_k(X)$ la sortie de la dernière couche, les probabilités estimées sont $\mathbb{P}[y = 1|\boldsymbol{x}] = \frac{exp(f_k(X))}{\sum_{j=1}^{K} exp(f_j(X))}$. Cette transformation est la même que celle utilisée pour des régression logistiques multinomiales.

On constate qu'il existe deux familles différentes : d'un côté les algorithmes de forêts aléatoires et de k-plus proches voisins qui génèrent un échantillonnage des données permettant d'estimer des probabilités grâce à la fréquence de chaque classe au sein des groupes d'observations formés (soit les k plus proches soit dans la même feuille), et les algorithmes de Boosting, réseaux de neurones et SVM qui appliquent une transformation au score de sortie pour construire une distribution de probabilité.

3.1.2.4 Autres supports de données

Cet algorithme a été conçu pour estimer des probabilités de classes disjointes. On présente ici les pistes de généralisations à des données de classes ordonnées et de données qualitatives continues.

Variables ordinales : ce cas particulier de la multiclassification correspond à des données pour lesquelles il existe un ordre entre les classes. Des stratégies pour traiter ce cas sont exposées ci-dessous [86] :

— classification ordinale : l'ordre entre les classes n'est pas pris en compte et le problème est traité comme une multiclassification classique, ce qui peut engendrer une perte de performance étant donné que l'information liée à l'ordre n'est pas exploitée. On peut compenser partiellement ces pertes en ajustant la fonction de poids pour pénaliser moins fortement les erreurs de classification quand elles concernent des classes d'ordres voisins.

— régression ordinale : afin de prendre en compte l'ordre des données, on peut les modéliser par une méthode de régression en numérotant les classes. Cette approche a le défaut d'ignorer la structure des données qui ne sont pas nécessairement numériques. Dans le cas où les données sont qualitatives, la distance entre les classes n'est pas connue ou n'existe pas, et ne sera donc pas respectée par la fonction d'erreur utilisée. Pour la prédiction de probabilités avec des arbres, la méthode s'adapte très bien à la régression en utilisant comme critère une fonction d'erreur adéquate, comme par exemple l'erreur quadratique. Une fois les arbres construits, la prédiction reste la distribution au sein de noeuds terminaux.

— décomposition dichotomique : une approche alternative proposée par Frank et Hall [73] construit $k - 1$ classifieurs binaires pouvant estimer des probabilités et dont le but est d'identifier la frontière entre les classes C_i et C_{i+1} pour $i = 1..k - 1$, et de calculer ainsi la fonction de répartition $F(i) = \mathbb{P}[C_y \leq C_i]$ en chaque point $i = 1..k - 1$. La classe prédite est celle dont la probabilité estimée est la plus grande, avec $\mathbb{P}[y \in C_i] = F(i) - F(i - 1)$. Un défaut de cette méthode est qu'elle peut conduire à des probabilités prédites négatives comme chaque modèle est entraîné indépendemment des autres. Cardoso et Pinto da Costa [44] proposent une variante de la méthode de Frank et al qui assure la validité des distributions de probabilités obtenues

en construisant des classifieurs binaires estimant la probabilité $p_i = \mathbb{P}[C_y > C_i | C_y > C_{i-1}]$. Un inconvénient de cette variante est que les classifieurs sont entraînés sur des sous-ensembles de données puisque les premières classes sont isolées à chaque étape.

Données qualitatives : l'extension à des variables continues est peu traitée. Frank et Bouckaert [72] proposent d'estimer des fonctions de densité conditionnelles en discrétisant la variable en différentes classes. Les probabilités d'appartenance de chaque classe sont calculées et utilisées pour estimer la fonction de densité à l'aide d'un estimateur de noyau par exemple.

Une autre option à explorer serait d'entraîner les arbres de probabilités comme des arbres de régression, ce qui ne nécessite pas d'avoir un support de valeur connu à l'avance. La distribution de probabilité obtenue à la sortie des noeuds terminaux devra ensuite être lissée pour permettre d'estimer des distributions de probabilités individuelles (par exemple avec un estimateur de noyau). Cette approche peut également être considérée pour des données de comptage où une approche multicatégorielle n'est pas envisageable étant donné que le support est infini.

3.1.3 Machine Learning ou modèles statistiques ?

Quelques articles ont étudié les performances des deux familles de modèles et ont conclu qu'aucune méthode n'était systématiquement meilleure et que chacune présente ses spécificités [28, 56, 114]. Les GLM et forêts aléatoires sont comparés ci-dessous selon différents critères.

Support des observations : le point de distinction le plus important est le support de la variable cible. Les GLM permettent de représenter une grande variété de données, à savoir dans le cas général binaires, continues ou de comptage, mais également des données tronquées ou censurées. Les observations doivent cependant pouvoir être modélisées par une loi de probabilité de référence ce qui requiert une forme lisse. Les forêts aléatoires peuvent être naturellement adaptées pour prédire des probabilités d'appartenance à des classes, mais l'algorithme ne peut pas être utilisé tel quel sur des données continues ou discrètes non bornées.

Interactions entre les variables : un GLM représente les paramètres de la loi choisie comme une régression linéaire des variables explicatives (en intégrant la fonction de lien). Les interactions détectées sont donc linéaires. Les forêts aléatoires se basent sur des séparations successives. Elles sont connues pour capturer plus efficacement les interactions non linéaires entre la cible et les features mais sont de fait moins efficaces quand la relation est linéaire.

Encodage des variables explicatives : comme dit précédemment, un GLM va exploiter des relations linéaires entre les prédicteurs et les paramètres de la loi. La relation doit donc être parfaitement définie pour que le modèle soit exact, et cela nécessite parfois de modifier l'encodage des variables, leur échelle ou encore de leur appliquer une fonction, par exemple carré ou inverse, si la relation n'est pas linéaire. Au contraire, les forêts aléatoires supportent différents supports de données et peuvent prendre en charge l'encodage automatiquement. Par exemple si les différentes modalités d'une variable sont numérotées, les arbres pourront les isoler automatiquement et de manière optimale. Les arbres sont également plus robustes aux valeurs extrêmes. Celles-ci ne risquent pas de faire dévier les résultats autant que dans un GLM.

Performances : les méthodes par maximum de vraisemblance ont tendance à être mieux calibrées, alors que les arbres vont produire des estimations plus précises, avec des bons classements par risque

[152]. Le choix de la méthode peut donc se fonder sur l'utilisation des résultats et les attributs souhaités des prédictions.

Interprétabilité : les deux approches présentées permettent d'interpréter partiellement l'impact des variables utilisées sur la variable cible. Comme les GLM reposent sur la modélisation du paramètre de loi par une fonction d'une régression linéaire, chaque coefficient estimé contient une information par sa valeur et son signe sur l'impact de la variable correspondante. Cependant plus le modèle est complexe (distributions à plusieurs paramètres, troncature, etc.) ou plus les variables explicatives sont corrélées, moins les effets sont interprétables directement.

Les arbres de décisions sont directement interprétables puisqu'ils construisent des règles de séparation simples pour prédire des futures valeurs. Les variables utilisées sont celles qui séparent au mieux et donc qui ont la contribution la plus importante. Cependant cette interprétabilité se perd en agrégeant les résultats dans une forêt. Il n'est plus possible d'appréhender facilement chaque séparation, d'autant plus que les variables sont choisies aléatoirement. Il est tout de même possible de mesurer de différentes manières l'importance de chaque variable dans la construction de la forêt. Par exemple on peut permuter aléatoirement les valeurs des variables utilisées et mesurer l'erreur induite dans l'arbre [56, 77], ou encore mesurer à chaque séparation l'amélioration que la variable apporte au critère de séparation [74] puis moyenner sur la forêt. Les variables ayant un mauvais score sont peu utilisées, soit car elles ne permettent pas de séparer efficacement, soit car elles ne peuvent pas être utilisées plusieurs fois (cas des variables binaires par exemple). L'interprétation reste difficile du fait de l'agrégation de l'information.

Sélection de features : la sélection de features consiste à rechercher un sous-ensemble de variables explicatives pour entraîner le modèle afin de le simplifier sans perdre en information. Cette étape est plus importante pour les modèles linéaires où l'ajout d'une variable supplémentaire même inutile permet toujours d'améliorer la vraisemblance et donc d'augmenter le risque de surapprentissage. Des variables additionnelles inutiles peuvent nuire à l'interprétabilité du modèle.

La sélection de variables est beaucoup moins importante pour les algorithmes d'arbres car elle est implicitement intégrée au processus de construction des arbres. Une variable inutile ne sera pas utilisée pour séparer les données, et des variables redondantes auront un score d'importance similaire ce qui n'influence pas l'interprétabilité des résultats. Par ailleurs les arbres de décision sont également efficaces dans le cas où le nombre de variables est très grand par rapport au nombre d'observations, ce qui n'est pas le cas des modèles linéaires. A l'inverse, les forêts donnent en général de moins bonnes performances pour un très faible nombre de variables [56, 77].

Hyperparamétrage : au delà de la sélection de features et du choix de la distribution, les modèles linéaires généralisés ne nécessitent pas d'ajustement supplémentaire lors de la construction du modèle. L'implémentation est alors facile, mais le risque que le modèle soit faux ou mal spécifié est plus important [28]. Au contraire, les algorithmes de Machine Learning nécessitent un hyperparamétrage conséquent mais les hypothèses sur la forme des données sont moins restrictives.

Transportabilité : cet argument avancé par Boulesteix et Schmid [28] est qu'un modèle paramétrique permet de définir une formule relativement simple et reproducible qui est réutilisable même en dehors de l'étude. La prédiction pour de nouvelles valeurs ne génère pas de temps de calcul important. Pour les forêts aléatoires ce n'est pas vrai : pour reproduire un modèle il faudrait disposer de la description de chaque arbre, avec les variables et les points de séparations associés à chaque noeud, ainsi que la valeur ou distribution contenue dans les noeuds terminaux.

3.2 Sélection et validation des machines probabilistes

La sélection de modèle et l'évaluation de la qualité des prédictions posent plusieurs difficultés quand la réponse est exprimée sous forme de probabilités individuelles car les méthodes utilisées dans des domaines connexes (statistiques classiques, cadre général de régression et de classification supervisées) deviennent inadéquates.

Dans le cas simplifié de l'estimation de densité de probabilité d'une variable aléatoire sans prise en compte de variables explicatives, il existe plusieurs tests permettant d'évaluer l'adhérence du modèle à un échantillon de données. Un des plus utilisés est le test de Kolmogorov-Smirnov qui mesure la distance entre les répartitions empirique et théorique des données, et dont on teste la significativité. Ces méthodes ne sont plus applicables pour le cas de l'estimation de probabilités conditionnellement à des variables. En effet, il n'existe parfois qu'une seule réalisation de chaque distribution prédite, ce qui ne permet pas de produire une distribution empirique à comparer avec la distribution estimée.

Contrairement à la plupart des approches d'apprentissage supervisé, la valeur de la variable de réponse et la prédiction ne sont pas homogènes, avec d'un côté une valeur binaire, catégorielle ou numérique et de l'autre une fonction de densité ou de masse. Cela demande d'utiliser des méthodes de validation spécifiques, puisque les comparaisons directes entre observations et prédictions ne sont plus possibles ni pertinentes.

Traditionnellement, et en particulier dans le cas binaire, trois approches sont utilisées pour évaluer la qualité des prédictions sous forme de probabilités [93, 164] :

— Les fonctions de scores : elles donnent une évaluation globale des modèles, et sont particulièrement utiles pour la sélection de modèle.

— La mesure de la discrimination : elle caractérise la propension du modèle à classer correctement les éléments en fonction de leur risque effectif.

— L'attestation de la calibration : c'est l'adéquation des probabilités prédites aux taux observés.

3.2.1 Mesures de la performance globale

Une fonction de score permet de quantifier la performance d'un modèle en comparant les probabilités prédites et la valeur observée. Ces fonctions de scores jouent plusieurs rôles : elles peuvent servir de fonction objectif lors de l'entraînement du modèle, mesurer la qualité des prédictions ou encore classer des modèles en compétition et sélectionner le meilleur candidat.

Soit Y la variable aléatoire étudiée à valeur dans Ω, de distribution $P_Y \in \mathcal{P}$ avec $\mathcal{P}$ un ensemble de distributions de probabilité. Soit $S : \mathcal{P} \times \Omega \to \mathbb{R}$ une fonction de score. Pour y une observation de Y et $\bar{P} \in \mathcal{P}$ une distribution candidate, on notera $S(\bar{P}, y)$ le score correspondant à l'évaluation de $\bar{P}$ pour y et $S(\bar{P}, Y)$ l'espérance mathématique du score. Le score S à maximiser est dit strictement propre si $S(P_Y, Y) \geq S(\bar{P}, Y)$ avec égalité si et seulement si $\bar{P} = P_Y$.

En pratique, pour n observations y_i associées aux probabilités prédites P_i, une machine de probabilité ou un modèle noté M est évalué grâce au score moyen $S(M) = \sum_i S(P_i, y_i)/n$.

Trois fonctions de score strictement propres sont particulièrement présentes dans la littérature et sont introduites dans la suite de cette section : les scores logarithmiques, les scores sphériques et les scores quadratiques. Un état de l'art plus approfondi de certains autres scores strictement propres peut être trouvé dans le rapport de Gneiting et Raftery [79]. Des formules génériques paramétrées décrivant des familles de scores propres sont données par Merkle et Steyvers [133]. Les preuves que ces scores sont strictement propres sont rappelées par Bröcker et Smith [33].

On étudiera chacun des scores avant tout dans un contexte de classification multicatégorielle, en rappelant à chaque fois les spécificités des cas particuliers à deux catégories et à catégories ordonnées,

mais aussi dans le cas d'une réponse continue. Pour chacun de ces cas et pour la durée de cette section, on utilisera les notations suivantes :

— cas catégoriel : soit $\boldsymbol{y} = (y_1, ..., y_n)$ un vecteur d'observations à valeurs dans $\{1, ...K\}$ symbolisant K classes mutuellement exclusives. Pour une observation y_i, on note $\boldsymbol{p_i} = (p_{i,1}, ...p_{i,K})$ les probabilités d'appartenance aux différentes classes estimées par le modèle M, dont p_{i,y_i} la probabilité de la classe observée, et $\delta_{i,j}$ variable valant 1 si y_i est dans la classe j, 0 sinon.

— cas numérique : soit $\boldsymbol{y} = (y_1, ..., y_n)$ un échantillon d'observations à valeurs réelles. Pour une observation y_i, f_i la fonction densité de probabilité estimée par le modèle M (ou fonction de masse pour des données de comptage).

3.2.1.1 Score logarithmique

Le score logarithmique d'une observation, quel que soit le support des valeurs, correspond au logarithme de la vraisemblance de la valeur observée. Ainsi dans le cadre catégoriel, $S_{log}(\boldsymbol{p}, y_i) = log(p_{i,y_i})$. Ce score s'adapte naturellement au cas continu avec $S_{log}(f_i, y_i) = log(f_i(y_i)))$. Ce score a la propriété d'être local, c'est-à-dire qu'il ne dépend que de la probabilité de la valeur observée.

Le score logarithmique de M aggrégé pour les observations $\boldsymbol{y}$ est aussi appelé log-vraisemblance ou cross-entropie :

$$l(M, \boldsymbol{y}) = \sum_{i=1}^{n} log(p_{y_i}) \tag{3.12}$$

ou dans le cas continu :

$$l(M, \boldsymbol{y}) = \sum_{i=1}^{n} log(f_i(y_i|x_i)) \tag{3.13}$$

Ce score est équivalent à la déviance globale $D = -2l(M, \boldsymbol{y})$ utilisée dans l'approche GAMLSS [163]. L'estimation des paramètres par maximum de vraisemblance dans les modèles linéaires généralisés et l'extension GAMLSS garantit que le modèle estimé est optimal d'un point de vue de la déviance ou de l'entropie. Le score logarithmique peut également être utilisé comme critère de séparation dans des algorithmes d'arbres de décision [74, 147].

3.2.1.2 Score quadratique, score de Brier et variantes

Cas catégoriel : Le score quadratique est plus traditionnellement défini pour des données catégorielles, avec :

$$S_q(\boldsymbol{p}, y_i) = 2p_{i,y_i} - \sum_{k=1}^{K} p_{i,k}^2 \tag{3.14}$$

Ce score n'est local que dans le cas d'une réponse dichotomique, puisque pour plus de deux catégories, il dépend également des probabilités des évènements non observés par le terme $\sum_{k=1}^{K} p_{i,k}^2$. Dans ce cas, il est possible qu'un premier modèle affectant une probabilité plus petite qu'un second à la valeur réelle obtienne malgré tout un meilleur score.

Pour évaluer les prédictions d'un modèle M sur un échantillon, on utilise plus généralement le score de Brier qui est un indice introduit par Brier en 1950 mesurant l'erreur quadratique d'une estimation de probabilités catégorielles [32].

$$\begin{aligned} BS(M, \boldsymbol{y}) =& \frac{1}{Kn} \sum_{i=1}^{n} \sum_{k=1}^{K} (p_{i,k} - \delta_{i,k})^2 \\ =& 1 - \frac{1}{Kn} \sum_{i=1}^{n} S_q(\boldsymbol{p_i}, y_i) \end{aligned} \tag{3.15}$$

Le score de Brier se généralise aux sorties discrètes ordonnées par le score des probabilités ordonnées noté RPS (*ranked probability score*) pour la prédiction d'une variable scalaire y [39, 137]. Comme le score de Brier, cet indicateur est basé sur les résidus quadratiques entre la fonction de probabilités cumulées estimée et la fonction de répartition empirique calculés en différents points. Soient L valeurs seuil ξ_l et les événements $\epsilon_l = \{y \geq \xi_l\}$. On note $o_{il} = 1$ l'évènement $y_i \geq \xi_l$ et p_{il} sa probabilité estimée. Le score RPS se calcule par :

$$\begin{aligned} RPS =& \frac{1}{nL} \sum_{i=1}^{n} \sum_{l=1}^{L} (o_{il} - p_{il})^2 \\ =& \frac{1}{2L} \sum_{l=1}^{L} BS(\epsilon_l) \end{aligned} \tag{3.16}$$

S'il a été initialement conçu pour des variables ordinales, il se généralise naturellement aux données de comptages.

Dans le cas où les probabilités estimées ne peuvent prendre qu'un nombre fini de valeurs, le score de Brier, et par extension le score des probabilités ordonnées, se partitionnent en deux composantes, la fiabilité, analogue à la calibration, et la résolution, proche de la discrimination [136]. Cette décomposition n'est pas étudiée plus en détail dans ces travaux dans la mesure où les GLM et les forêts aléatoires estiment des distributions individuelles, c'est-à-dire qu'il y a potentiellement autant de distributions différentes que d'éléments dans l'échantillon.

Cas continu : le *continuous ranked probability score* CRPS peut être utilisé pour mesurer la distance entre une densité de probabilité et une observation de cette loi [79] :

$$CRPS(F, y) = -\int_{-\infty}^{\infty} (F(u) - \mathbb{1}[u \geq y])^2 du \tag{3.17}$$

3.2.1.3 Le score sphérique :

Ce score est moins utilisé et présent dans la littérature, tant pour optimiser les modèles que pour les évaluer. Il se définit de la manière suivante :

$$S(\boldsymbol{p_i}, y_i) = \frac{p_{y_i}}{||p||} = \frac{p_{y_i}}{\sqrt{p_{i,1}^2 + ... + p_{i,K}^2}} \tag{3.18}$$

3.2.1.4 Comparaison et utilisation de ces scores

Ces trois règles de score sont très utilisées, en particulier les scores logarithmiques et quadratiques, cependant il y a peu d'éléments permettant de guider le choix d'un des critères plutôt qu'un autre :

— tous ces scores s'adaptent aux différents supports de données (binaires, catégorielles, continues). Le RPS est cependant à notre connaissance le seul score spécifiquement dédié aux variables ordinales, ce qui évite de traiter les données comme appartenant à des classes indépendantes.

— ces scores sont particulièrement utilisés pour classer des estimateurs de probabilités et sélectionner le meilleur modèle candidat. Ces trois scores produisent généralement des classements fortement corrélés, mais ils peuvent beaucoup varier localement [23, 133]. Machete a étudié plus en détail d'où venaient les différences de classements [126]. Il prouve que dans le cas binaire, le score logarithmique pénalise la surestimation de l'issue la plus probable plus que sa sous-estimation (pour une déviation égale) alors que le score sphérique la favorise. Le score quadratique pour sa part ne pénalise que par rapport à la valeur absolue de la déviation.

— le score logarithmique est le seul à être local, ce qui signifie qu'il ne dépend que de la probabilité de l'évènement observé. Il va ainsi tendre à préférer des estimateurs maximisant la probabilité des observations indépendamment des probabilités des évènements non observés. Les scores quadratiques et sphériques ne sont pas locaux, car ils intègrent la somme des carrés des probabilités de chaque évènement. Ce terme supplémentaire peut engendrer des cas d'égalité, c'est-à-dire deux estimateurs associant la même probabilité à l'évènement observé mais recevant un score différent. Il est aussi possible que deux estimateurs aient le même score alors qu'un des deux avait assigné une probabilité supérieure à l'évènement observé. Un avantage de la localité est cependant qu'elle privilégie des distributions lisses, même s'il existe des valeurs non observées mais possibles [33].

— il peut être intéressant d'évaluer les performances du modèle en se basant sur un autre critère que celui qui a été utilisé pour l'entraîner. En particulier, un certain nombre de modèles sont basés sur l'optimisation de la vraisemblance ou l'entropie, qui sont équivalentes à l'optimisation du score logarithmique. Il est parfois conseillé de varier les outils d'analyse des résultats [59].

— certaines valeurs extrêmes ou absurdes peuvent avoir un effet plus fort sur les scores. Par exemple, une valeur absurde de très faible densité peut pénaliser très fortement la vraisemblance du modèle et fausser les scores logarithmiques. Dans le cas de distributions avec asymptote en zéro, des variations faibles des observations peuvent également occasionner des variations importantes du score. Le RPS peut être faussé dans le cas de données de comptage à valeurs très élevées : quelques observations extrêmes vont dominer le reste des observations et biaiser le score moyen [59].

3.2.2 Analyse de la calibration

3.2.2.1 Les différentes niveaux de calibration

La calibration se définit comme l'adéquation des taux d'évènements observés aux taux de risque prédits. En pratique sa définition est plus complexe, avec plusieurs niveaux de calibration [93, 167] :

— calibration en espérance (*calibration in the large*) : le taux d'évènements observés dans les données doit être égal au risque estimé moyen

— calibration faible : on vérifie qu'il n'y a pas de surestimation ou sous-estimation systématique des risques. Pour cela on étudie l'intercept de calibration, par exemple avec un test de recalibration de Cox où on doit avoir une pente de un et un intercept nul.

— calibration modérée : la calibration est testée au sein de groupes de probabilités prédites similaires. Les tests d'Hosmer-Lemeshow et les graphes de calibration, qui sont présentés ci-dessous, rentrent par exemple dans ce cadre là.

— calibration forte : les risques prédits doivent être bons pour chaque combinaison de valeurs des variables explicatives, ce qui est impossible dans le cas où au moins une de ces variables est continue.

Van Calster et al [167] considèrent la calibration forte comme un objectif irréaliste, car elle repose sur un modèle parfaitement spécifié. Au contraire, la calibration faible est insuffisante. Pour des prédictions qui seront utilisées pour prendre des décisions, une calibration modérée est souhaitée.

3.2.2.2 Les graphes de calibration

Cette approche consiste à agréger les observations en groupes pour permettre de comparer graphiquement la moyenne des probabilités prédites au sein du groupe avec le taux d'évènements effectivement observés. Les sujets sont triés par probabilités d'évènement croissantes, séparés en g groupes soit par valeur fixes soit par quantiles, et un point est ajouté au graphe pour chaque groupe. On considère qu'un modèle est calibré si les points sont proches de la diagonale. Plusieurs diagnostics peuvent être menés à partir de ces graphes :

— Des points sous la diagonale caractérisent une surestimation de la probabilité estimée, alors que des points au-dessus de la diagonale sont une sous-estimation. Il n'est pas rare d'observer un biais systématique si tous les points ou une grande partie se concentrent d'un côté de la diagonale, malgré des écarts faibles. Une autre forme de biais courante est une allure du graphe de calibration en sigmoïde : cela signifie que le modèle évite d'estimer des probabilités extrêmes, et aura tendance à surestimer les probabilités faibles et sous-estimer les probabilités élevées.

— Il est également possible de visualiser la discrimination en étudiant un graphe de calibration. En effet, plus un modèle est discriminant, plus les prédictions qu'il produit permettent d'ordonner correctement les individus en fonction de ce qui est observé. En pratique, cela se caractérise par des graphiques où les points sont alignés et peu dispersés.

— Le positionnement des points le long de la diagonale peut également être étudié. A écarts à la diagonale égaux, un modèle où les points se répartissent mieux le long de la diagonale doit être privilégié. En effet, cela signifie que le modèle arrive efficacement à capter l'information des variables explicatives et à estimer des risques plus précis qu'un autre modèle où les points se superposeraient ou seraient très proches.

Le nombre de groupes doit être choisi judicieusement. Chaque groupe doit contenir assez d'observations pour réduire la variance de la probabilité moyenne observée au sein du groupe, sans quoi des écarts à la diagonale pourraient être interprétés à tort comme des défauts de calibration. Il faut cependant avoir suffisamment de groupes pour assurer l'homogénéité des probabilités estimées au sein de chaque groupe. On recommande d'avoir au minimum 10 ou 20 groupes pour assurer une couverture suffisante des probabilités estimées et des groupes homogènes mais moins de 100 groupes sans quoi le graphique perd en lisibilité. Si les probabilités estimées se concentrent sur un intervalle de valeurs limité à quelques déciles, g peut là encore être réduit.

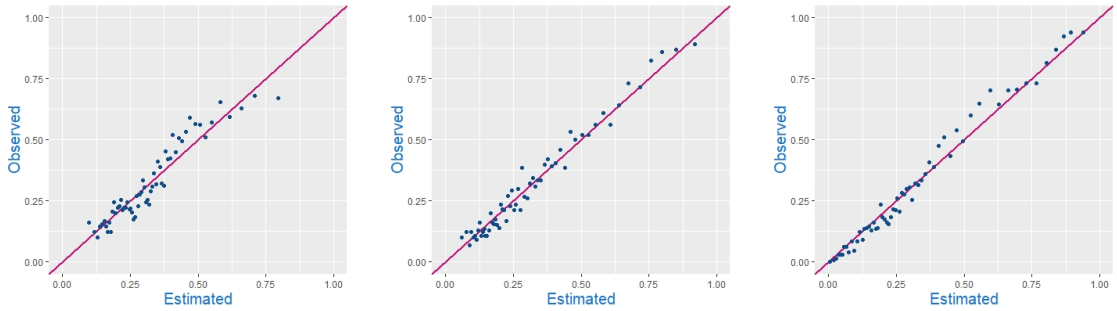

FIGURE 3.2 – Graphes de calibration

La figure 3.2 expose les graphiques de calibration de trois modèles différents appliqués à la même base de données (données réelles de retards de trains). Les graphiques sont tracés avec 60 groupes contenant chacun environ 130 observations. Pour cet exemple le modèle à privilégier serait celui de

droite. Le modèle le plus à gauche est le moins bon : les points sont parfois éloignés de la diagonale, ils sont regroupés sur un intervalle réduit de valeurs et se superposent. Cela indique qu'il estime des probabilités similaires à toutes les observations et capte mal les motifs permettant d'identifier des risques plus extrêmes. Les graphiques du milieu et de droite correspondent à des modèles qui ont un meilleur pouvoir prédictif. Celui du milieu a cependant une moins bonne couverture de l'intervalle $[0, 1]$, les probabilités estimées ont une plus grande variance (points dispersés autour de la diagonale) et est donc aussi moins discriminant. Il ne marque pas de biais à part une légère sous-estimation des valeurs extrêmes. Le modèle de droite présente une légère forme de sigmoïde (points sous la diagonale en dessous d'une probabilité estimée de environ 0.4, et au dessus pour des valeurs supérieures.

3.2.2.3 Le test d'Hosmer-Lemeshow

Cadre classique : le test d'Hosmer-Lemeshow [100] est un test du χ^2 visant à attester l'adhérence de régressions logistiques. Pour chaque observation $y_i \in \{0, 1\}$, on étudie une probabilité binaire p_i d'observer un évènement (ou probabilité de succès). Comme pour construire les graphes de calibration, on trie les sujets par probabilités prédites croissantes et les séparer en g groupes. Plusieurs options de séparation sont possibles : soit par rapport à des valeurs fixes et chaque groupe contient des probabilités comprises dans un intervalle donné, soit par rapport aux quantiles de la loi, formant des groupes de même taille. Les auteurs conseillent plutôt un découpage basé sur les quantiles de la loi, en particulier quand les probabilités prédites sont petites ou grandes, et concentrées dans quelques déciles [98].

La statistique d'Hosmer-Lemeshow est donnée par :

$$C_g = \sum_{i=1}^{g} \frac{(O_{1,i} - E_{1,i})^2}{E_{1,i}} + \frac{(O_{0,i} - E_{0,i})^2}{E_{0,i}} \tag{3.19}$$

Avec $O_{1,i}$ et $E_{1,i}$ (resp. $O_{0,i}$ et $E_{0,i}$) les nombres observés et prédits de succès (resp. echec) dans le groupe i. Sous l'hypothèse nulle que le modèle suit les données, cette statistique suit une loi du χ^2 à $g - 2$ degrés de liberté. Un modèle sera considéré comme calibré si on a un test non significatif : cela signifie que le test a échoué à rejeter l'hypothèse nulle. En pratique, la probabilité p d'observer C_g sous hypothèse nulle et calculée. Si $p \leq 0.05$ (ou un autre niveau), on considérera que le modèle n'est pas calibré.

Le test standardisé : ce test repose sur un test du χ^2 qui est sensible à la taille de l'échantillon. En effet, le pouvoir du test augmente avec n, ce qui peut amener à rejeter des modèles ne présentant que des très faibles déviations ou erreurs de modélisation. La conclusion du test dépend également du choix du nombre de groupes puisque le pouvoir du test décroît quand g augmente. Une version standardisée du test d'Hosmer-Lemeshow est proposée afin de corriger cette dépendance pour des bases de données jusqu'à 25 000 observations [146]. Pour n sujets, m succès, il est suggéré d'utiliser :

$$g = \max\left(10, \min\left(\frac{m}{2}, \frac{n-m}{2}, 2 + 8(\frac{n}{1000})^2\right)\right) \tag{3.20}$$

Le terme $2 + 8(\frac{n}{1000})^2$ vise à uniformiser le pouvoir du test dans le cas général, avec une valeur minimale de 10 groupes imposée. C'est une valeur de référence utilisée dans la littérature et en dessous de 6 le nombre de groupes peut être insuffisant pour que la statistique soit distribuée selon la loi χ_{g-2} [146]. Cette approximation peut également échouer quand le nombre de groupes est trop grand ou que les taux de succès (ou d'échecs) sont trop importants, d'où les termes en $\frac{m}{2}$ et $\frac{n-m}{2}$ pour assurer approximativement un nombre minimal de représentants de chaque classe dans les groupes. La statistique peut tout de même avoir une valeur trop importante à cause de déviations dans les premiers et derniers déciles.

La sensibilité du test dans le cas de probabilités extrêmes a été étudiée spécifiquement par Hosmer et al [99]. Ils concluent que le test est relativement fiable quand moins de la moitié des probabilités estimées sont inférieures à 0.1. Dans le cas où la majorité des prédictions sont dans le premier ou dernier décile, la statistique d'Hosmer-Lemeshow peut avoir une valeur trop importante qui ferait échouer le test. Paul et al [146] affirment que dans la plupart des simulations l'hypothèse du χ^2 a été généralement respectée malgré des probabilités extrêmes.

Alternatives au test standardisé : Paul et al. [146] déconseillent le test d'Hosmer-Lemeshow standardisé pour les bases de données de plus 25000 observations en raison de la trop forte croissance du nombre de groupes. Ils proposent deux idées pour dépasser cette limite : l'échantillonnage et l'utilisation de plus petites valeurs de groupes, si le test n'est pas significatif, le modèle peut être considéré comme calibré.

L'échantillonnage apparaît comme une solution plus stable : on sélectionne des sous ensembles de données, par exemple de taille 1000, qu'on teste avec 10 groupes pour rentrer dans le cadre classique du test d'Hosmer Lemeshow, et le nombre de tests significatifs obtenus est compté. Si le modèle est bon, un taux de 5% de tests significatifs devrait être obtenu. Cette approche a été peu étudiée, mis à part Bartley [18] qui l'utilise dans sa thèse. Il expérimente en simulant de grandes bases de données dont il connaît la calibration et les décompose en 100 sous-ensembles de taille 1000, 2000 et 5000 obtenus par tirages avec remise. On compte le nombre de tests significatifs qui doit être bas pour de bons modèles. Le constat est qu'en général avec des échantillons de taille plus grande le test est plus souvent significatif. Il constate que même pour un modèle parfait, il y a souvent plus de 5% des tests qui sont significatifs.

Les tests de calibration en analyse de survie : plusieurs tests différents ont été proposés pour étendre le test d'Hosmer-Lemeshow au cadre d'analyse de survie, prenant à la fois en compte le caractère continu de la variable, et la potentielle présence de données censurées. Sans la censure, ces tests sont très proches d'un test d'Hosmer-Lemeshow appliqué à plusieurs points de la fonction de répartition pour en vérifier la qualité. Dans ces cas là, les probabilités de survie après un temps t sont étudiées, les données sont agrégées en groupes au sein desquels on compare la probabilité estimée avec l'estimation de Kaplan-Meier [93]. Demler et al. [66] proposent une comparaison des principales méthodes. On compte notamment le test de Nam and d'Agostino [61], le test de Gronnesby and Borgan [85] et enfin son équivalent développé par May and Hosmer [132]. Ces exemples d'application sont à notre connaissance les seuls cas où le test d'Hosmer-Lemeshow ait été adapté pour étudier la calibration d'une distribution de probabilité non bianire.

3.2.2.4 Transformation PIT et résidus de quantiles

PIT : La méthode de transformation PIT (*Probability integral transform*) [17, 69] permet de créer un ensemble d'observations indépendantes et uniformément distribuées à partir de données dont on dispose de la fonction de répartition.

La PIT peut être utilisée pour attester de la calibration d'un modèle de prédiction de densité. Soit y_i une réalisation de la variable aléatoire Y, et F_i sa fonction de répartition, la transformation PIT s'obtient par :

$$u_i = F_i(y_i)$$

Si le modèle est correct, alors $U \sim U(0,1)$. On peut tester cela à l'aide d'un test de Kolmogorov-Smirnov. En effet, pour toutes variables continues Y et U telles que $U = F_Y(Y)$ et F_Y la fonction de répartition de Y, et pour tout $u \in [0,1]$ on a :

$$\begin{aligned} F_U(u) &= \mathbb{P}(U \leq u) \\ &= \mathbb{P}\left(F_Y(Y) \leq u\right) \\ &= \mathbb{P}\left(Y \leq F_Y^{-1}(u)\right) \\ &= F_Y\left(F_Y^{-1}(u)\right) \\ &= u \end{aligned} \tag{3.21}$$

F_u est la fonction de densité cumulée d'une loi uniforme dans $(0, 1)$.

Cette définition n'est plus valable pour les données de comptage mais on peut l'adapter en randomisant la transformation [59]. Pour P la fonction de masse prédite, y_i la valeur observée et v une variable uniforme standard indépendante de Y, la nouvelle transformation est :

$$\begin{aligned} u_i &= P(y_i - 1) + v\left(P(y_i) - P(y_i - 1)\right) \quad \text{si} \quad y_i \geq 1 \\ u_i &= vP(0) \quad \text{si} \quad y_i = 0 \end{aligned} \tag{3.22}$$

De même, sous hypothèse que le modèle est correct, la calibration des prédictions peut se vérifier en testant que u suit bien une loi uniforme.

RQR : Une autre forme de résidus similaires est parfois utilisée, notamment pour les modèles linéaires généralisés et les modèles GAMLSS [70]. Il s'agit des résidus de quantiles randomisés (RQR).

Soit $F(y, \mu_i, \phi)$ la loi cumulée de $P(\mu_i, \phi)$. Si F continue, les résidus de quantile sont définis par :

$$r_{q,i} = \Phi^{-1}[F(y_i, \mu_i, \phi)]$$

où Φ est la fonction de répartition de la loi normale standardisée. Il s'agit d'une normalisation des résidus formés par la transformation PIT. Cette normalisation est intéressante car il existe plus de tests statistiques pour vérifier qu'une variable suit une loi normale plutôt qu'une loi uniforme et leur spécificité à la loi normale les rend plus efficaces. Cependant, comme pour le test d'Hosmer-Lemeshow, la taille de l'échantillon compromet la validité des tests. Une approche plus robuste est d'échantillonner afin de réaliser les tests sur des sous-ensembles de données de taille standard, dans des conditions où le pouvoir de ces tests est connu.

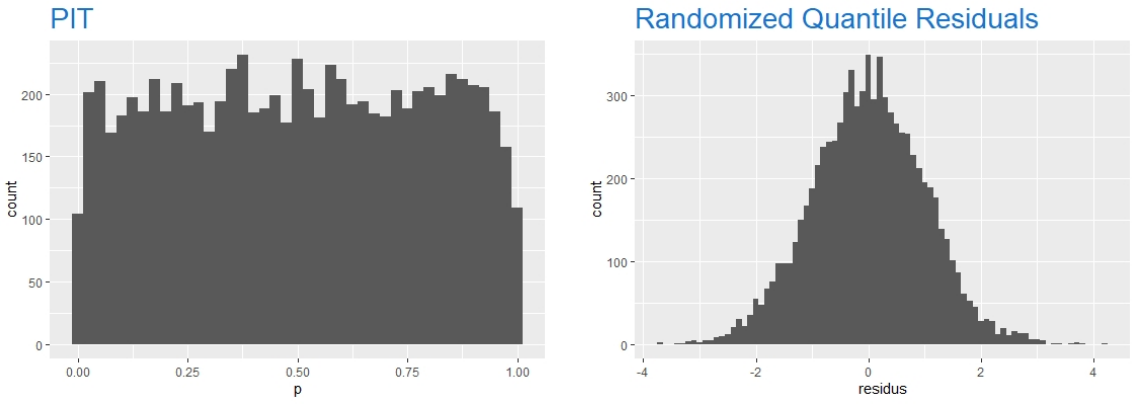

FIGURE 3.3 – Visualisation des résidus PIT et RQR

Deux visualisations de ces transformations calculées sur un même ensemble d'estimations de distributions discrètes sont données dans la figure 3.3.

3.2.2.5 Comparaison

L'attestation de la calibration est une étape capitale pour l'utilisation des probabilités pour la prise de décision. Les avantages et inconvénients de ces méthodes sont résumés dans le tableau 3.2.

Test d'Hosmer-Lemeshow standardisé	Randomized quantile residuals, PIT	Graphes de calibration
Avantages : — Test standardisé pour ajuster le pouvoir du test à la taille de l'échantillon — Représentation graphique grâce aux graphes de calibration	Avantages : — Un seul test nécessaire pour l'ensemble de la distribution — Représentation graphique des résidus	Avantages : — Facile à représenter et à interpréter — Pouvoir discriminant visible graphiquement
Inconvénients : — Standardisation incomplète — Échecs possibles de la standardisation pour les évènements rares — Nécessite plusieurs tests à différents points de la distribution	Inconvénients : — Tests de normalité non standardisés — Graphiques peu interprétables en cas d'échecs	Inconvénients : — Pas de garantie statistique — Besoin de représenter les graphes à différents points de la distribution

Tableau 3.2 – Comparaison des analyses de calibration

Les graphes de calibration sont simples à tracer et à interpréter, nous recommandons de toujours les visualiser pour avoir une première idée des performances du modèle. Dans le cas d'une prédiction multicatégorielle, la qualité des probabilités peut s'évaluer en *one-versus-all*, c'est-à-dire en évaluant la probabilité d'appartenir à une classe donnée à la fois. Pour des données ordinales ou numériques on peut tracer les graphes à certains seuils donnés. Plus le nombre de graphes augmente, plus l'interprétation des résultats est complète mais elle est également plus difficile à mener. Ces graphes sont importants pour le cas où les tests statistiques échouent, afin de juger selon le problème étudié si les déviations dans les probabilités prédites sont acceptables ou non pour le problème étudié.

La statistique d'Hosmer-Lemeshow est basée sur les écarts quadratiques entre le taux d'évènements estimé et la valeur observée dans chaque groupe, sans considération du sens des déviations. On conseille d'afficher un graphe de calibration en complément pour vérifier la présence d'un biais d'une forme en sigmoïde qui pourrait ne pas être détectée.

Pour les approches utilisant des tests statistiques, nous recommandons plutôt l'utilisation d'échantillonage des données quand la base est assez grande pour assurer une standardisation des conditions d'applications du test et baisser la variance des résultats. C'est en particulier vrai pour le test d'Hosmer-Lemeshow dont l'issue dépend fortement du découpage en groupe qui est appliqué.

3.2.2.6 Correction de la calibration

Il est possible de partiellement contrôler la calibration des probabilités estimées en les ajustant a posteriori. Deux méthodes classiques ont été proposées pour le cas binaire : l'échelle de Platt et la régression Isotonique [141].

Soit $f(\boldsymbol{X})$ la sortie du modèle, par exemple la probabilité non calibrée. La méthode proposée par Platt estime par maximum de vraisemblance les paramètres A et B tels que $\mathbb{P}[y = 1|f(X)] = \frac{1}{1+exp(Af(X)+B)}$. La régression isotonique construit une fonction constante par morceaux croissante prenant en entrée la probabilité d'évènement prédite et donnant en sortie la nouvelle probabilité calibrée. La fonction m retenue doit réduire les écarts aux carrés entre la probabilité estimée calibrée $m(f(\boldsymbol{X}))$ et y. Dans les deux cas il est conseillé d'utiliser des données n'ayant pas servi à construire les modèles pour éviter un surapprentissage.

Ces méthodes permettent de corriger les déformations de probabilités, par exemple quand les graphes de probabilités ont une forme de sigmoïde, ce qui est caractéristique des algorithmes de Boosting et de SVM. Les probabilités obtenues par forêts aléatoires sont moins concernées et les régressions logistiques donnent en général des probabilités très bien calibrées du fait de l'optimisation directe par maximum de vraisemblance [141, 147].

3.2.3 Analyse de la discrimination

Des prédictions sont appelées discriminantes quand elles permettent de différencier efficacement les sujets à valeur plus élevée. On parle également de concordance, dans le sens où les probabilités estimées doivent être concordantes avec les valeurs observées.

La mesure la plus classique de la discrimination est l'indice de concordance ou statistique c. Pour deux observations Y_1 et Y_2 avec comme valeurs estimées par le modèle respectivement $\hat{Y_1}$ et $\hat{Y_2}$, la probabilité de concordance vaut :

$$c = \mathbb{P}\left(\hat{Y_2} > \hat{Y_1} | Y_2 > Y_1\right) \tag{3.23}$$

On peut l'estimer en calculant la proportion de couples d'observations pour lesquels le sujet avec l'observation la plus importante a également la valeur prédite la plus grande. Cette statistique atteste du respect de l'ordre des éléments entre les prédictions et les observations.

La formulation générale de la concordance 3.23 implique que les estimations évaluées peuvent se ramener à une grandeur simple $\hat{Y}$ dont l'ordre doit être cohérent avec le classement des observations. Plusieurs variantes ont été imaginées selon les cas d'application.

- variables binaires : la concordance est particulièrement étudiée dans le cas binaire en prenant comme prédiction la probabilité de succès. Cette quantité est alors égale à l'aire sous la courbe ROC (*Receiver Operating Characteristic curve*) [89].
- variables qualitatives non ordonnées : elles peuvent être étudiées par comparaison une à une ou une contre toutes et se ramener au cas binaire pour induire un ordre.
- variables ordinale : les classes sont numérotées et $\hat{Y}$ est l'indice de classe prédite par le modèle. La concordance dans une paire est définie par le respect de l'ordre entre les classes prédites et les classes observées [45].
- variables numériques :
  - pour les GLM, le plus classique est d'utiliser l'espérance de Y comme estimation $\hat{Y}$ [186]
  - pour les modèles de Cox à risque proportionnel (cf 3.1.1.2), Gönen et Heller [80] proposent une expression de la concordance utilisant le temps de survie estimé, qui est une fonction monotone du risque. D'Agostino et Ham ont proposé une généralisation de la statistique C au cadre de l'analyse de survie en conservant l'idée qu'il s'agit de la proportion de couples concordants à un instant t [61]. Contrairement à Gönen et Heller, la discrimination est donc étudiée à un instant précis. L'intégration de la censure dans le calcul de la statistique distingue leur mesure du cas binaire.

3.3 Conclusion

3.3.1 Synthèse

Ce chapitre a exposé différentes méthodologies utilisées dans la littérature pour estimer des probabilités, ainsi que les différentes techniques de validation associées. Plusieurs avantages à travailler sur des probabilités plutôt que sur des prédictions ponctuelles ont été identifiés, en particulier pour les biostatistiques, cependant ce domaine manque encore de structure.

L'estimation de probabilité est encore largement menée à l'aide de méthodes paramétriques, le plus souvent des régressions logistiques ou des modèles de Cox. Les modèles plus complexes qui permettent une meilleure adéquation aux données, tels que les modèles GAMLSS, sont plus récents, et on constate que les méthodes de validation de résutats sont rarement adaptées à ces approches (en particulier les cas non binaires, avec plusieurs paramètres ou avec un paramètre différent de la moyenne). L'alternative algorithmique pour la prédiction de probabilité est elle aussi très récente, et uniquement dédiée aux cas binaires et catégoriels.

Le choix entre une approche algorithmique ou paramétrique dépend principalement du support des variables, puisque les algorithmes de forêts aléatoires présentés ici ne peuvent pas être utilisés pour modéliser des variables continues. Les besoins de l'étude (portabilité, performance, interprétabilité etc.) ainsi que les données disponibles (besoin d'encodage ou de sélection, types d'interactions, etc.) sont également à prendre en compte.

La sélection du meilleur modèle pour la prédiction de probabilité se fait facilement à l'aide de scores, en particulier avec un score logarithmique et quadratique. Ces scores sont cependant relatifs : ils permettent d'identifier l'estimateur qui donne les meilleures prédictions mais ne garantissent pas leur qualité.

La qualité des prédictions s'évalue en fonction de leur calibration et discrimination. La discrimination se mesure facilement à l'aide de la concordance. La calibration est plus complexe à vérifier : plusieurs méthodes sont possibles, mais chacune a des limites (pas de garantie par test statistique, résultats difficiles à interpréter, tests peu fiables).

Le chapitre suivant utilise les outils introduits ici sur les données de retards de train. On illustrera les différents avantages et inconvénients des méthodes de modélisation, et on proposera une méthodologie d'estimation et validation des distributions de probabilité pour ce cas d'étude.

3.3.2 Méthodologie choisie pour les retards de trains

Les données étudiées dans ces travaux sont des relevés de retards de trains, mesurés en minutes, positifs et bornés. On cherche à appliquer les méthodes statistiques présentées ci-dessus pour estimer des lois de probabilités de retard.

Les probabilités prédites seront utilisées pour choisir entre les différents scenarios dans les planifications de quais et itinéraires en gare afin de limiter les conflits dus aux retards et gagner en robustesse. Le modèle doit être calibré, afin de ne pas sous-estimer le risque et bien jauger les marges nécessaires. Il est cependant plus important qu'il soit discriminant dans la mesure où différents niveaux de risque doivent être identifiés pour prendre des décisions là où les conséquences sont les plus importantes.

A première vue, l'approche par GLM et l'approche par forêt aléatoire sont toutes les deux valides. Les modèles linéaires généralisés constituent une solution naturelle puisque les retards de trains sont connus pour suivre une distribution appartenant à la famille exponentielle (cf 2.4). Une distribution discrète adéquate devra être choisie, et le modèle entraîné. Pour que l'algorithme de forêt aléatoire soit appliqué ici, il faut considérer les retards comme des données catégorielles, ce qui fait perdre une partie de l'information inhérente aux données, mais assure également plus de flexibilité. Les forêts aléatoires

ont cependant le mérite de nécessiter moins de prétraitement des données et de mieux détecter des motifs complexes entre les prédicteurs et les retards de trains

Dans les deux cas, on recommande de construire et sélectionner le modèle final en se basant sur un des scores proposés précédemment. C'est le cas directement pour des estimations de maximum de vraisemblance, comme pour les GLM, et il suffit de choisir un critère adapté pour construire la forêt, comme par exemple la cross-entropie qui correspond au score logarithmique. La capacité prédictive et l'utilité des modèles seront évaluées directement via la concordance en utilisant le retard moyen et en vérifiant la calibration, en particulier sur des données futures.

Cependant les méthodes de vérification de la calibration utilisant des tests statistiques ne sont pas très adaptées ici : elles sont instables du fait de leur dépendance en la taille de l'échantillon, et sont peu interprétables. Une déviation entraînera le rejet de la calibration même si elle est acceptable concernant la prise de décisions. Dans notre cas, il est important que les probabilités soient bien calibrées, mais des déviations sont acceptables. On se concentrera donc sur l'étude des graphes de calibration.

Chapitre 4

Estimation du risque de retard de trains : le cas de la gare Montparnasse

Ce chapitre présente la méthodologie et les résultats obtenus pour l'analyse et la prédiction de retards de trains pour une grande gare au trafic hétérogène. Quel que soit le type de train et le sens de circulation considérés, la grande majorité des retards sont nuls ou de l'ordre de quelques minutes. Ces petits retards sont plus souvent signes d'une instabilité du réseau (petits retards accumulés durant les arrêts, signaux fermés, etc.) que de réels incidents dont les conséquences sont plus lourdes et les origines imprévisibles. Dans un cadre d'adaptation des planifications, c'est-à-dire quelques jours à mois avant les circulations, les informations disponibles sont limitées, et en particulier les variables ayant le plus d'impact sur le retard ne sont pas accessibles (cf 2.4). Par exemple la météo du jour, les incidents en cours, les retards aux arrêts précédents ou encore l'affluence ne sont pas connus, ce qui ne permet pas d'estimer précisément le retard.

Compte tenu de ce constat, le cadre de travail proposé ici s'articule sur deux choix de modélisation importants :

— Les modèles présentés visent à estimer la probabilité de retard des trains conditionnellement à leurs caractéristiques plutôt que de prédire une valeur ponctuelle. Estimer la valeur précise du retard avant la veille est irréaliste étant donné les informations disponibles. Par ailleurs les méthodes classiques de régression ou de classification tendent à prédire la même valeur pour tous les trains (retard nul ou faible), sans modéliser la dispersion des valeurs. Estimer la probabilité de retard permet de quantifier efficacement le risque associé à chaque valeur, et pas seulement de renvoyer la valeur la plus probable ou le retard moyen estimé.

— Seuls les petits retards sont modélisés, c'est-à-dire les retards d'une à quelques dizaines de minutes. Les retards plus importants sont peu probables et sont très peu corrélés aux données disponibles à long-terme. Ils biaiseront les modèles du fait qu'ils n'ont pas les mêmes origines que les petits retards qui sont des signes d'instabilité du réseau. Par ailleurs, seuls les petits retards doivent être pris en compte pour des aspects de robustesse, les trains très retardés étant pris en charge de manière très spécifique. Les modèles présentés ici sont donc tronqués à droite pour ne représenter que les retards inférieurs à un seuil donné. Les trains en avance (retards négatifs) sont considérés comme à l'heure.

4.1 Présentation du cas d'étude

Les méthodes statistiques présentées précédemment sont appliquées aux données de retards de trains en gare dans ce chapitre. L'enjeu final est de réussir à identifier correctement les motifs dans les retards et de proposer une réorganisation des affectations de quais et d'itinéraires en gare qui en minimise l'impact. Le cas d'étude utilisé est celui de la gare Montparnasse à Paris.

4.1.1 Gare de Paris Montparnasse

Cette gare terminale est formée par 28 voies réparties sur deux sites : le site principal et la gare de Vaugirard. Plus de détails sur la gare et ses circulations peuvent être trouvées en annexe A. Elle accueille un trafic varié, avec des circulations commerciales qui transportent des passagers et des circulations techniques. Les circulations techniques comprennent principalement des trains dont le départ ou l'arrivée correspond à un trajet depuis (ou vers) le technicentre, des trajets à vide pour permettre les roulements de matériel ou certains trains de travaux, mesure, etc. Les trains techniques sont horairisés pour la gare de Montparnasse, cependant leurs horaires sont rarement respectés dans la mesure où ces circulations sont bien plus flexibles et moins contraintes que celles transportant des passagers. Les retards des circulations techniques sont alors beaucoup plus dispersés, avec une grande partie de trains en avance et des valeurs très importantes en valeur absolue. Cette étude se concentre donc sur les trains commerciaux dont les retards sont à la fois beaucoup moins bruités et dont les enjeux sont plus importants pour le niveau de service de l'entreprise.

Les trains commerciaux circulant à Montparnasse sont :

- Les trains TGV : ils desservent les régions Bretagne, Centre-Val de Loire, Pays de la Loire et Nouvelle-Aquitaine, ils correspondent à environ 36% du trafic
- Les trains régionaux (15% des circulations) :
  - TER Centre-Val de Loire vers Le Mans, Chartres et Nogent-le-Rotrou
  - TER Normandie et anciens Intercités vers Argentan et Granville
- Les trains Transilien de la ligne N. Ils assurent la desserte vers Mantes-la-Jolie, Dreux et Rambouillet, en intégrant notamment les noeuds de Versailles et Sèvres et représentent 49% des trains commerciaux.

4.1.2 Contribution

Ce chapitre propose une méthodologie complète de traitement des données de retards de trains dans le but de quantifier l'incertitude sur les horaires. Cette méthodologie inclut la construction et la sélection d'un modèle d'estimation du risque, mais également l'étude de sa stabilité sur des données de test qui appartiennent au "futur". Les grandes étapes de cette méthodologie sont explicitées dans la figure 4.1.

Les trois premières étapes concernent le travail préliminaire à effectuer avant toute modélisation. La construction des bases de données finales, de la description des données brutes aux choix d'encodage, est explicitée dans la section 4.2. Les principaux résultats de l'analyse de ces bases de données sont montrés dans la section 4.3, avec en particulier des considérations sur la distribution que suivent les retards de train et des graphiques montrant l'impact des variables identifiées à l'étape précédente.

La section 4.4 décrit les étapes de construction des modèles. Les deux familles de méthodes présentées dans le chapitre 3 sont testées sur les données de retard, à savoir les modèles linéaires généralisés (GLM) et des forêts aléatoires (RF).

Enfin, les résultats de validation et d'évaluation sont donnés dans la section 4.5. La validation permet de sélectionner le modèle final qui sera utilisé en pratique pour estimer les risques individuels

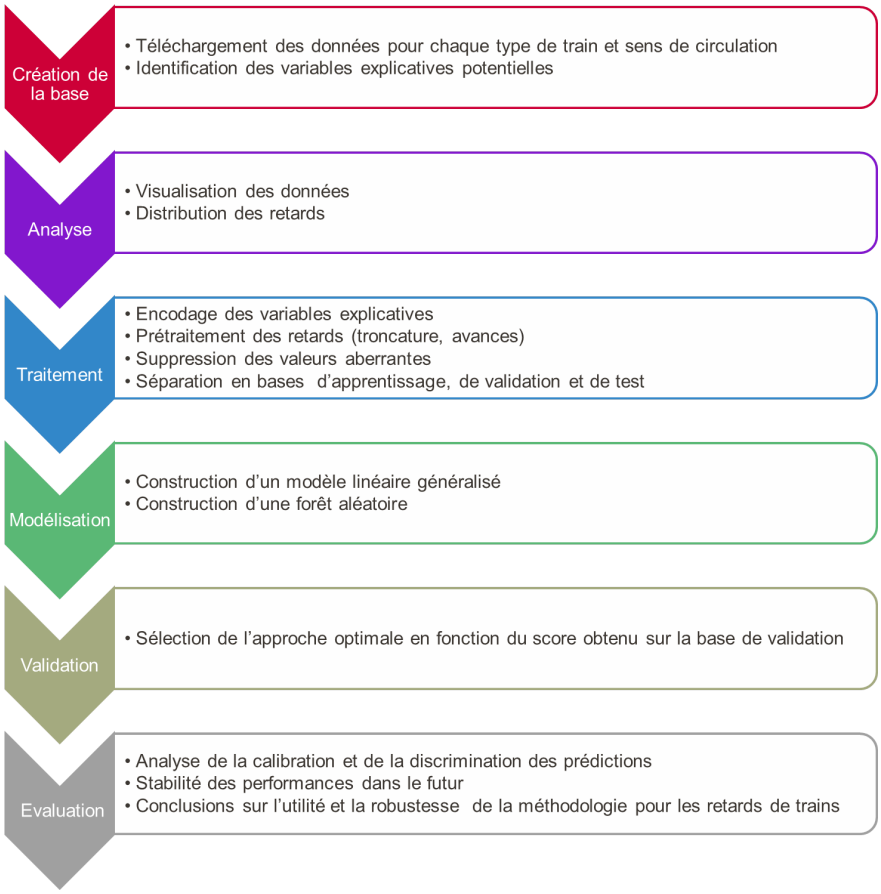

FIGURE 4.1 – Méthodologie de prédiction

de retards. L'étape d'évaluation des performances n'a pas d'influence sur les décisions de modélisation mais vise à établir la confiance qu'on peut avoir dans les prédictions en utilisant des données postérieures à celles d'apprentissage. Dans l'hypothèse d'une industrialisation, ces données de test ne seront pas disponibles au moment de la construction du modèle, mais elles sont utilisées ici pour valider l'utilité et la fiabilité de la méthodologie.

Les limites et perspectives de ces travaux sont discutées dans la section 4.6.

4.2 Création de la base de données

Cette section décrit succinctement les différentes étapes de constitution de la base de données utilisée pour la suite de ces travaux. On traite en premier lieu l'extraction des données, puis le choix des variables explicatives et enfin la gestion des exceptions et anomalies statistiques. La liste des variables est explicitée dans l'annexe B.

Circulation		Observations						
Numéro	Date	CI	Date observation	Type horaire	Catégorie statistique	Structure suiveuse	Série tracée	Ecart horaire
867281	29/08/2016	PAU	29/08/2016 16:31	O	RBB	SNCF TER	Z51500	0
867432	29/08/2016	PAU	29/08/2016 16:46	T	RCB	SNCF TER	X73500	-2
867228	29/08/2016	PAU	29/08/2016 17:04	T	RBB	SNCF TER	Z51500	-1
8585	29/08/2016	PAU	29/08/2016 17:04	A	LVA	SNCFVOYAGE	TGVAT	1
8585	29/08/2016	PAU	29/08/2016 17:07	D	LVA	SNCFVOYAGE	TGVAT	2
867435	29/08/2016	PAU	29/08/2016 17:11	O	RCB	SNCF TER	X73500	0
14148	29/08/2016	PAU	29/08/2016 17:15	A	NBA	SNCFINTERC	7200	2
14148	29/08/2016	PAU	29/08/2016 17:21	D	NBA	SNCFINTERC	7200	5
867285	29/08/2016	PAU	29/08/2016 17:26	O	RBB	SNCF TER	Z51500	0
872707	29/08/2016	PAU	29/08/2016 17:31	T	RBT	SNCF TER	Z7300	1
867023	29/08/2016	PAU	29/08/2016 17:39	O	RCB	SNCF TER	X73500	3
872740	29/08/2016	PAU	29/08/2016 17:53	O	RBT	SNCF TER	7200	0
867434	29/08/2016	PAU	29/08/2016 18:27	T	RCB	SNCF TER	X73500	0
867437	29/08/2016	PAU	29/08/2016 18:34	O	RCB	SNCF TER	X73500	0
14151	29/08/2016	PAU	29/08/2016 18:43	A	NBA	SNCFINTERC	7200	3
867024	29/08/2016	PAU	29/08/2016 18:43	A	RBB	SNCF TER	Z51500	0
14151	29/08/2016	PAU	29/08/2016 18:44	D	NBA	SNCFINTERC	7200	1
867024	29/08/2016	PAU	29/08/2016 18:47	D	RBB	SNCF TER	Z51500	0
867287	29/08/2016	PAU	29/08/2016 18:58	O	RBB	SNCF TER	Z51500	0

FIGURE 4.2 – Exemple d'extraction sur l'outil e-Brehat

4.2.1 Données brutes

Base de données Brehat : Brehat est un outil de SNCF Réseau permettant le suivi des circulations et de leurs retards par centralisation en temps réel des données remontées par des balises qui détectent les passages des trains. Elles sont disposées sur les voies à des *points remarquables* ou PR du réseau ferré. Ces balises ne sont pas toujours positionnées au même niveau sur les quais ou le long de voies parallèles ce qui peut ajouter un léger bruit aux valeurs mesurées. Ces observations sont enregistrées et constituent une base de données de retards couvrant l'ensemble des circulations et du territoire [4].

Pour un PR donné, les données disponibles sont la liste des circulations enregistrées à ce point précis avec comme informations disponibles le numéro du train, l'heure et la date d'observation, le type de mouvement (passage, arrivée, départ), l'entreprise ferroviaire, le type de matériel, et surtout l'écart horaire en minutes observé par rapport à l'horaire théorique enregistré. Un exemple d'une extraction d'une base de données Brehat est présentée dans l'image 4.2.

Extractions : plusieurs extractions de données ont été faites pour les zones suivantes entre les dates du 1ier juillet 2017 et le 31 mars 2019 :

- Bâtiment Voyageur de la gare Montparnasse, grâce aux balises placées au niveau des quais
- A l'entrée dans l'avant gare de Montparnasse, soit avant la zone de routage jusqu'aux quais
- Gares des réseaux TGV Atlantique, Transilien ligne N, TER Centre et Normandie avec connexion à Paris Montparnasse

Les données d'avant gare sont utilisées pour les trains arrivant à Montparnasse et celles du bâtiment voyageur sont utilisées pour le départ car elles n'incluent pas les retards créés en zone d'avant gare. Les statistiques calculées dans ce chapitre seront ensuite exploitées pour la robustesse des operations en gare (GOV, routages), elles ne doivent donc pas inclure les minutes créées dans le périmètre de la gare Montparnasse. Le modèle de reconstitution du retard en avant gare à partir des différents PR est donné en annexes A.2.1 et B.1.1.

Traitement préliminaire : les données récoltées sur les différents points du réseau sont utilisées pour reconstituer le parcours planifié de chaque train : le numéro de train associé à la date forment un identifiant unique des circulations commerciales. Croiser les données des différentes gares en fonction

du numéro de train et de la date permet donc de connaître les horaires des arrêts planifiés, et d'en déduire l'origine, la desserte, la densité du trafic et les temps d'arrêts prévus.

Données complémentaires : cette base de données a été uniquement complétée par le calendrier des vacances scolaires. Quelques tests ont été effectués en incluant des données météorologiques (température, précipitations, vent) dont les résultats étaient probants et en accord avec la littérature (cf 2.4). Cependant ces données ne sont disponibles que quelques jours en avance sous forme de prévisions, ce qui rentre difficilement dans notre cadre de travail. D'autres bases complémentaires seraient également pertinentes mais posent parfois des problèmes d'accessibilité de la donnée, comme la planification des travaux, les roulements de matériel roulant ou les opérations de maintenance.

4.2.2 Variables explicatives

Plusieurs variables explicatives sont créées à partir de la base de données brutes. Une description individuelle de chaque variable est donnée en annexe B.1 avec notamment l'encodage choisi. Certaines de ces variables ne sont utilisées que pour les arrivées ou départs.

- Les variables de type de circulation : le type d'activité (TGV, TN, TER) et le matériel programmé.
- Les variables temporelles : description de la plage horaire, type de jour, vacances scolaires.
- Les variables de mission : ligne, Origine/Destination, nombre d'arrêts, temps d'arrêt programmé, marges, etc.
- Les variables de densité : densité aux stations visitées (nombre de trains en circulation en même temps) et densité en ligne (nombre de trains sur la ligne, espacements avec le train précédent et le train suivant).

Ces variables sont principalement issues des archives des circulations et correspondent aux types de variables utilisées dans la littérature et dont la corrélation aux retards a été montrée (cf 2.4). Elles sont cependant très corrélées entre elles, au sein même d'une catégorie (par exemple l'origine du train ou son nombre d'arrêts contraignent la durée du trajet) mais également entre les catégories (la densité du trafic est liée aux heures de pointe et certaines origines ne sont desservies qu'à certains horaires).

4.2.3 Hypothèses et restrictions

Plusieurs hypothèses de modélisations ont été faites pour cadrer ce sujet et sont justifiées ici. Elles permettent notamment d'augmenter l'homogénéité des données ou d'adhérer mieux aux besoins de l'étude.

4.2.3.1 Séparation des types de trains

La figure 4.3 expose les différents profils des retards des trains commerciaux selon le sens du mouvement (arrivée et départ) et le type de train (TGV, TN, TER). Si les profils au départ sont globalement homogènes, avec légèrement plus de retards pour les TER, on observe des différences importantes dans les histogrammes des retards de trains à l'arrivée en fonction de l'entreprise ferroviaire en charge.

Chaque type de train est exploité de manière spécifique par des entreprises ferroviaires différentes avec des différences de matériel, de vitesse, etc. Les expérimentations sont menées en traitant séparément selon le sens du mouvement et le type de train afin d'augmenter l'homogénéité des données. Les causes des retards et leurs amplitudes varient beaucoup selon les différents cas, les mélanger risque de dissimuler des motifs qui ne seraient présents que pour un type de train.

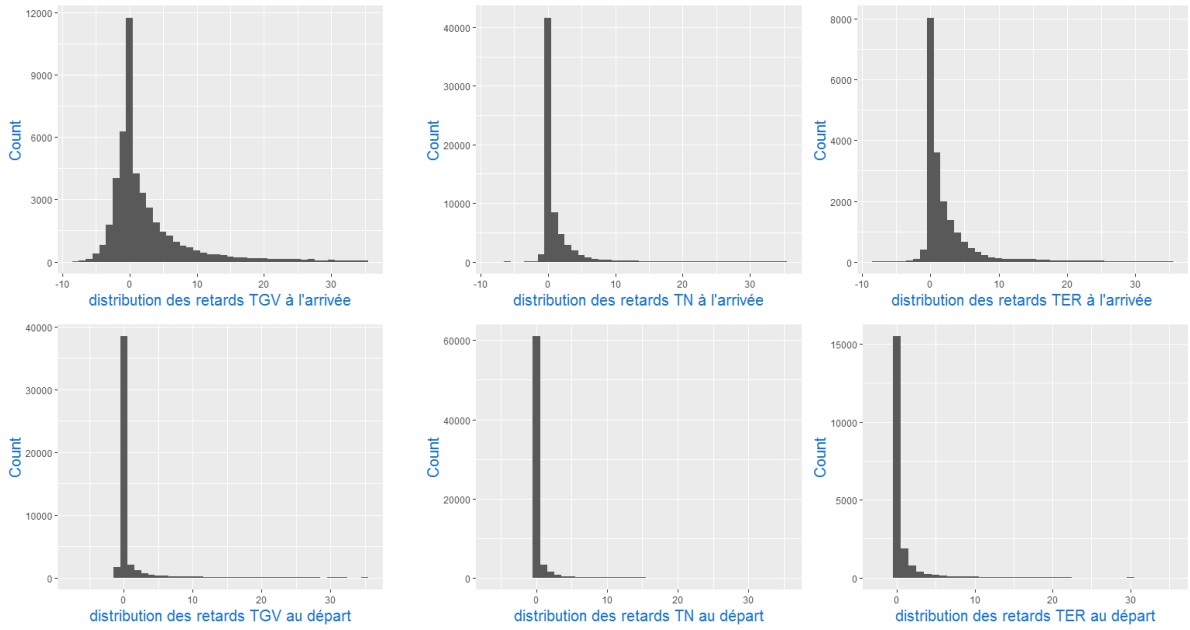

FIGURE 4.3 – Histogramme des retards observés

Comme expliqué précédemment seuls les retards des trains passagers sont modélisés, ceux des trains techniques étant beaucoup plus dispersés et moins contraints par le système, leur prédiction serait à la fois beaucoup plus complexe et bien moins utile dans notre cas.

4.2.3.2 Troncature et modélisation des petits retards

Une hypothèse forte de modélisation a été faite dans ces travaux concernant l'amplitude des retards à étudier. Elle consiste à ne modéliser que les *petits retards*, c'est-à-dire les retards inférieurs à un seuil donné dépendant du type de train. Plusieurs raisons permettent de justifier ce choix.

Argument de robustesse : l'objectif final de cette thèse est d'exploiter les archives de retards pour augmenter la robustesse aux retards des planifications d'occupation des voies. Cependant il n'est pas réaliste d'essayer de se protéger contre des retards importants dont la probabilité d'occurrence est très faible et qui nécessiteront des mesures exceptionnelles pour être pris en charge en opérationnel. La robustesse consiste plus à trouver une solution qui restera réalisable si les paramètres d'entrées changent légèrement.

Argument ferroviaire : les petits retards et les grands ne dépendent en général pas des mêmes causes. Un retard important peut être la suite d'un incident voyageur ou d'une panne alors qu'un petit retard peut n'être que le signe d'une instabilité du système qu'on cherche à mesurer ici. Ainsi, plusieurs ordres de grandeur de retards ne devraient pas être modélisés simultanément dans la mesure où ils ne répondent pas des mêmes mécanismes de création.

Argument de modélisation : deux méthodes de prédictions des risques ont été présentées dans le chapitre précédent. Celle utilisant des forêts aléatoires fonctionne en considérant les retards comme

des classes mutuellement exclusives. L'algorithme n'est possible qu'avec un nombre fini de classes, ce qui est le cas pour des retards en minutes tronqués.

Les troncatures choisies sont de 20 minutes pour les TGV à l'arrivée, 10 minutes au départ, 15 minutes pour les TER à l'arrivée, 8 minutes au départ, 10 minutes pour les TN à l'arrivée et 5 minutes au départ. Ces seuils excluent entre 3 et 6% des retards. Ils ont été établis après concertation avec des experts ferroviaires, et intègrent à la fois les contraintes liées au problème (au delà de 20 minutes les retards ne sont plus pertinents pour la robustesse), les spécificités d'exploitation (par exemple les TN sont régulièrement supprimés au delà de 10 minutes de retard ce qui crée des absences d'observations si un seuil supérieur est utilisé) et le pourcentage de la base qui est supprimé.

4.2.3.3 Cas des trains en avance

Une seconde hypothèse importante faite ici est de considérer les trains arrivant en avance comme étant à l'heure. Ce phénomène ne concerne pas les trains au départ qui ne sont pas autorisés à partir plus tôt que prévu.

Comme on peut le voir sur les histogrammes des retards sur la figure 4.3, les avances concernent presque exclusivement les TGV qui peuvent avoir de 1 à 5 minutes d'avance à l'arrivée. En théorie des trains en avance peuvent également générer des conflits en gare et en avant gare. On fait cependant l'hypothèse qu'un train grande vitesse arrivé en avance aurait eu le temps d'ajuster sa vitesse en cas de besoin pour adhérer aux horaires prévus puisque les arrêts sont espacés les uns des autres. Cette avance était connue des agents et opérateurs qui ont laissé le train circuler comme cela probablement car cela ne générait pas de conflit.

Cette hypothèse n'a pas beaucoup d'impact pour une modélisation utilisant des forêts aléatoires puisqu'il suffit de créer des classes supplémentaires pour chaque valeur négative et d'ajouter une troncature à gauche. Pour une modélisation avec des GLM, le problème devient plus compliqué puisqu'il n'existe pas de distribution classique permettant de modéliser une variable discrète à valeurs positives et négatives. Une option serait de translater les valeurs pour travailler sur un domaine positif ou de composer un GLM en trois parties avec une partie pour les valeurs négatives et une seconde pour les valeurs positives, et une régression logistique qui donne la probabilité d'être positif ou négatif.

4.2.3.4 Indépendance des retards

Les techniques de modélisations utilisées requièrent l'indépendance des observations. Dans le cas des retards de trains, cette hypothèse peut être fausse. On peut considérer que des trains circulant des jours différents ou sur une infrastructure distincte sont indépendants. Pour une même journée, cette hypothèse ne tient plus : le retard d'un train peut impacter la ponctualité des circulations proches.

La troncature permet de réduire les dépendances entre les retards en isolant les incidents majeurs qui ont un impact fort, et grâce à une valeur seuil inférieure à la fréquence de circulation des trains d'une même desserte. Par exemple, si un TGV est retardé à Bordeaux, ce retard n'aura pas ou peu d'impact sur le train suivant passant par Bordeaux et avec un terminus à Paris. Ce raisonnement n'est pas valable pour les TN, car la circulation en zone dense est très sensible à l'affluence dans le train. Cette affluence est perturbée par un retard antérieur. Cependant, il est fréquent qu'un TN soit supprimé pour permettre un retour rapide à la normale et éviter les effets de propagation.

L'élément de dépendance le plus important concerne les trains d'un même type à l'arrivée car ils peuvent partager la même infrastructure en fin de trajet. S'ils n'ont pas la même desserte, il est possible qu'ils arrivent à quelques minutes d'intervalle par les mêmes voies, et dans ce cas là le retard du premier train peut impacter le second.

4.2.3.5 Gestion des exceptions

La base de données dont on dispose contient un grand nombre de scenarios non représentatifs de l'exploitation en situation nominale. Les cas suivants ont été identifiés :

- Les jours d'incidents majeurs : dans ces cas-là, la gare est complètement bloquée, de nombreux trains sont supprimés et ceux qui circulent sont fortement retardés. Par ailleurs les jours qui ont suivi ces incidents ne correspondent pas à des situations nominales car la reprise du trafic se fait parfois en capacité réduite. Pour cette raison, les jours du 29 au 31 juillet 2017 et du 27 juillet 2018 au 3 août 2018, ainsi que les jours du 31 janvier et du 25 mars 2019 sont supprimés.
- Les jours de grève de 2018 : durant les grèves, très peu de trains sont en circulation et sont peu retardés car le réseau est en sous exploitation. Ces situations ne coïncident pas du tout avec la réalité des opérations. Les jours de grève entre mars et juin 2018 ne sont pas conservés.
- Les jours de neige extrême du 7 et 8 février 2018 sont supprimés de la base de données.
- Le mois d'août 2018 se caractérise par des taux de retards anormaux pour les TN et TER/IC. Les jours du 1ier au 18 août 2018 sont supprimés des bases d'entraînement, mais les tests sur cette période sont tout de même donnés à titre indicatif.
- Les données antérieures au 1ier juillet 2017 ne sont pas conservées car le réseau TGV a été très fortement modifié à cette date avec le prolongement de la LGV Atlantique jusqu'à Bordeaux et Rennes, ce qui a changé les dessertes, l'infrastructure est moins partagée et plusieurs temps de trajets ont fortement diminué. D'importantes différences dans les profils de retards avant et après cette date ont été constatées, ce qui risque de biaiser les modèles puisque les observations avant l'ouverture de la ligne ne sont plus pertinentes pour expliquer les retards futurs.

4.2.4 Gestion de la temporalité

L'influence temporelle est complexe à exploiter dans la modélisation. La temporalité d'une circulation fait partie des éléments de contexte permettant d'identifier des motifs susceptibles de se reproduire. La chronologie des trains n'est cependant pas exploitée directement, et va parfois à l'encontre de certaines hypothèses de modélisation, comme par exemple l'hypothèse d'indépendance et d'identité des distributions des retards sur laquelle peuvent être construits des modèles statistiques. Une autre difficulté concerne le choix des données d'apprentissage et de la méthodologie de sélection optimale. Le modèle doit être capable de prédire dans le futur, c'est-à-dire sur des données temporellement différentes de celles utilisées pour le construire. L'intégration du temps dans la méthodologie est détaillée ci-dessous.

4.2.4.1 Évolution des retards au cours du temps

Le fonctionnement général du réseau et du trafic est stable. Concernant l'horizon tactique, le plan de transport est construit selon un nombre restreint de scenarii prédéfinis. L'année est divisée en 4 grandes périodes (proches des saisons annuelles) et les mêmes journées types sont utilisées durant toute la période, à l'exception de certains jours particuliers (départs en vacances, jours fériés, etc.). Les missions et horaires des trains sont donc dépendants du jour de la semaine, de la période dans l'année et de certains éléments de contexte. Les décisions stratégiques, comme les choix des lignes et des fréquences et les modifications structurelles de l'infrastructure, sont rares et prévues très en amont. Dans les cas où il n'y a pas de modification de ce type, les plans de transport sont souvent repris d'une année à l'autre.

Malgré cette grande stabilité de fonctionnement, il n'est pas exclu que le réseau évolue légèrement au cours du temps. Par exemple certains travaux peuvent fragiliser le trafic pendant une période

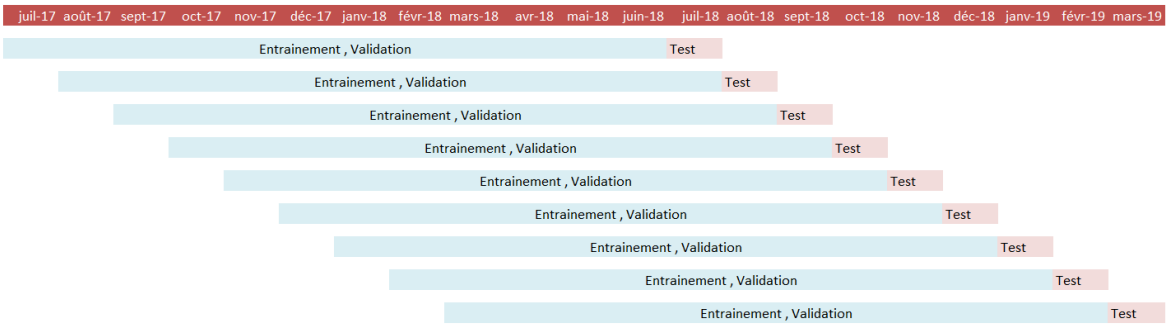

FIGURE 4.4 – Séparation temporelle des bases d'apprentissage et de test

donnée, ou au contraire le fluidifier une fois terminés, ce qui crée des changements imprévisibles dans les profils de retards. Certains changements de politiques locales peuvent également avoir un impact. Ces modifications ne peuvent pas être détectées avec les données dont on dispose, mais vont générer des similitudes plus fortes entre des évènements proches dans le temps.

Les modifications structurelles du réseau ou du plan de transport changent en profondeur les profils de retards et ne sont pas étudiées plus en détail. On recommande cependant de ne pas intégrer des données antérieures à un changement majeur. Cette décision a par exemple été prise sur ce cas d'étude en ne prenant pas en compte les observations précédant le prolongement de la LGV Atlantique.

4.2.4.2 Choix de représentation

Approche adoptée : on choisit de ne pas exploiter l'ordre chronologique entre les données dans la construction du modèle mais de représenter la temporalité via des variables explicatives corrélées au retard. Les probabilités de perturbations sont estimées conditionnellement au contexte temporel, via l'heure, le type de jour, les vacances ou encore la saison. L'évolution du réseau est intégrée en mettant régulièrement la base de données à jour. Ainsi, des changements légers dans le fonctionnement du réseau sont ajoutés au fur et à mesure, et des motifs qui ne sont plus pertinents disparaissent progressivement.

Plages d'étude : le modèle doit être construit sur une période donnée, puis être utilisé pour estimer des probabilités d'une période ultérieure pour laquelle on cherche à améliorer les planifications. On propose de travailler sur des prédictions menées à l'échelle d'un mois, ce qui est acceptable d'un point de vue industriel, mais assure tout de même une mise à jour régulière des données et des modèles.

Étant donné la périodicité sous-jacente, trois options ont été considérées. La première était de travailler sur la même période de prédiction mais de l'année passée, l'autre de construire le modèle sur les quelques mois précédents, pour intégrer uniquement les dernières tendances de retards sur le réseau, et enfin utiliser toute l'année précédente. Les trois options ont été testées mais c'est la dernière qui donnait les meilleures performances. Travailler sur une année entière assure d'avoir déjà rencontré les situations que l'on cherche à prédire (à la fois sur le même mois l'année précédente et sur les données plus récentes) et permet d'avoir plus de données ce qui stabilise les prédictions. Dans le cas d'un changement majeur dans les données, on recommandera la seconde approche le temps d'acquérir un an de données, et ensuite on appliquera de nouveau la troisième approche.

4.2.4.3 Bases d'apprentissage, de validation et d'évaluation

Chaque modèle est donc construit sur l'année qui précède le mois où le modèle sera utilisé pour prédire les risques de retard. Une difficulté concerne le choix du modèle optimal. En effet, pour chaque période deux modèles en compétition, un modèle linéaire généralisé (GLM) et une forêt aléatoire (RF), sont construits pour estimer les probabilités de retards. Le choix du modèle optimal doit se faire sur des données n'ayant pas servi construire le modèle afin d'éviter des phénomènes de surapprentissage.

Une option classique est de diviser la base de travail d'un an en une base d'apprentissage et une base de validation séparées dans le temps, par exemple apprendre sur 11 mois, valider sur le dernier, puis prédire sur le mois suivant (mois de test). Cette méthode a été exclue car elle privait l'apprentissage d'observations potentiellement proches de celles à prédire. Il est important que des motifs proches de ceux de la base à prédire aient déjà été rencontrés par le modèle, ce qui est compromis si on se prive des données les plus récentes. Les données sont donc séparées aléatoirement dans ces travaux : pour chaque base d'un an, 20% des observations sont mises de côté pour permettre le choix du modèle optimal. Cette stratégie ne permet donc pas de sélectionner le modèle en se basant sur sa capacité à généraliser dans le futur, mais permet de le juger de manière plus exhaustive sur sa capacité à représenter les différentes périodicités des retards.

Les performances prédictives des modèles créés sont étudiées également a posteriori sur les données du mois où ils sont appliqués (base de test). Ces tests ne peuvent pas être utilisés pour affiner la stratégie ou choisir le modèle étant donné que la base de test ne serait pas disponible dans des conditions réelles d'utilisation. Ils visent cependant à instaurer une confiance en la méthodologie en étudiant son fonctionnement au cours du temps. Deux aspects sont importants :

— Étudier la qualité et l'utilité des prédictions : il faut assurer que les modèles construits arrivent à extraire des données futures une information fiable qui puisse avoir une plus-value pour prendre des décisions.

— Assurer que la méthodologie est robuste : les modèles doivent parvenir à apprendre sur des données futures malgré l'évolution possible au cours du temps et avoir un comportement stable indépendemment de la période

Les données exploitées ici rassemblent les observations du 1ier juillet 2017 au 31 mars 2019, ce qui permet la formation de 9 périodes de test. Les différents découpages sont rappelés dans la figure 4.4. Une base d'apprentissage (base de validation exclue) sur une année compte environ 25000 TN, 18000 TGV et 7500 TER ou IC à l'arrivée ou au départ. Un mois de données utilisé en test sans suppression de dates perturbées correspond en général à 2500 TGV, 3300 TN et entre 850 et 1000 TER.

4.3 Analyses des retards

Une analyse préliminaire des retards de trains est proposée ici, avec tout d'abord des résultats concernant la modélisation des retards par des distributions de probabilité classiques, puis la représentation par des arbres de décisions et enfin des visualisations de la relation entre la ponctualité et certaines variables explicatives.

4.3.1 Distribution des retards

L'objectif de cette partie est d'identifier les distributions paramétriques classiques ayant la meilleure adhérence aux données de retard. On se place dans le cadre de travail précédemment décrit où les retards négatifs sont considérés comme nuls et les valeurs extrêmes dépassant le seuil de troncature lié au type de circulation sont supprimées. Les données sont partitionnées en 6 classes différentes : les TGV à l'arrivée, les TGV au départ, les TN à l'arrivée, les TN au départ, les TER Centre et Normandie

à l'arrivée et enfin les TER Centre et Normandie au départ. Chacune de ces classes correspond à une famille de circulations homogène tant au niveau des retards que de l'exploitation ferroviaire. Une valeur de troncature propre est associée à chaque classe.

La méthodologie appliquée à chacune de ces classes est la suivante : on génère les distributions discrètes tronquées candidates à l'aide du package *GAMLSS* en R [155, 163], la fonction *fitdist* est appliquée pour optimiser les paramètres de chaque distribution par maximum de vraisemblance sur les échantillons de retards de la période juillet 2017 à juin 2018, ces distributions sont ensuite comparées sur la base du critère d'information d'Akaike (AIC) et de la complexité de la distribution. En effet, une distribution ayant un trop grand nombre de paramètres est également plus difficile à manipuler, notamment pour entraîner des modèles linéaires généralisés. Les coefficients modélisant chaque paramètre de la loi sont estimés un à un en supposant les autres fixés (par exemple on entraîne les coefficients pour la localisation en supposant ceux d'échelle et de forme fixés) jusqu'à convergence de la vraisemblance. Les itérations entre chaque régression (cf 3.4) peuvent être longues et nombreuses, et ne pas converger en temps raisonnable.

(a) TGV à l'arrivée

Distributions	***df***	***AIC***
Zero Inflated Poisson Inv. Gaussian	3	64 700
Zero Inflated Negative Binomial	3	64 726
Negative Binomial	2	64 890
Poisson inverse gaussian	2	66 464
Zero inflated Poisson	2	77 187
Geometric	1	71 021
Poisson	1	115 355

(b) TGV au départ

Distributions	***df***	***AIC***
Zero Inflated Poisson Inv. Gaussian	3	24 711
Negative Binomial	2	24 702
Zero Inflated Negative Binomial	3	24 704
Poisson inverse gaussian	2	25 183
Zero inflated Poisson	2	26 475
Geometric	1	32 011
Poisson	1	43 144

(c) TN à l'arrivée

Distributions	***df***	***AIC***
Zero Inflated Poisson Inv. Gaussian	3	56 333
Zero Inflated Negative Binomial	3	56 336
Negative Binomial	2	56 374
Poisson inverse gaussian	2	56 798
Zero inflated Poisson	2	57 447
Geometric	1	59 738
Poisson	1	73 753

(d) TN au départ

Distributions	***df***	***AIC***
Zero Inflated Poisson Inv. Gaussian	3	23 002
Zero Inflated Negative Binomial	3	23 003
Negative Binomial	2	23 004
Poisson inverse gaussian	2	23 107
Zero inflated Poisson	2	23 194
Geometric	1	26 149
Poisson	1	29 327

(e) TER et IC à l'arrivée

Distributions	***df***	***AIC***
Zero Inflated Poisson Inv.Gaussian	3	24 253
Negative Binomial	2	24 289
Zero Inflated Negative Binomial	3	24 291
Poisson inverse gaussian	2	24 294
Zero inflated Poisson	2	27 530
Geometric	1	24 805
Poisson	1	32 987

(f) TER et IC au départ

Distributions	***df***	***AIC***
Zero Inflated Poisson Inv. Gaussian	3	12 239
Negative Binomial	2	12 244
Zero Inflated Negative Binomial	3	12 246
Poisson inverse gaussian	2	12 253
Zero inflated Poisson	2	12 646
Geometric	1	13 240
Poisson	1	15 602

TABLEAU 4.1 – Comparaison de l'adhérence des distributions sur les échantillons

La distribution obtenant globalement les meilleures performances est la loi binomiale négative notée NBI. Plus d'informations concernant cette loi peuvent être trouvées par la suite dans la sous-section 4.4.1.1. Les variantes avec inflation de zéro donnent parfois de meilleurs résultats mais sont écartées dans ces travaux en raison de la trop grande complexité de la loi qui a 3 paramètres. En effet, l'ajout

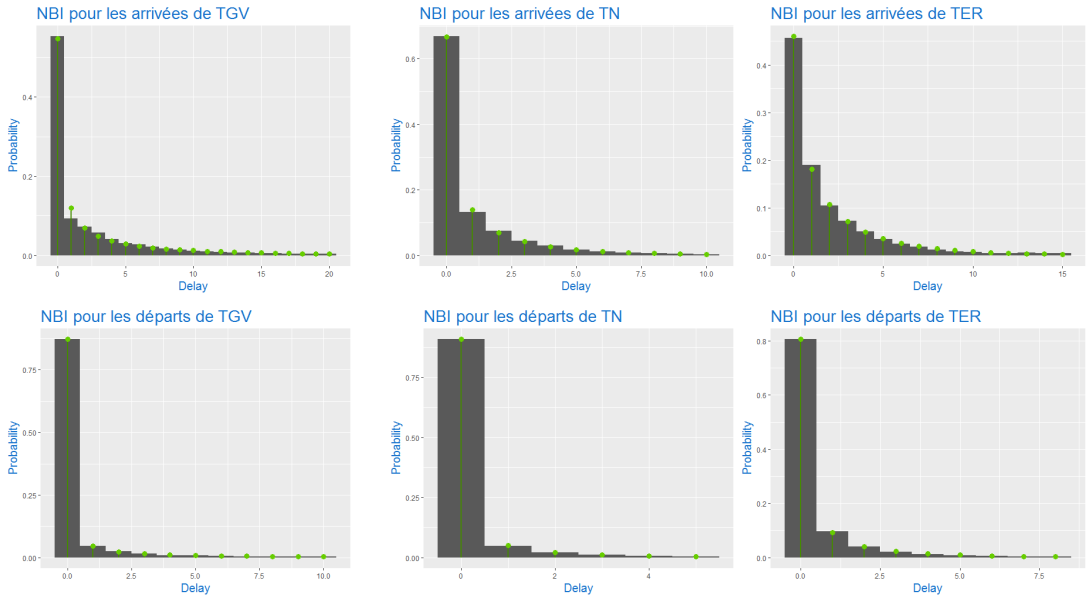

FIGURE 4.5 – Modélisation des retards par une loi Négative Binomiale (NBI)

d'un paramètre dans le modèle augmente considérablement les temps de calcul. Quelques tests ont été faits avec ces lois pour des GLM, cependant l'algorithme de parvenait pas à converger en temps raisonnable et rendait les étapes de sélections de variables impraticables, sans gain de performance net. Des expériences sur les temps de calcul en fonction de la complexité de la loi sont présentées en annexe C.1. La figure 4.5 montre graphiquement l'adhérence de la loi NBI aux différents types de circulations.

Cas des distributions continues : seules des distributions discrètes sont testées ici, contrairement à ce qui a été fait dans un grand nombre d'articles du domaine (cf 2.4.1.1). Plusieurs expérimentations préliminaires ont été faites sur les données de la gare Montparnasse avec des distributions continues dont une partie des résultats peut être trouvée dans les proceedings suivants [63, 64]. Dans le cas où seule une distribution continue était utilisée, il fallait prendre en charge les données nulles puisque les distributions comme la loi Weibull n'admettent pas de masse en zéro (cf 3.1.1.2). Les deux approches utilisées étaient l'ajout d'un bruit strictement positif aux données nulles [64] et la modélisation en deux parties [63]. Les résultats obtenus étaient cependant de bien moins bonne qualité qu'en utilisant des lois discrètes, en particulier avec la méthode du bruitage des données nulles. Plusieurs choses expliquent cela. Tout d'abord les données utilisées sont en minutes, la granularité est telle qu'utiliser une loi continue adhère mal aux données. Ensuite, le choix du bruit à ajouter sur les valeurs nulles a une influence très importante sur la distribution obtenue, même pour des variations infinitésimales. En effet, l'asymptote en zéro des lois candidates, comme la loi Weibull, accorde un poids très important aux valeurs proches de zéro dans la fonction de vraisemblance. Une variation de l'ordre du millième sur le bruit ajouté aux retards nuls n'est pas visible dans les données mais peut déséquilibrer les contributions à la vraisemblance des observations non nulles et nuire à la qualité du modèle.

4.3.2 Représentation par arbres de probabilité

Un moyen de visualiser facilement les variables explicatives ayant le plus d'impact sur le retard est d'utiliser des arbres de décision. Cette méthodologie est appliquée sur la base de données créée en section 4.2 en utilisant la fonction *ctree* du package partykit [101]. Les données sont séparées dans chaque noeud selon la variable et le seuil indiqués. La figure 4.6 représente l'arbre de probabilité construit à partir des données de retards à l'arrivée des trains de Montparnasse et la figure 4.7 ceux au départ. Les noeuds terminaux contiennent la fréquence relative de chaque occurrence de retard. Les retards sont ici tronqués à 15 minutes pour les arrivées et 10 minutes pour les départs, et les valeurs négatives sont considérées comme nulles.

Cette représentation des données présente plusieurs avantages. Le premier est qu'elle permet de générer automatiquement des règles expertes simples. Par exemple on peut estimer à l'aide de l'historique des retards que les trains Transiliens n'arrivant pas le matin et avec un temps d'arrêt maximal planifié compris entre 3 et 5 minutes incluses ont un taux de trains à l'heure (0 minute de retard) de 80%, ce qui correspond au chemin par les noeuds 1, 17, 16, 22 et 24 de la figure 4.6. Le second est qu'elle met en évidence les critères et points de séparation les plus discriminants pour l'analyse de la ponctualité.

Dans ce cas-ci, les données sont séparées très rapidement selon le type de circulation (Transilien, TGV, TER), ce qui confirme l'importance de modéliser les retards séparément. Les variables liées aux heures de pointe (densité ou plage horaire du matin par exemple) et les variables liées à la mission (nombre d'arrêt, origine, temps d'arrêt) sont très utilisées.

4.3.3 Influence des variables explicatives

Cette partie présente une analyse préliminaire sur le rôle des variables sélectionnées. Seuls quelques variables et types de mouvement sont présentés. L'objectif est de justifier l'importance de mener une étude séparant les types de mouvement et d'établir la relation entre certaines variables et le risque de retard.

Densité du trafic : elle peut être mesurée en ligne, en comptant l'écart en minutes avec le train précédent et le train suivant sur la même ligne, par exemple ligne Bordeaux pour tous les trains allant jusqu'à Bordeaux et au delà. Elle est également évaluée en gare en comptant le nombre de trains en circulation (arrivée ou départ) dans la gare considérée. Les variables les plus utilisées sont *densitePrec20* et *densiteSuiv5* (ou *densiteSuiv15*) qui comptent respectivement le nombre de train en mouvement à Montparnasse dans les 20 minutes précédant l'horaire du train ou dans les 5 minutes (ou 15 minutes) le suivant. D'autres variantes mesurées à l'origine ou dans les différentes gares visitées sont possibles.

Un exemple est donné dans la figure 4.8 pour les trois types de trains et les deux types de mouvement, avec en abscisse le nombre de trains en circulation 5 minutes après le train observé et en ordonnée la proportion de trains ayant un retard non nul. On observe une corrélation nette entre les deux variables, mais qui dépend fortement des données considérées. Par exemple pour les arrivées des TN, le taux de trains retardés augmente presque proportionnellement à la densité, alors qu'il diminue pour les trains régionaux.

Mission : une partie du retard peut s'expliquer par les caractéristiques de la mission. L'élément principal est l'origine du train puisqu'elle conditionne la distance parcourue, la durée du parcours, parfois même l'heure puisque certaines gares ne sont desservies qu'à certains moments de la journée.

Le graphique 4.9 présente des variations des $50^{\text{ème}}$, $70^{\text{ème}}$ et $90^{\text{ème}}$ percentiles du retard, calculés sur la base complète non tronquée, en fonction de l'origine et du type de circulation, comme c'est le cas

pour Rambouillet. Le choix de représenter des percentiles plutôt que la moyenne et l'écart type vient du fait que les données contiennent beaucoup d'exceptions, comme des trains avec plusieurs heures de retard, qui pourraient fausser l'interprétation des résultats, contrairement aux percentiles qui sont plus stables.

Certaines tendances sont propres au type de train, par exemple les Transiliens sont largement plus ponctuels que les trains grande vitesse et trains régionaux avec une médiane toujours nulle et des retards extrêmes ($90^{\text{ème}}$ percentile) inférieurs à 5 minutes. Sans surprise, les trains grande vitesse sont ceux dont les retards extrêmes sont les plus importants, en particulier pour les gares se situant à une grande distance de Paris comme Toulouse, Hendaye et Tarbes.

La mission peut également se caractériser par d'autres variables, comme la durée du trajet, la marge (excès de temps de trajet par rapport à la durée habituelle pour une même desserte), le nombre d'arrêts, le temps d'arrêt planifié, etc.

Variables temporelles : la figure 4.10 montre la ponctualité (seuil de ponctualité à 0 minute) en fonction de la valeur d'indicateurs binaires liés à la temporalité de la circulation. Des variations dans la ponctualité peuvent être observées, et là encore les résultats dépendent du type de train et de son sens. Par exemple, les trains pour lesquels la variable *MATIN* vaut 1 (arrivée entre 7h et 10h) ont près de 30% de risque supplémentaire d'être en retard si le train est un Transilien ou un TER alors que dans le cas des TGV la différence de proportion de trains retardés le matin est de moins de 10%.

Conclusion : les relations entre retards et les différentes variables explicatives sont en général complexes et difficilement interprétables. Quelques motifs peuvent cependant être identifiés à la main par des graphiques simples. Les relations trouvées doivent cependant être interprétées avec précaution. Ces variables explicatives sont proposées car qu'elles représentent des indicateurs d'instabilité du réseau, comme les heures de pointe ou la densité, mais elles ne permettent pas de prouver des relations de causalité.

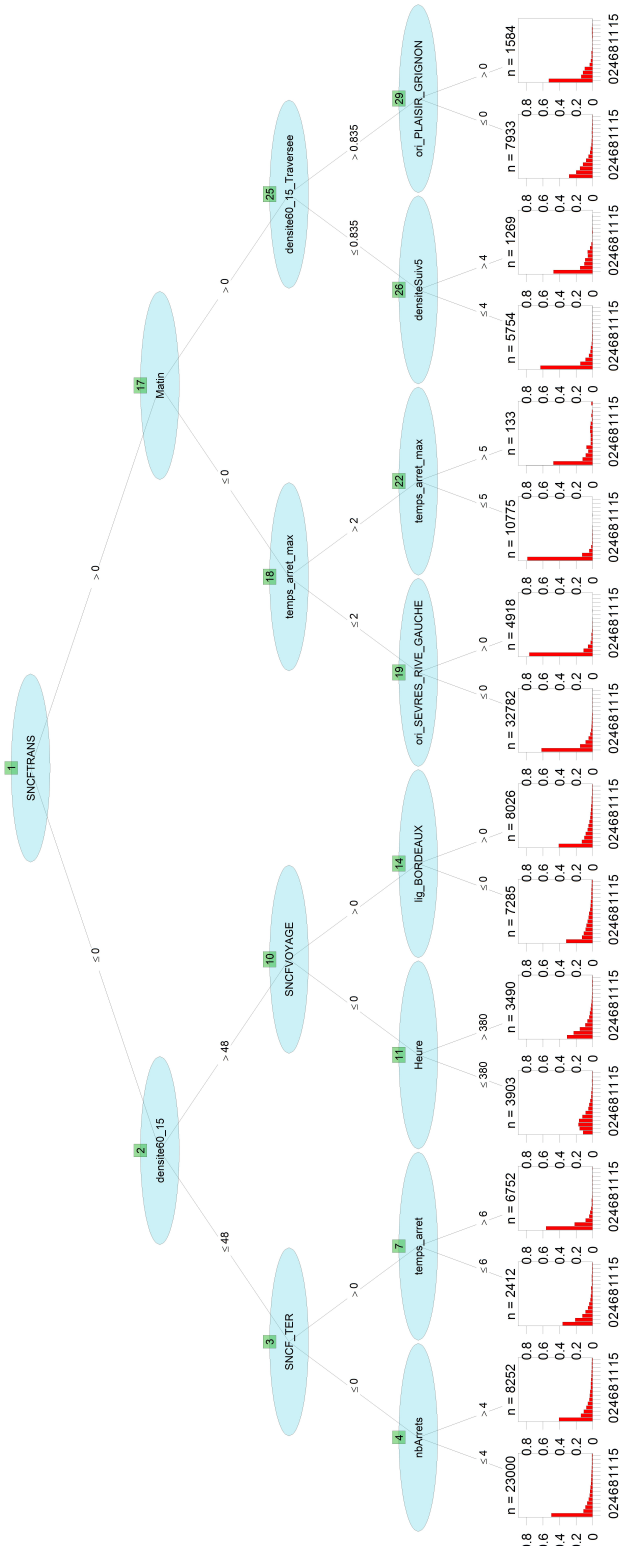

FIGURE 4.6 – Arbre de probabilité des données de retard à l'arrivée

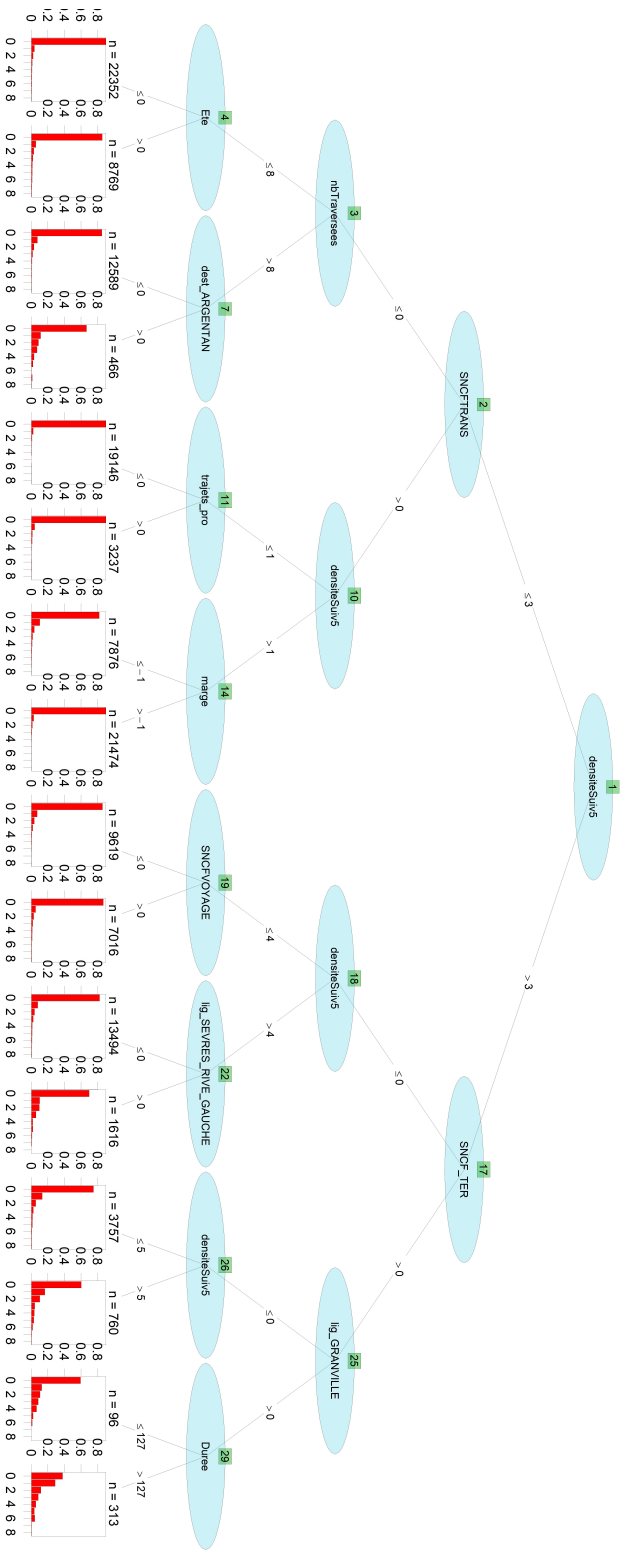

FIGURE 4.7 – Arbre de probabilité des données de retard au départ

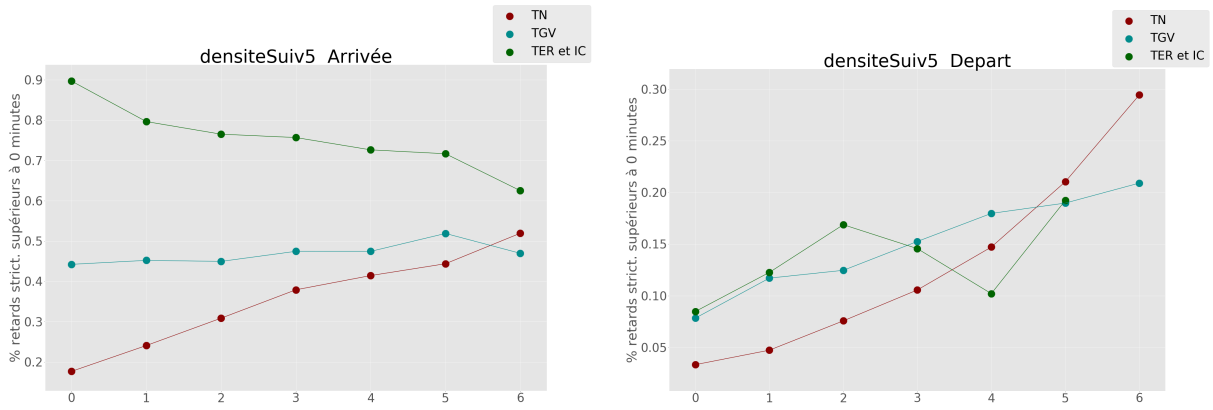

FIGURE 4.8 – Densité en gare et retards

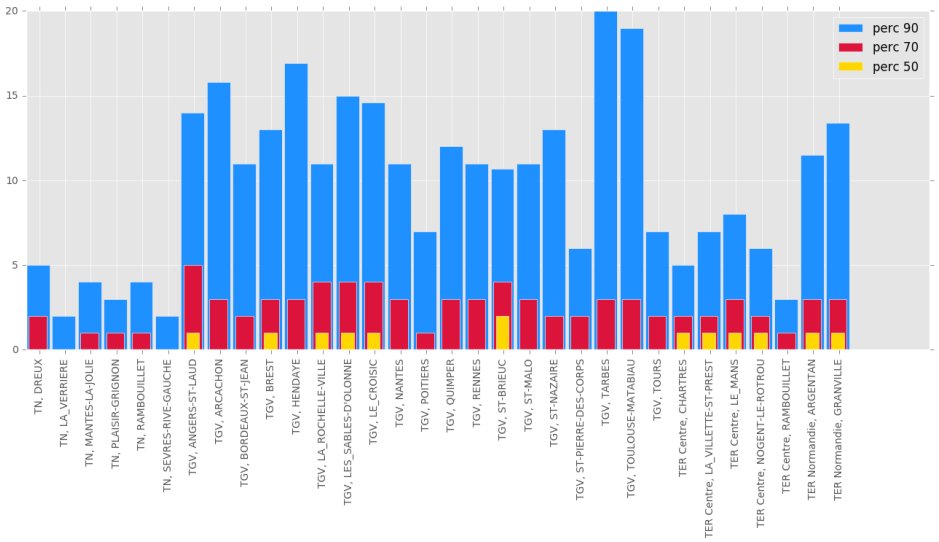

FIGURE 4.9 – Percentiles du retard selon l'origine

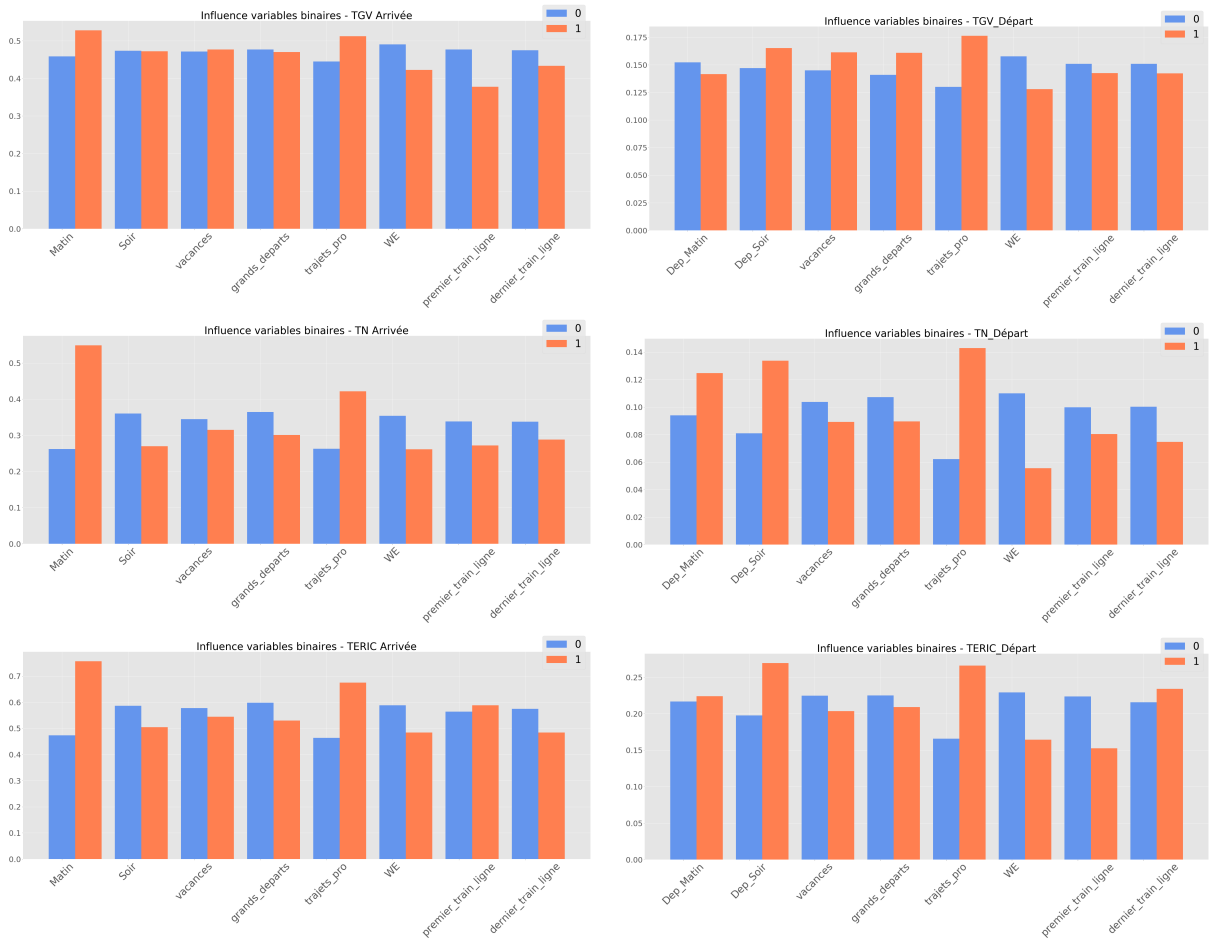

FIGURE 4.10 – Relation aux indicateurs temporels

4.4 Construction du modèle optimal

Cette section décrit les étapes de construction des deux modèles en compétition pour estimer les distributions de probabilités de retard conditionnellement aux variables explicatives introduites précédemment. On traitera en particulier des stratégies d'encodage et de sélection des variables, de l'hyperparamètrage choisi, de l'étape de validation interne pour limiter le surapprentissage et enfin les pistes d'amélioration des modèles.

4.4.1 Modèles Linéaires Généralisés

Comme expliqué dans le chapitre précédent, un GLM est composé de plusieurs briques, à savoir une variable de réponse dont on fait l'hypothèse qu'elle suit une distribution paramétrique connue, un ensemble de variables explicatives et une fonction de lien qui permet de connecter ces variables aux paramètres de la loi de réponse. Ces différentes composantes sont détaillées ici. Une attention particulière est apportée au choix des variables explicatives.

4.4.1.1 Architecture du modèle

La loi négative binomiale : comme cela a été attesté plus haut, la meilleure loi candidate est la loi négative binomiale. Dans GAMLSS [155], le paramétrage de cette loi est :

$$\mathbb{P}[Y = y|\mu, \sigma] = \frac{\Gamma\left(y + \frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)\Gamma\left(y+1\right)}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^{y}\left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma} \tag{4.1}$$

pour $y = 0, 1, 2, .., \mu > 0$ et $\sigma > 0$

Cette loi est intéressante car elle est plus adaptée pour modéliser des cas de surdispersion ou de sousdispersion que des lois plus classiques comme la loi de Poisson [97]. Une difficulté se pose cependant pour l'interprétation des résultats, en effet, les probabilités prédites ne sont pas une fonction monotone des variables comme pour une loi exponentielle et les variables interviennent à la fois dans μ et dans σ. L'impact d'une variable explicative donnée sur la prédiction est donc difficilement quantifiable.

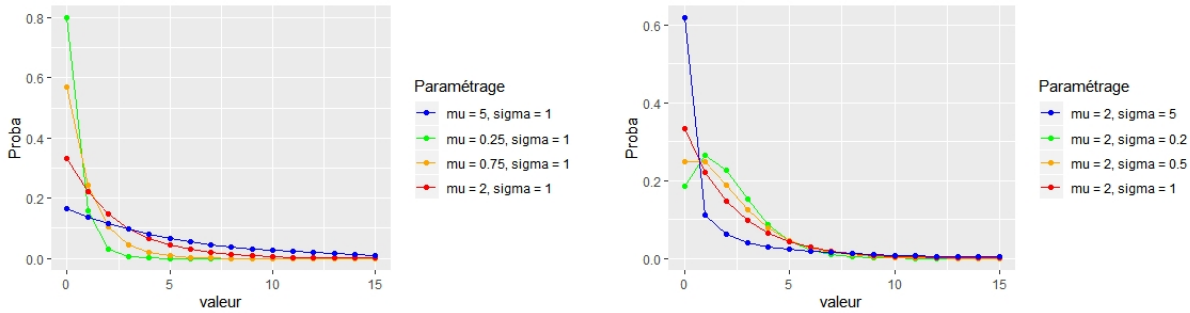

FIGURE 4.11 – Loi négative binomiale

Modèle GAMLSS : dans la méthodologie GAMLSS, la fonction de lien utilisée pour des paramètres de loi strictement positifs est le logarithme. Le modèle linéaire généralisé qui est entraîné ici est de la forme :

$$\begin{aligned} Y &\backsim \mathcal{NBI}_{tr}(\mu, \sigma) \\ log(\mu) &= X_\mu \beta_\mu \\ log(\sigma) &= X_\sigma \beta_\sigma \end{aligned} \tag{4.2}$$

avec Y l'échantillon de retards tronqués de l'ensemble d'entraînement, X_μ et X_σ les matrices covariées et $\mathcal{NBI}_{tr}$ la loi de distribution négative binomiale tronquée à droite.

4.4.1.2 Sélection et encodage des variables

Encodage des données : les bases de données utilisées pour l'approche par GLM et l'approche par forêts aléatoires sont majoritairement identiques à l'exception de l'encodage de certaines variables. Dans le cas d'un GLM, la relation entre le retard et les autres variables est modélisée via une interaction linéaire entre ces variables et les paramètres de la loi. Pour cette raison, les modifications suivantes sont apportées à la base de données :

- la variable *Heure* n'est pas conservée, mais est indirectement représentée par des variables binaires symbolisant les plages horaires
- la variable *Numéro* qui est catégorielle mais avec un encodage numérique n'est pas gardée.
- les variables numériques continues sont standardisées (durées, temps d'arrêts, marges, variables de densité, ...). Cela permet notamment de pouvoir interpréter plus directement les valeurs des coefficients dans les régressions du GLM et les comparer entre eux pour étudier l'impact des variables correspondantes.

Nécessité de la sélection de variables : il est généralement conseillé de sélectionner les variables explicatives pour ne conserver que celles apportant le plus d'information sur la variable cible. La simplicité du modèle permet de faciliter l'interprétation des résultats et de réduire les temps de calcul pour l'entraînement et la prédiction. Utiliser un sous-ensemble de variables permet aussi de réduire la variance d'un modèle linéaire et de minimiser le surapprentissage. C'est particulièrement vrai quand l'information est redondante entre les différentes variables.

Guyon et Elisseff [87] rappellent différentes stratégies classiques pour sélectionner les variables. La première est le classement des variables selon un critère comme la corrélation à la cible, et la sélection des meilleures variables. Cela nécessite peu de temps de calcul mais risque de sélectionner des variables corrélées entre elles. Les approches gloutonnes par sélection pas-à-pas (ou *stepwise*) sont classiquement utilisées : à chaque étape, plusieurs modèles sont construits en ajoutant ou enlevant une variable au modèle courant, puis la variable qui aura donné le meilleur modèle est sélectionnée. Le modèle est mis à jour avec cette variable, et l'opération est réitérée jusqu'à convergence ou critère d'arrêt. Certains algorithmes et modèles d'apprentissage sélectionnent naturellement les variables lors de l'entraînement, et le sous-ensemble utilisé peut être récupéré. C'est par exemple le cas avec des arbres de décision CART ou une régression LASSO. Enfin, on peut choisir une représentation plus optimale des données : plutôt que choisir un sous-ensemble des variables, on va chercher à réduire la dimension en transformant les variables. Il est possible de procéder par combinaison des features et réduction matricielle de la dimension, comme avec une analyse en composantes principales, ou encore par clustering en remplaçant plusieurs variables similaires par un cluster.

On propose ici d'utiliser des approches gloutonnes pour choisir les variables à utiliser. Le but de la procédure est de produire un sous-ensemble de features construit itérativement et qui maximise la vraisemblance en pénalisant le nombre de degrés de liberté du modèle, c'est-à-dire le nombre de coefficients à estimer.

Les autres techniques sont en comparaison moins attractives, bien que également plus rapides à mettre en place. En effet, le classement des variables est déconseillé ici car plusieurs variables sont très corrélées entre elles, par exemple la durée du trajet et le nombre de gares visitées, et ne prendre que celles qui ont le plus de poids individuellement risque de ne sélectionner que des variables ayant un rôle similaire. L'utilisation d'algorithme CART ou LASSO pour isoler un sous-ensemble de variables est pertinent mais pourrait ne pas prendre en compte correctement la façon dont les variables interagissent avec la cible, en particulier il n'y a pas de distinction entre le rôle pour les paramètres de forme et de localisation. Cette démarche est plus rapide mais moins exhaustive que les sélections pas-à-pas. Utiliser une analyse en composante principale permettrait par exemple de regrouper les variables les plus corrélées entres elles, mais n'est valable que pour les variables non binaires qui sont peu nombreuses ici. Par ailleurs la réduction de la dimension et la transformation de variables font aussi perdre en interprétabilité.

Méthodologie utilisée : on travaille sur la base d'apprentissage complète en partant du modèle nul (aucune variable ni pour μ ni pour σ) et à chaque itération on ajoute la variable donnant le modèle avec la meilleure amélioration du critère BIC (Critère d'information bayésien $BIC = -2l - k\log(n)$, avec l la logvraisemblance, n le nombre d'observations et k le nombre de degrés de liberté du modèle). Le critère BIC permet de contraindre la complexité du modèle.

En utilisant une distribution à plusieurs paramètres, comme la binomiale négative, la sélection doit se faire sur chacun des paramètres. Étant donné les temps de calcul pour le modèle complet, on opte pour une sélection *forward*, en ajoutant les variables unes à unes sans possibilité de supprimer des variables déjà présentes. Trois stratégies sont possibles :

— effectuer les sélections sur chacun des paramètres jusqu'à convergence : toutes les variables de μ sont choisies jusqu'à ce qu'on ne puisse plus en ajouter, puis on procède de même sur σ. On recommence jusqu'à ce qu'aucun ajout ne soit possible ni sur μ ni sur σ.

— ajouter les variables une à une pour chacun des paramètres : on cherche une variable à ajouter pour μ, puis une pour σ, et on réitère jusqu'à convergence.

— ajouter la même variable à chaque étapes aux formules de μ et σ jusqu'à convergence.

Au vu des expérimentations menées, la troisième option n'est pas à privilégier. Elle permet d'accélérer la procédure puisqu'une seule recherche est menée, cependant les ensembles de variables sélectionnées sont souvent très différents selon le paramètre représenté. Les deux premières approches ont été testées, mais on privilégie la seconde qui construit les deux ensembles de variables au fur et à mesure. Elle permet d'équilibrer les formules de μ et de σ, et d'intégrer l'information là où elle est la plus utile. La première approche va tendre à saturer l'ensemble de variables modélisant μ, quitte à intégrer des variables qui amélioreraient plus le modèle si elle était ajoutées à l'ensemble de σ.

Cette procédure rentre dans le cadre de validation interne : on sélectionne le modèle (dont ses variables) parmi un ensemble de candidats. Le critère BIC se compose d'un terme évaluant l'efficacité du modèle (score logarithmique) et d'un terme qui en pénalise la complexité. Cette pénalité permet de limiter les effets de surapprentissage.

Le tableau 4.2 donne les formules obtenues pour les estimations de μ et σ pour les différentes bases de données en entraînant sur juillet 2017 à juin 2018, ainsi que le temps de calcul. Les temps de calcul sont assez stables d'une période à l'autre, cependant les ensembles de variables peuvent évoluer, étant donné que de nouveaux motifs peuvent apparaître au fur et à mesure que des données sont ajoutées à la base d'apprentissage.

La figure 4.12 liste les variables sélectionnées pour chaque période pour μ et pour σ pour les TGV et TN à l'arrivée, avec l'ordre d'ajout dans le modèle. Les variables avec la plus grosse contribution à la vraisemblance sont ajoutées en premières. On constate que les variables sélectionnées sont bien

set	temps	μ	σ
TN arr	3h31	$densite20\_5\_Traversee + temps\_arret\_max + marge + Dep\_Matin + densiteSuiv5 + SEVRES\_RIVE\_GAUCHE + grands\_departs + temps\_arret + PLAISIR\_GRIGNON + LA\_VERRIERE + trainSuivLigne + Matin + mardi + lundi + nbArrets + RAMBOUILLET + nbArrets : RAMBOUILLET + nbTraversees + trajets\_pro$	$Matin + SEVRES\_RIVE\_GAUCHE + temps\_arret + marge + PLAISIR\_GRIGNON + temps\_arret\_max + densite20\_5\_Arret + DREUX + densite20\_5\_Traversee + nbArrets + Hiver + DREUX : nbArrets + temps\_arret\_min + densitePrec20 + densiteSuiv5 + nbTraversees + vacances + Journee + X7200$
TGV arr	2h08	$Duree + Dep\_Matin + vendredi + Dep\_Journee + POITIERS + marge + TOURS + densitePrec20 + Matin + trainPrecLigne + Automne$	$trajets\_pro + LA\_ROCHELLE\_VILLE + nbTraversees + densite20\_5\_Arret + densite20\_5\_origine + TGV\_D + marge + Dep\_Matin + trainPrecLigne + LES\_SABLES\_D\_OLONNE + ST\_BRIEUC$
TER arr	37min	$densitePrec20 + temps\_arret + X7200 + densite20\_5\_Arret + vacances + GRANVILLE + NOGENT\_LE\_ROTROU + Automne + marge + densite20\_5\_Traversee + Nuit + trajets\_pro + Dep\_Matin + nbTraversees + temps\_arret\_max + Dep\_Nuit + jeudi + trainPrecLigne$	$Matin + temps\_arret + densite20\_5\_Traversee + nbTraversees + Dep\_Nuit + trainSuivLigne + Dep\_Matin + Ete + CHARTRES + Dep\_Journee + WE + dernier\_train\_ligne + ARGENTAN + marge$
TN dep	1h15	$densiteSuiv5 + marge + MANTES\_LA\_JOLIE + mardi + Dep\_Journee + densiteSuiv15 + WE$	$densiteSuiv5 + marge + trajets\_pro + densitePrec60 + Dep\_Nuit + Dep\_Journee + Hiver$
TGV dep	2h32	$densiteSuiv15 + Duree + Matin + TGV2N2$	$Ete + densiteSuiv15 + grands\_departs + Automne$
TER dep	22min	$ARGENTAN + densiteSuiv15 + GRANVILLE + trajets\_pro$	$densiteSuiv5 + ARGENTAN + trainSuivLigne + Hiver$

TABLEAU 4.2 – Exemple de sélection de variables

plus stables pour les TN, avec conservation de l'ordre d'ajout pour les plus importantes, d'une période à l'autre alors que pour les TGV il y a moins de variables sélectionnées et elles ne sont pas ajoutées systématiquement. Les seules variables TGV ajoutées à toutes les périodes sont *Duree* et *Dep_Matin* pour μ et *LA_ROCHELLE_VILLE* et *nb_Traversees* pour σ. Pour les TN il y en a une vingtaine.

4.4.1.3 Perspectives d'amélioration

Ajout de termes non linéaires : le package GAMLSS offre la possibilité d'ajouter des termes non linéaires dans les équations de μ, σ, ν, τ. Ces termes permettent de représenter le paramètre comme une combinaison de fonctions des variables, et non plus seulement une combinaison linéaire de ces variables. Les fonctions proposées sont par exemple le spline cubique, des fonctions polynomiales ou encore des effets aléatoires. Quelques tests ont été menés, cependant les termes non linéaires requièrent plus de coefficients à estimer, et n'apportent quelque chose que si la relation entre la cible et la variable explicative est mieux modélisée par ce nouveau terme.

Augmentation du nombre de paramètres : les performances peuvent éventuellement être améliorées en utilisant un modèle plus complexe qui représente mieux les données. En particulier dans le cas de données avec une forte composante nulle, les modèles à inflation de zéros et modèles en deux parties peuvent permettre une meilleure adhérence aux données (cf 3.1.1.2).

Sur les données de retards de trains, les modèles à inflation de zéros posent cependant une limite de complexité puisque 3 paramètres doivent être optimisés en parallèle, ce qui est plus long à calculer. En pratique, ces modèles ne convergent pas sur des échantillons de retards testés (plusieurs dizaines de milliers d'observations). Les résultats d'expérimentations sur la relation entre la complexité de la loi et le temps de calcul sont donnés dans l'annexe C.1.

Les modèles en deux parties sont plus rapides à optimiser : la première partie est une régression logistique simple entraînée sur la base complète. L'apprentissage est rapide et sans problème de convergence. La seconde partie est ici une loi négative binomiale entraînée sur l'échantillon tronqué en zéro, ce qui dans notre cas permet de réduire de presque de moitié le nombre d'observations à traiter. Quelques résultats obtenus sur les TGV avec des modèles en deux parties peuvent être trouvés ici [63]. Dans les expériences menées, le modèle en deux parties obtenait cependant des résultats un peu moins bons que le modèle simple, mais la sélection de features pas à pas était fortement accélérée par le caractère séquentiel de la modélisation. Des tests gagneraient à être faits pour étudier cette stratégie sur les autres types de circulation.

4.4.2 Modélisation par Forêts aléatoires

4.4.2.1 Construction de la forêt

Encodage des données : contrairement aux modèles linéaires généralisés, les arbres de décisions ne modélisent pas des interactions linéaires mais fonctionnent par séparations successives des données. L'encodage est moins important ici. En particulier, les variables continues ne sont pas standardisées et la variable *Heure*, mesurée en minutes, est conservée en plus des encodages de plages horaires. En effet, on peut s'attendre à ce que l'algorithme de forêt aléatoire sépare automatiquement la variable pour isoler de nouvelles plages. Dans le cas de la variable *Jour*, les valeurs peuvent être ordonnées. On utilise un encodage numérique de 1 à 7, commençant à lundi et terminant à dimanche.

Implémentation : la forêt est construite en utilisant le classifieur *RandomForestClassifier* dans le package *sklearn* sous python [147]. Les bases d'apprentissage utilisées pour construire les forêts sont les mêmes que pour les GLM. Le critère de séparation utilisé est l'entropie, c'est-à-dire qu'à chaque noeud, la variable et le seuil choisis sont ceux qui maximisent la vraisemblance en utilisant les fréquences relatives de chaque valeur contenues dans les noeuds fils comme estimation des probabilités.

Hyperparamétrage : les forêts aléatoires imposent de choisir les hyperparamètres du modèle, en particulier le nombre de variables à sélectionner aléatoirement à chaque séparation et la complexité de l'arbre. L'hyperparamétrage est mené en étudiant plusieurs combinaisons de valeurs pour le nombre de variables m et la taille minimale des noeuds terminaux d. Cette étape n'est faite qu'une fois par type de mouvement, en se basant sur les données de juillet 2017 à juin 2018. Les valeurs choisies sont celles donnant les meilleures performances en validation croisée *k-fold*, c'est-à-dire en partitionnant la base d'apprentissage en k échantillons et en construisant k modèles chacun entraîné sur une base formée de $k-1$ échantillons et évalué sur le dernier.

Cette étape est implémentée en utilisant la fonction *GridSearchCV* de la librairie *sklearn*, avec comme fonction de score l'entropie (score logarithmique) avec une validation croisée avec $k = 10$ et des forêts de 500 arbres. Les paramètres recommandés sont indiqués dans le tableau 4.3.

Validation interne : pour chaque période, une validation croisée avec $k = 10$ est également utilisée pour construire plusieurs modèles alternatifs. La forêt finale est celle qui atteint la meilleure erreur (entropie) sur les données isolées de l'apprentissage.

Mouvement	Nombre variables total	Nombre de variables m	Taille des feuilles d
TGV arrivée	63	30	20
TN arrivée	52	20	15
TER/IC arrivée	53	25	30
TGV départ	54	30	10
TN départ	45	15	10
TER/IC départ	43	25	30

TABLEAU 4.3 – Hyperparamétrage conseillé

4.4.2.2 Importance des variables

Les forêts aléatoires permettent l'estimation de l'importance des variables utilisées dans la construction des arbres. Cette mesure d'importance repose sur le calcul à chaque séparation de l'apport de la variable explicative au critère d'optimisation si elle était utilisée pour séparer le noeud [147].

Un défaut de cette approche est cependant qu'elle favorise les variables continues. En effet, une variable binaire, même très importante, ne pourra être utilisée qu'une fois pour séparer les variables dans un arbre alors qu'une variable continue sera utilisée plusieurs fois et aura donc potentiellement plusieurs apports non nuls au critère. Par ailleurs, les variables, en particulier binaires, sont parfois redondantes, et si elles ont déjà été séparées selon une variable, l'autre devient potentiellement inutile et verra donc son importance diminuer. C'est particulièrement vrai pour les différentes catégories d'une même variable avant encodage, l'arbre les considère comme indépendantes : si la variable *Matin* est utilisée tôt dans l'arbre pour séparer, les variables *Journée* et *Soir* auront une contribution nulle dans toute la partie de l'arbre contenant les observations circulant le matin.

L'importance des variables moyennée sur les 9 périodes est donnée dans les figures 4.13 et 4.14.

4.4.2.3 Perspectives d'amélioration

L'algorithme de forêts aléatoire est flexible et nécessite très peu d'hypothèses sur la forme des données à représenter. De fait, il existe peu de pistes d'amélioration du modèle. Les deux plus prometteuses sont :

— Modifier la structure du modèle : les retards sont traités ici comme des classes indépendantes, le problème peut être vu comme une régression ou une classification de variables ordinales (cf 3.1.2.4). La prise en compte de l'ordre entre les valeurs peut apporter en précision. Ce point est cependant à vérifier, a priori la probabilité de retard d'un train est décroissante passé la mode, cependant il n'est pas exclu que certains motifs dans les retards soient associés à des valeurs précises. Traiter les classes comme indépendantes permet de détecter ce cas là. On peut également imaginer des approches mixtes où plusieurs parties de la forêts seraient construites avec des structures différentes, permettant de à la fois tirer parti de l'ordre entre les valeurs et d'identifier des motifs non ordonnés.

— L'encodage des variables catégorielles peut être revu pour apporter plus d'information. Ici un encodage *One Hot* est utilisé, c'est-à-dire qu'il existe une variable binaire pour chaque catégorie, comme les jours de la semaine, les origines des trains, etc. avec parfois en complément un encodage numérique naturel (la variable *Numéro* et la variable *Jour*). Un encodage *One Hot* donne parfois de moins bons résultats avec des algorithmes d'arbres [1]. Par exemple si le nombre de catégories est élevé, l'arbre doit aller profondément pour séparer ses variables et est très asymétrique, ensuite un grand nombre de variables augmente les temps de calcul. Un encodage numérique, en numérotant arbitrairement chacune des classes, est parfois préférable. Une approche recommandée est de numéroter les catégories en fonction de la proportion de succès

ou de la valeur moyenne des observations appartenant à cette catégorie, ce qui aide le modèle à séparer optimalement la nouvelle variable encodée numériquement [74]. Quelques tests ont été menés en utilisant ce nouvel encodage sur l'origine des trains, cependant les performances sont équivalentes. Un examen plus approfondi de l'encodage idéal serait à faire. L'algorithme pourrait également gagner en ne dupliquant pas l'information comme cela est fait ici, puisqu'à chaque étape des variables sont sélectionnées aléatoirement. Trop de variables binaires alors qu'une variable presque équivalente continue existe risque de noyer l'information. Enfin, le choix de la variable *Numero* est à considérer avec vigilance : elle contient une information riche car un numéro de train est associé à un trajet spécifique (desserte, horaire, période ou jours de circulation) mais les numéros de train sont susceptibles de changer, en particulier à la fin du service annuel en décembre.

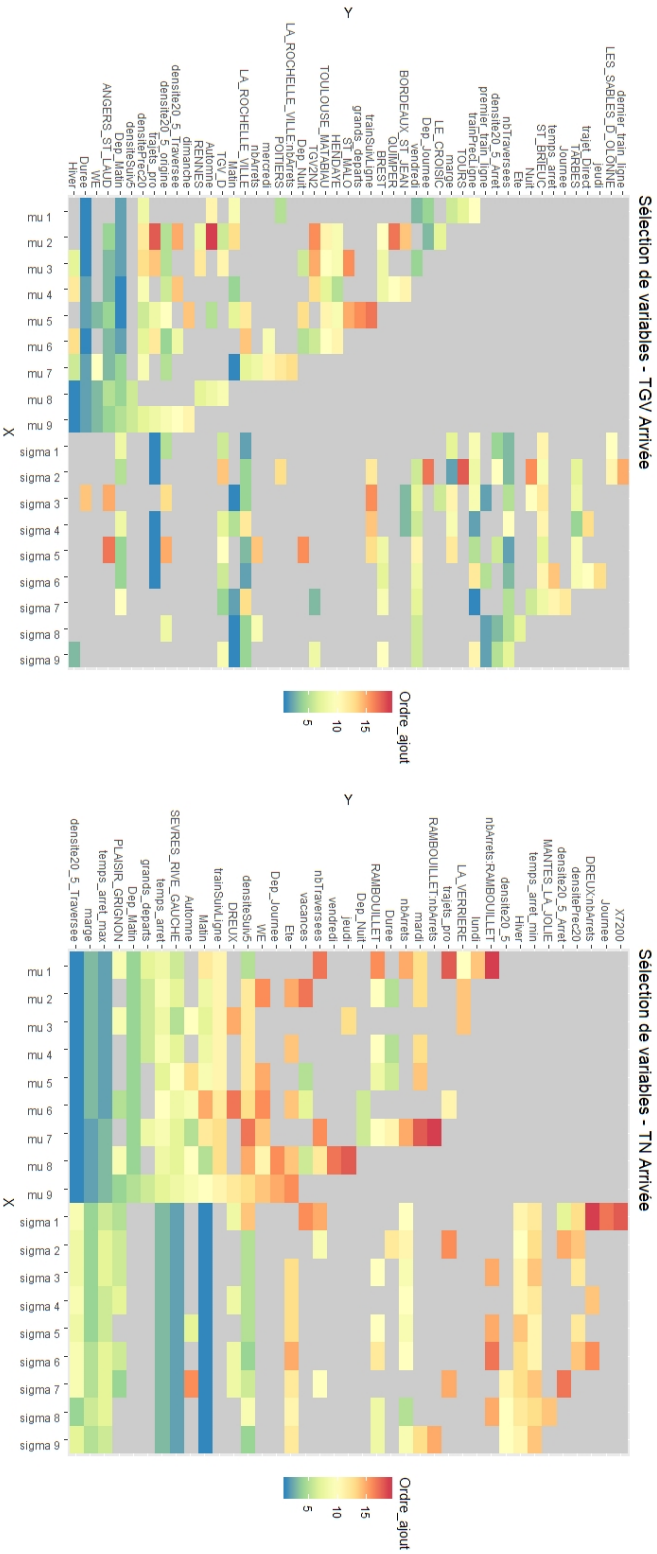

FIGURE 4.12 – Variables sélectionnées - TGV et TN arrivée

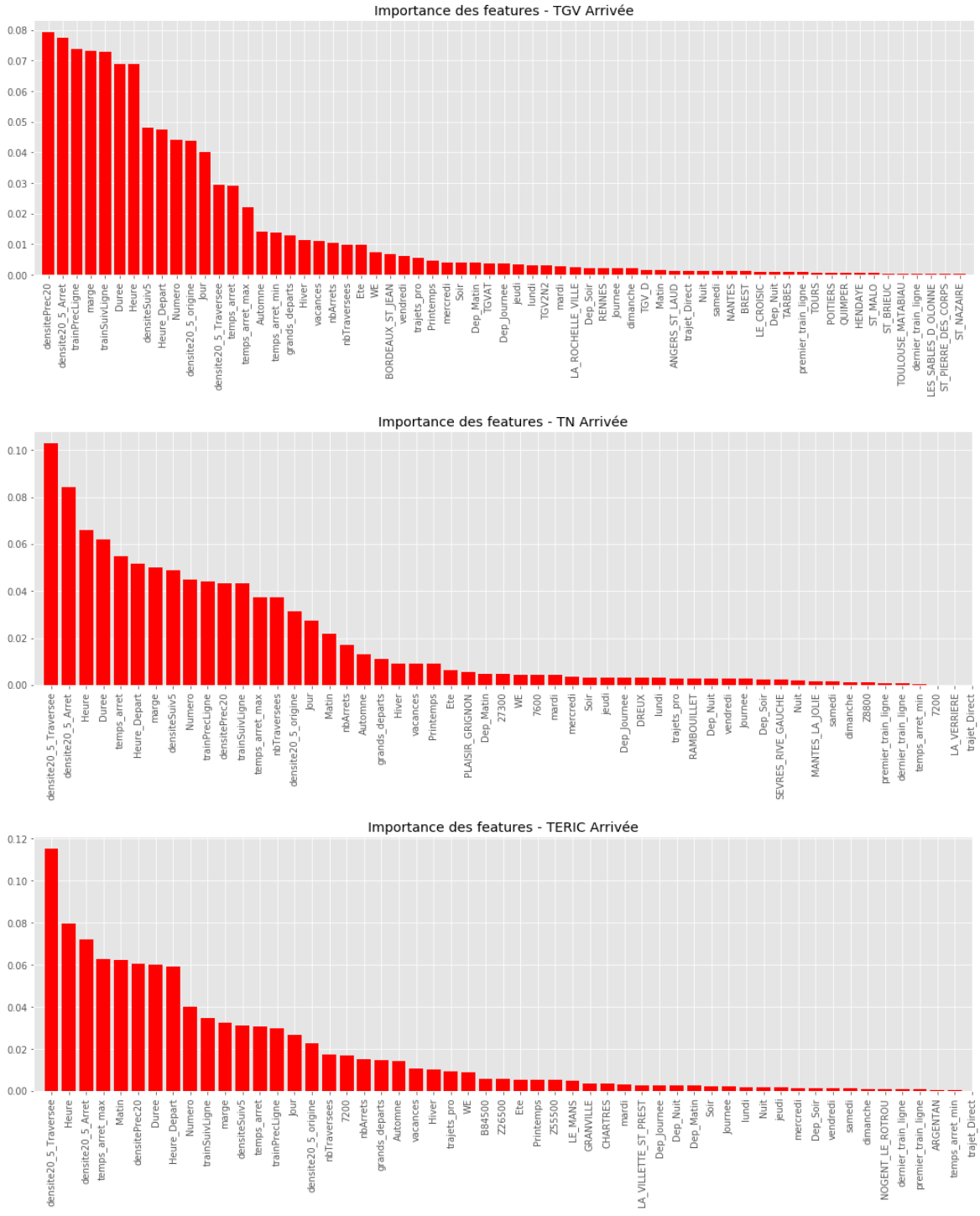

FIGURE 4.13 – Importance des features - arrivée

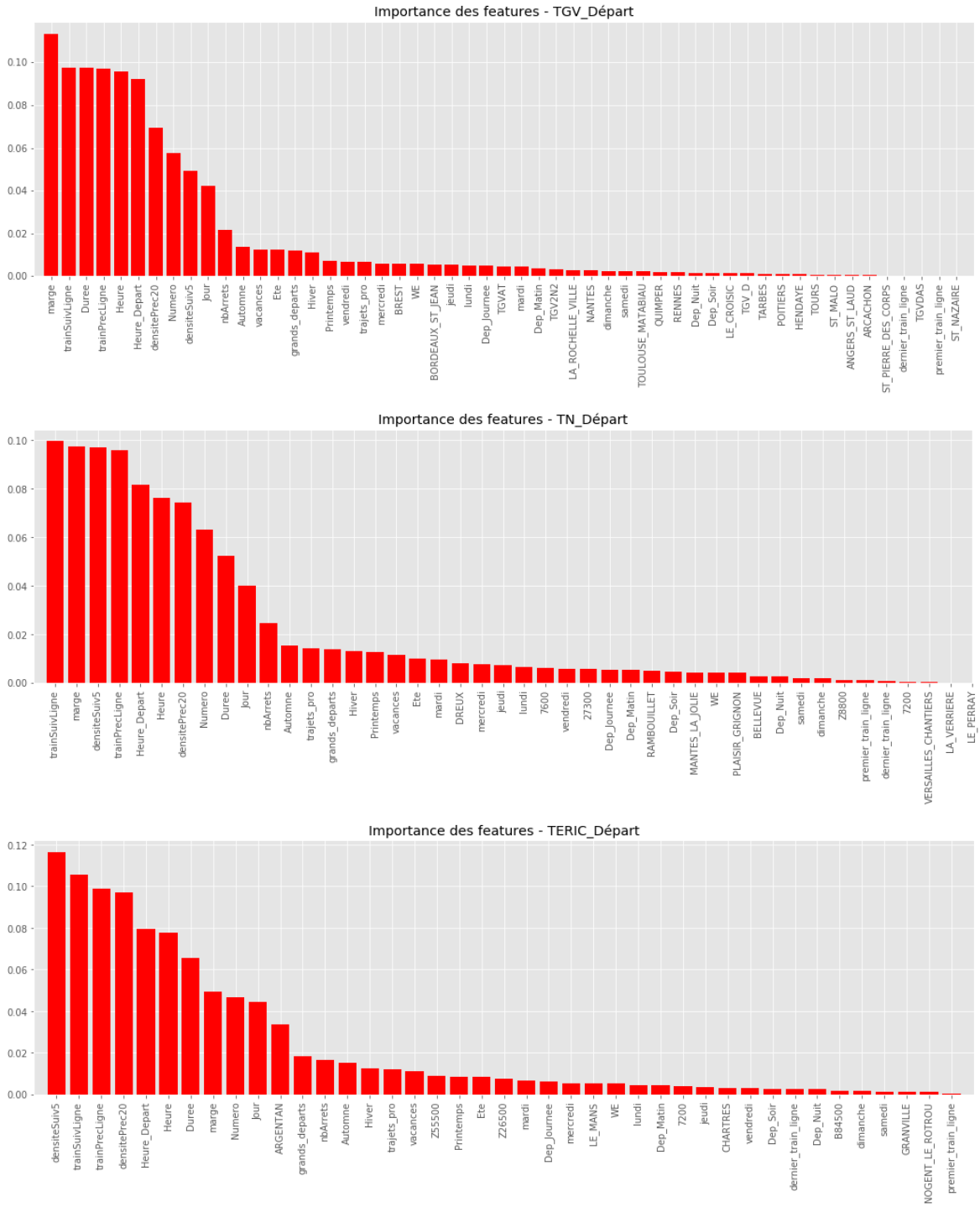

FIGURE 4.14 – Importance des features - départ

4.5 Résultats

L'analyse des expérimentations s'articule en deux étapes. La première est une validation externe, c'est-à-dire qu'on a isolé au préalable des données de la base d'apprentissage, et on utilise ces données pour identifier quel modèle a les meilleures performances. La deuxième étape consiste à évaluer les capacités prédictives du modèle final sur les données de test qui sont postérieures à celles d'apprentissage et de validation. Ces données ne sont pas disponibles au moment où le modèle est construit, dans la mesure où l'objectif de la méthode est de donner des estimations de risque pour ces données du futur et de les utiliser pour adapter les planifications en gare. On souhaite cependant évaluer l'utilité et la stabilité de cette méthodologie en regardant si les motifs appris sur une année se répètent en général assez le mois suivant pour utiliser ces prédictions.

4.5.1 Validation et sélection des modèles finaux

L'utilisation de données qui n'ont pas servi à construire le modèle est nécessaire pour évaluer les capacité à généraliser à des données inconnues. On utilise pour cela le score RPS (*Ranked Probability Score*, cf 3.2.1.2). Le RPS mesure l'erreur quadratique entre les distributions estimées et la distribution empirique, ce qui permet de classer des modèles en compétition selon leur efficacité globale (la calibration et discrimination des performances sont implicitement évaluées dans le score). Ce score, compris entre 0 et 1, est rappelé dans la formule 4.3, avec T la valeur de troncature, n le nombre d'observations évaluées, $o_{i,t}$ valant 1 si $y_i < t$ et $p_{i,t}$ la probabilité de cet évènement estimée par le modèle. Il est préféré ici par rapport à un score logarithmique car les modèles ont tous deux été optimisés par rapport à une mesure logarithmique (entropie, déviance), et que ces mesures ne prennent pas en compte la distribution dans son ensemble, alors que le RPS n'est pas un score local, et respecte l'ordre entre les valeurs.

$$RPS = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (o_{i,t} - p_{i,t})^2 \tag{4.3}$$

On présente les scores obtenus pour les deux modèles finaux dans le tableau 4.4 pour les différentes bases de données et périodes d'apprentissage. On ajoute pour élément de comparaison un modèle empirique noté M_0 qui utilise, comme estimation de distribution de probabilité, la fréquence de chaque valeur au sein de la base d'apprentissage, sans prise en compte du contexte. Ce modèle de référence simple permet de quantifier ce qui est appris par les autres modèles.

On constate que à information équivalente, les forêts aléatoires obtiennent des scores largement meilleurs que les modèles linéaires généralisés, quelles que soient la période et les données étudiées. Les GLM apprennent moins mais il y a tout de même une amélioration significative par rapport au modèle de référence dans le cas des arrivées, mais moindre dans le cas des départs. Cette faible amélioration se justifie à la fois par la difficulté du problème, puisque que les trains au départ sont très ponctuels. Comme la gare Montparnasse est terminale, les trains ont peu d'historique expliquant le retard, ce qu'on peut constater avec le faible nombre de variables ajoutées aux modèles lors de l'étape de sélection pas à pas.

4.5.2 Qualité des prédictions de test

Dans cette étape, on évalue la qualité des modèles sélectionnés, soit en l'occurrence les forêts aléatoires dans tous les cas étudiés. Comme présenté au chapitre précédent, deux aspects sont recherchés : la calibration et la discrimination.

Arrivée										
		Juillet	Aout	Sept.	Octobre	Nov.	Dec.	Janvier	Fevrier	Mars
TGV	RF	0.0764	0.0797	0.0807	0.0801	0.0819	0.0813	0.0818	0.0798	0.077
	GLM	0.0848	0.0877	0.0882	0.0878	0.0892	0.0887	0.0887	0.0872	0.0829
	M_0	0.118	0.121	0.1201	0.1214	0.1203	0.121	0.1182	0.1146	0.1123
TN	RF	0.0523	0.0505	0.0488	0.0505	0.0518	0.0509	0.0502	0.0534	0.0512
	GLM	0.0578	0.0556	0.0543	0.0557	0.0577	0.0562	0.0556	0.0592	0.0568
	M_0	0.0668	0.0644	0.0632	0.0642	0.0661	0.0661	0.0652	0.0682	0.0658
TER	RF	0.0692	0.0668	0.073	0.0683	0.0718	0.0694	0.0688	0.0678	0.0654
	GLM	0.0742	0.0734	0.0796	0.0751	0.0761	0.0733	0.0728	0.0715	0.0698
	M_0	0.0834	0.0803	0.0872	0.0816	0.0849	0.0828	0.0797	0.0817	0.0784

Départ										
		Juillet	Aout	Sept.	Octobre	Nov.	Dec.	Janvier	Fevrier	Mars
TGV	RF	0.0373	0.0373	0.0363	0.0349	0.0329	0.0317	0.0303	0.031	0.029
	GLM	0.0432	0.0438	0.0406	0.0398	0.0388	0.0373	0.0355	0.0316	0.0325
	M_0	0.0438	0.0444	0.0411	0.0402	0.0392	0.0375	0.0358	0.0319	0.0327
TN	RF	0.028	0.0276	0.0276	0.0267	0.0265	0.0255	0.0257	0.0254	0.0246
	GLM	0.0332	0.0351	0.0301	0.033	0.0299	0.0285	0.0285	0.0305	0.0278
	M_0	0.0347	0.037	0.0312	0.0345	0.0314	0.0296	0.0297	0.0316	0.0289
TER	RF	0.0494	0.0559	0.0501	0.0529	0.0509	0.0479	0.046	0.0389	0.0428
	GLM	0.0526	0.0592	0.0535	0.0562	0.0542	0.0506	0.0492	0.0413	0.0455
	M_0	0.0544	0.0617	0.0553	0.059	0.0552	0.0521	0.0511	0.0425	0.0476

Tableau 4.4 – Validation externe sur RPS

	TGV_A	TN_A	TER_A	TGV_D	TN_D	TER_D
GLM	26%	14%	10%	1%	4%	3%
RF	33%	22%	16%	11 %	12 %	9%

Tableau 4.5 – Pourcentage d'amélioration moyen par rapport au modèle de référence

La discrimination signifie qu'on est en mesure de classer les trains selon leur risque de retard. Cet attribut est important car on souhaite arbitrer différents scénarios d'affectations de voie et itinéraire en se basant sur ces risques, par exemple privilégier les cas où les voies occupées par des trains à fort risque de retard sont laissées disponibles un peu plus de temps pour limiter le risque de propagation. Si le modèle n'est pas discriminant, il n'y aura pas d'amélioration de la robustesse des planifications. La calibration, qui correspond à la fiabilité de la valeur prédite est également un prérequis pour la prise de décision, surtout qu'ici les trains sont séparés par type et sens de circulation. Il faut que les probabilités soient calibrées pour que les propriétés de discrimination soient conservées au moment où les différentes probabilités seront utilisées simultanément. Par exemple si le risque est largement surestimé pour les TN et sous estimé pour les autres, même si les modèles sont individuellement discriminants, les probabilités n'auront plus de sens utilisées ensemble car le biais de calibration empêchera de classer correctement selon le risque. La calibration est également nécessaire pour doser les marges en fonction des risques. Par exemple, ce n'est pas utile de protéger très fortement le train avec le plus haut risque de retard si des trains avec un risque à peine plus faible sont sacrifiés. La capacité résiduelle en gare doit être répartie le plus équitablement possible selon les risques, ce qui nécessite une bonne calibration.

4.5.2.1 Discrimination

La discrimination est mesurée à l'aide de la concordance, calculée en utilisant la valeur moyenne comme estimateur du retard. La concordance correspond à la probabilité qu'en sélectionnant aléatoirement deux trains T_1 et T_2 tels que le retard de T_1 soit strictement supérieur au retard de T_2, alors le retard moyen estimé de T_1 soit supérieur à celui de T_2. Les scores sont donc compris entre 0 et 1 où 1 équivaut à un modèle parfait. Un modèle complètement non discriminant, comme par exemple le modèle de référence M_0 qui estime la même probabilité de retard à tous les trains d'une catégorie, a une concordance égale à 0.5 par défaut car on a une chance sur deux de classer correctement les trains. Les scores de concordance sont donnés dans le tableau 4.6.

	Juillet	Août	Septembre	Octobre	Novembre	Décembre	Janvier	Février	Mars
TGV_A	0.6	0.58	0.65	0.68	0.68	0.65	0.66	0.68	0.67
TN_A	0.79	0.72	0.79	0.81	0.79	0.77	0.81	0.83	0.78
TER,IC_A	0.73	0.62	0.72	0.71	0.74	0.69	0.71	0.75	0.73
TGV_D	0.57	0.68	0.73	0.73	0.75	0.69	0.77	0.8	0.82
TN_D	0.82	0.81	0.83	0.84	0.87	0.86	0.87	0.89	0.89
TER,IC_D	0.77	0.68	0.75	0.75	0.72	0.74	0.75	0.8	0.79

Tableau 4.6 – Étude de la concordance

A titre informatif, les GLM qui ont été exclus lors de la validation externe ont une concordance qui est systématiquement inférieure de quelques centièmes à dixièmes aux forêts aléatoires.

Les capacités prédictives sont très variables d'un type et sens de circulation à l'autre. Les modèles des transiliens ont la meilleure discrimination, en particulier au départ avec des scores allant jusqu'à 0.89. Les variations d'un mois à l'autre de la concordance sont en partie liées à des motifs absents certains mois et présents à d'autres qui permettent de mieux discriminer les trains, par exemple des motifs de vacances ou de saisons. La seule exception concerne le mois d'août 2018 qui a été anormalement perturbé sur les trains transiliens et régionaux, obtenant donc des concordances basses qui n'ont pas pu être anticipées.

4.5.2.2 Calibration

Graphes de calibration : ces graphes agrègent les observations en groupes de trains ayant un risque estimé similaire, afin de comparer ce risque au taux moyen d'évènement observés. Si les deux valeurs sont proches, on considère que le modèle est calibré, c'est-à-dire que les estimations sont fiables. Ces graphiques sont utilisés pour évaluer la calibration, cependant pour plus de visibilité, on ne présentera ici des graphes que pour deux périodes et deux seuils (1 minute et 5 minutes pour les arrivées, 1 minute et 3 minutes pour les départs). Les autres graphes peuvent être trouvés en annexe D.

Les périodes représentées sont les mois de juillet (figures 4.15 et 4.16) et de février (figures 4.17 et 4.18). Le choix est arbitrairement fait sur les deux mois ayant respectivement les moins bonnes performances et les meilleures en terme de concordance.

Chaque graphe représente les probabilités d'avoir un retard supérieur ou égal au seuil t agrégées en g groupes. Le nombre de groupes a été choisi de sorte à avoir environ 50 trains par groupe. Avoir beaucoup de groupes permet de mieux se rendre compte de la répartition des probabilités prédites et d'augmenter l'homogénéité au sein des groupes, mais augmente également la dispersion du graphe puisque les probabilités estimées et observées sont moyennées sur de petits échantillons.

De manière générale, les graphes de calibration doivent surtout être étudiés pour des seuils faibles (1 à 5 minutes maximum), puisque pour des valeurs supérieures, les probabilités estimées se concentrent

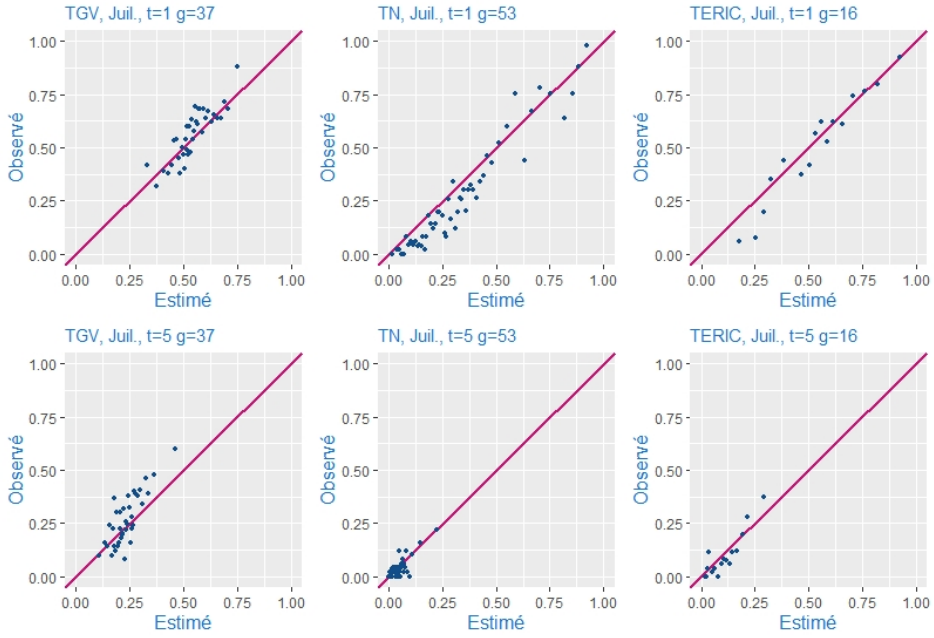

FIGURE 4.15 – Graphes de calibration arrivée - Juillet

sur un intervalle réduit, les erreurs sont faibles et les déviations sont peu interprétables à l'oeil nu (voir graphes pour $t = 3$ ou $t = 5$).

L'analyse des graphes sur les différentes périodes montre qu'à l'exception du mois d'août où les perturbations étaient anormales, les différents modèles sont assez bien calibrés.

— Les TGV à l'arrivée sont ceux pour lesquels les modèles sont le plus souvent biaisés. Par exemple dans le cas du mois février, les points sont très majoritairement sous la diagonale ce qui est caractéristique d'une surestimation des retards. Ce biais a également été observé pour les mois d'août, janvier, mars. Les graphiques des mois de septembre et octobre montrent une sous estimation des probabilités basses (inférieures à 0.5 pour $t = 1$) mais une meilleure calibration pour les valeurs supérieures. Les modèles des autres mois sont bien calibrés.

— A l'exception des tests effectués sur les prédictions du mois d'août, où les observations sont anormalement élevées, les modèles pour les Transiliens à l'arrivée donnent des prédictions bien calibrées, avec cependant une légère forme de sigmoïde caractéristique des estimations de forêts aléatoires (points sous la diagonale pour les valeurs plus faibles, et au dessus après).

— De même, les prédictions faites sur les trains régionaux et intercités à l'arrivée sont bien calibrées, avec également une légère forme de sigmoïde sur certains mois.

— Tous les modèles pour les trains au départ donnent des probabilités calibrées indépendemment du type de circulation, avec des écarts à la diagonale toujours réduits.

Approches statistiques : deux approches reposant sur des tests statistiques pour l'évaluation de la calibration ont été présentées dans le chapitre précédent : le test d'Hosmer-Lemeshow et les résidus de quantiles randomisés. De nombreux essais utilisant ces méthodes sur les modèles présentés ci-dessus ont été entrepris, cependant ils ont mis en évidence plusieurs défauts :

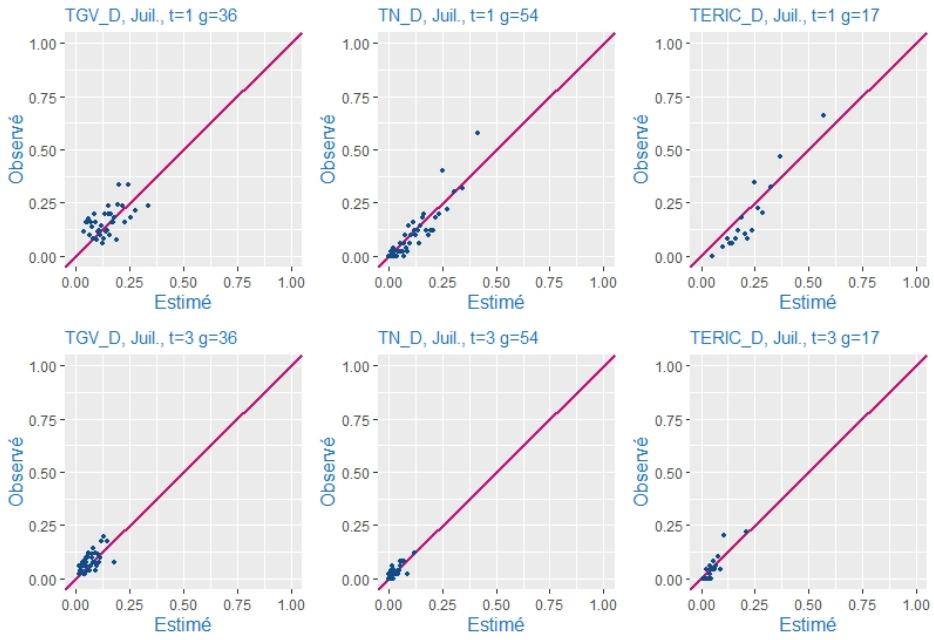

FIGURE 4.16 – Graphes de calibration départ - Juillet

— ces tests sont très sensibles à la taille de l'échantillon considéré. Dans cette thèse, les échantillons correspondent aux données d'un mois, dont la taille peut être variable selon le type et sens de circulation. Une standardisation est connue pour le test d'Hosmer-Lemeshow en jouant sur le nombre de groupes à utiliser pour calculer la statistique. Pour les RQR, il est préférable d'échantillonner et de compter la proportion de tests où la p valeur est inférieure à 0.05. Cependant on a observé que l'évaluation dépendait fortement des conditions d'expérimentations, comme par exemple la durée de test et le type de train puisqu'on ne dispose pas d'autant d'observations selon l'activité.

— ils sont peu interprétables et peuvent rejeter des modèles pour de petites déviations jugées acceptables pour la prise de décision. En particulier dans le cas du RQR, il est possible de procéder à une analyse graphique du modèle en construisant un histogramme des résidus, cependant il est bien moins interprétable que les graphes de calibration.

On préconise donc de se contenter de tracer des graphes de calibration pour vérifier qu'il n'y a pas d'écarts de calibration qui compromettent la qualité des prédictions. En particulier, dans ce cas d'étude où la majorité des retards sont nuls ou très faibles, on peut se contenter d'évaluer la calibration sur les premières minutes car c'est là que se concentre la masse de la distribution. Les écarts de calibration y sont donc les plus sévères.

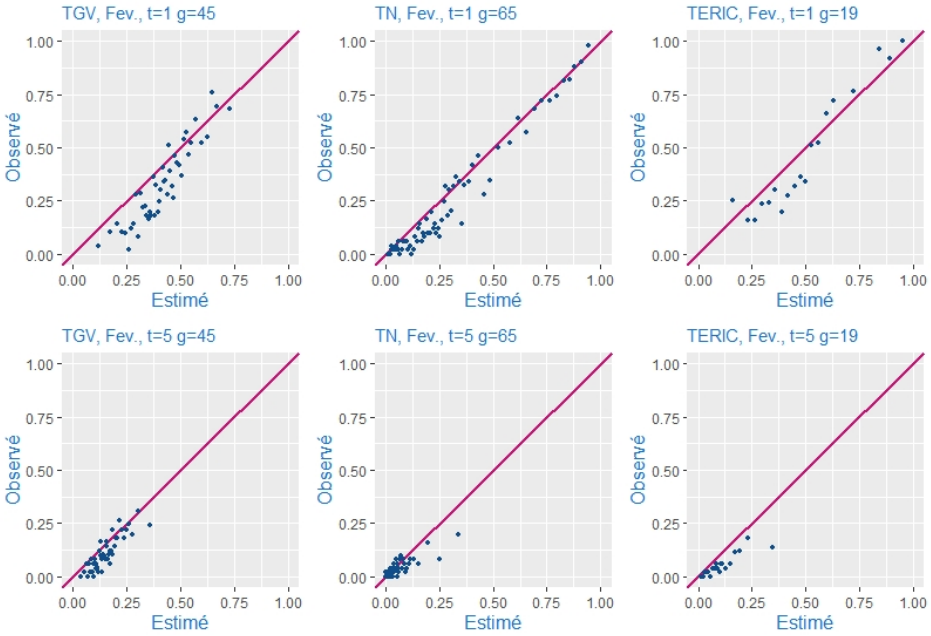

FIGURE 4.17 – Graphes de calibration arrivée - Février

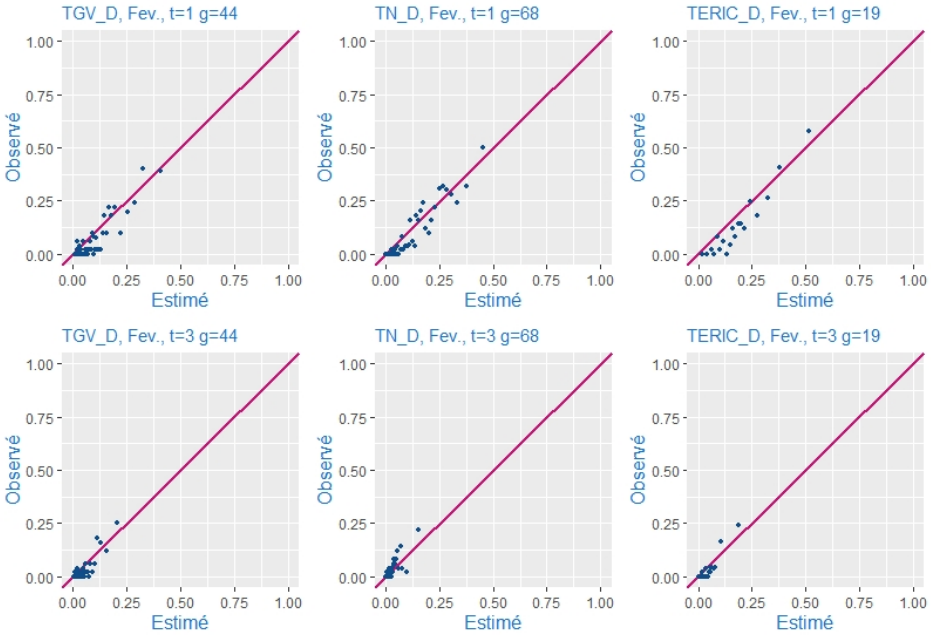

FIGURE 4.18 – Graphes de calibration départ - Février

4.6 Discussions

4.6.1 Choix de l'approche

L'utilisation de modèles linéaires généralisés a été motivée en premier lieu par le constat que les données de retards sont fortement asymétriques et bien représentées par une loi binomiale négative. Les forêts aléatoires ont été introduites après comme méthode alternative, bien que le support des données et l'objectif d'estimation de probabilités aient nécessité d'adapter l'algorithme.

Les expérimentations précédentes ne soutiennent cependant pas cette intuition initiale quant à l'adéquation des GLM. Plusieurs points négatifs peuvent être rappelés :

— Difficulté de construction : pour une liste de variables explicatives donnée, les forêts aléatoires ne nécessitent que de fixer les valeurs de quelques hyperparamètres. Pour les GLM la méthodologie de construction est beaucoup plus complexe. Les variables sont fortement corrélées et parfois non informatives. Une sélection pas-à-pas a été menée ici mais elle demande également un paramétrage (choix du critère d'optimisation, ici le critère BIC, coordination des sélections pour les différents paramètres,...). Selon la base de données étudiée, la sélection peut également prendre un temps très conséquent, allant jusqu'à plusieurs heures de calcul. Plusieurs modèles alternatifs sont possibles, à savoir changer de loi ou utiliser des modèles en deux parties ou à inflation de zéros.

— Manque de robustesse : les différentes expériences menées sur des GLM montrent que le rapport aux variables est parfois instable. Les variables sélectionnées peuvent varier beaucoup selon les données d'apprentissage. C'est par exemple ce qu'on observe en comparant les ensembles de variables obtenus pour les différentes périodes.

— Performances : on a pu constater que les forêts aléatoires avaient des performances générales supérieures aux modèles linéaires généralisés quel que soit le cas considéré (type de train, sens de circulation, période d'apprentissage). En particulier, dans le cas des trains au départ, les GLM améliorent très peu le modèle de référence contrairement aux forêts.

Le choix de l'approche est cependant fortement conditionné par le format des données. Un GLM s'adapte à tous types de données (continues, discrètes, tronquées, etc.) quand les forêts aléatoires utilisées ici ne peuvent prédire des probabilités que dans un cadre binaire ou multiclasse.

4.6.2 Limites liées aux données

Les travaux présentés se basent principalement sur les relevés de passages des trains, qui fournissent à la fois la cible (le retard) et une grande partie des variables explicatives. Malgré quelques possibilités d'amélioration des modèles, les capacités prédictives sont toujours bornées par la qualité des données sur lesquelles ils s'appuient. On revient ici sur plusieurs défauts des données utilisées. Les limites identifiées sont de trois ordres : soit inhérentes à ce qui est observé (le réseau ferroviaire, les retards), soit dépendantes du système de collecte des données, soit liées aux décisions de modélisation.

4.6.2.1 Limites du problème

Évolution temporelle : une des principales limites liées aux données est que le système ferroviaire évolue au cours du temps. Certaines mesures sont très structurantes et risquent de fausser les modèles. Cela a par exemple été le cas dans cette étude avec l'ouverture de la LGV Atlantique au 1^{ier} juillet 2017 qui a rendu caduques les observations antérieures à cette date (les temps de parcours ont fortement diminué, certaines portions du trajet sont passées en voies dédiées ce qui limite le risque de conflits, les missions ont été modifiées). La méthodologie présentée ici n'est pas robuste à de grandes modifications

du réseau comme celle-ci, et demande un temps d'adaptation pour recueillir de nouvelles données. D'autres modifications moins importantes ont tout de même lieu de manière locale sur le réseau et peuvent affecter dans une moindre mesure les profils de retards des trains. Il peut s'agir de changement de politiques locales, fin de travaux, etc.

Exceptions : on se concentre ici sur les motifs de petits retards récurrents en conditions normales dont on cherche à estimer la distribution de probabilité. La notion de *conditions normales* est cependant difficile à définir. Les évènements les plus importants comme les grèves, gros incidents ou travaux importants sont faciles à identifier, cependant il existe beaucoup d'anomalies dans la base qui ne peuvent pas être détectées et ajoutent du bruit.

Impact des opérations en gare : l'objectif de ce travail est d'estimer les probabilités de perturbation pour améliorer la robustesse des opérations en gare. Les retards créés en gare à cause d'un manque de robustesse des planifications ne doivent donc pas être inclus puisqu'on cherche à anticiper leur création. C'est pour cette raison que les données sont collectées avant l'arrivée dans le périmètre de gare pour les trains à l'arrivée et au niveau des quais pour les trains au départ. Cependant pour les trains au départ il n'est pas possible d'identifier dans les données quels sont les retards liés à un train maintenu à quai pour aider à résoudre un conflit, ou s'il ne peut pas s'engager car une ressource est indisponible, et quels sont les retards liés à l'instabilité de la production (temps d'occupation excessif, passagers bloquant les portes,etc.). Les probabilités surestimeront donc légèrement le risque réel mais forceront en contrepartie à protéger plus ces trains lors de l'adaptation robuste des plannings.

4.6.2.2 Système de collecte

Qualité de l'enregistrement : les retards utilisés ici sont collectés via l'outil de suivi des circulations *Brehat* [4]. Plusieurs bases de données sont créées à partir de ces enregistrements bruts, celles utilisées ici donnent des retards tronqués à la minute. Utiliser une base de données avec une granularité plus fine n'est pas toujours possible ni pertinent. Les balises ne sont pas toutes positionnées au même niveau sur les voies, ce qui peut déformer mécaniquement les retards observés. Étant donné que les modèles construits ici se basent sur une représentation macroscopique du réseau, on n'est pas en mesure d'identifier et corriger ces écarts liés au positionnement des balises. Ces écarts sont compensés en grande partie par l'utilisation d'un pas de mesure plus élevé, comme la minute. Ce pas est cependant trop important, surtout pour les trains au départ qui ont très majoritairement un retard d'au plus quelques minutes, cela contribue à déséquilibrer les données en créant de fortes masses en zéro.

Données manquantes : Brehat permet d'assurer un suivi de toutes les circulations mais il arrive qu'il y ait des dysfonctionnements des balises ou de l'archivage, et les trains concernés ne sont donc pas enregistrés. Cela arrive le plus souvent dans des petites gares où le problème peut durer un certain temps, mais également parfois pour des gares plus importantes. A notre connaissance les données relevées à Paris ne sont pas concernées mais des exemples ont été détectés sur des gares du parcours où le passage ou arrêt n'est pas enregistré, ce qui peut fausser quelques variables explicatives (nombre d'arrêts, origine, desserte, etc.).

4.6.2.3 Limites liées aux modèles

Dépendance des retards : les modèles utilisés ici considèrent les retards comme indépendants les uns des autres, ce qui en pratique n'est pas vrai. Cette hypothèse est valable sur des réseaux complètement différents (par exemple un TGV arrivant à Montparnasse n'a a priori pas d'effet sur le retard au départ d'un transilien quittant la gare) ou sur des dates différentes (sauf incidents majeurs,

les retards d'un train n'a pas d'impact ou n'est pas lié aux retards observés à d'autres moments). La troncature permet de limiter la dépendance entre les retards d'une même journée mais on ne peut pas exclure certaines exceptions, notamment entre des trains consécutifs dans une période courte sur une même infrastructure.

Choix de troncature : les troncatures appliquées ici ont été décidées à la fois pour adhérer aux contraintes industrielles (les retards importants ne sont pas pertinents pour la robustesse) et par interêt statistique (les valeurs hautes sont liées à des causes rares et variées, et non à une instabilité récurrente du réseau). Les valeurs sont donc de fait dépendantes du type de circulation car les ordres de grandeurs des retards varient beaucoup en fonction de cette partition. Par exemple, il est très rare d'observer 15 minutes de retard pour un transilien ou un train au départ mais cette valeur est plus fréquente sur des lignes TGV. Une valeur unique de troncature simplifierait cependant fortement la modélisation et l'évaluation, et permettrait d'harmoniser les probabilités prédites.

Données réelles : le système de collecte ne relève des informations que pour les circulations réalisées mais ce serait plus rigoureux d'utiliser en complément les planifications théoriques pour apporter en précision sur les mesures de certaines variables. En particulier, les variables de densité en gare et de densité en ligne sont calculées par rapport aux horaires théoriques des trains qui ont effectivement circulé, mais ces valeurs peuvent être légèrement modifiées par la présence de trains supprimés, comme cela peut être le cas pour des transiliens, ou encore des trains ajoutés au dernier moment, comme des mouvements techniques en gare.

4.6.3 Résultats selon le type de train

Cette partie recense des conclusions selon le type de train. Les principaux résultats concernent les TGV et TN à l'arrivée car la méthodologie a surtout été développée sur ces deux bases avant d'être appliquée aux autres données. On se concentre sur les trains à l'arrivée, car pour les trains au départ les motifs sont principalement liés à la densité en gare de Montparnasse et à l'heure de circulation.

TGV : malgré des profils de circulations très variés, que ce soit par l'origine, la durée du trajet ou la temporalité, les estimations de probabilités des TGV sont de moins bonne qualité que pour les autres types de circulation, que ce soit au départ ou à l'arrivée. En effet, on constate que les probabilités estimées par le modèle final sont moins discriminantes et que le graphe de calibration associé est plus dispersé dans le cas des départs, et plus biaisé pour les arrivées, avec dans les deux cas une moins bonne couverture de l'intervalle [0,1]. Le modèle ne parvient pas à identifier avec certitude les trains les plus ponctuels et les trains les plus retardés, ce qui se caractériserait par des probabilités proches de 0 ou de 1 dès un seuil de 1 minute, comme cela peut être le cas pour les Transiliens.

Les faibles capacité prédictives sur TGV viennent a priori d'une plus grande instabilité des opérations et d'un bruit plus fort. En effet, les TGV parcourent de très grandes distances sur un réseau très étendu et plus complexe, où différents types de trains interagissent, or l'hétérogénéité du trafic est connue pour contribuer à l'instabilité du réseau et à la propagation des retards. Certains arrêts sont susceptibles d'être allongés, par exemple pour attendre une correspondance ou pour une coupe-accroche qui prend du retard. Cependant, les TGV disposent de plusieurs minutes de marges pour parcourir les longues distances entre les arrêts, et parfois des arrêts longs, ce qui donne des opportunités de rattraper le retard. Ainsi les processus de propagation, de rattrapage ou d'aggravation des retards pour les TGV sont complexes et dépendent très fortement de ce qu'il se passe sur le réseau à large échelle, ce qui limite la récurrence des motifs.

Concernant les variables importantes, la saisonnalité a une plus grande influence pour les TGV que les autres trains. Cela mériterait une analyse plus approfondie pour identifier les raisons de cette corrélation à la saison. Un nouveau découpage des données pourrait alors être pertinent : si les observations provenant d'autres saisons biaisent les résultats pour les TGV, on pourrait se concentrer sur les périodes similaires au lieu de conserver une année entière. Les quelques tests effectués en utilisant d'autres périodes d'apprentissage ne vont cependant pas dans ce sens.

TN : c'est sur cette base de données que le modèle est le plus prometteur. On constate en effet grâce aux graphes de calibration qu'on est capable de discerner de manière automatique et fiable des profils de retards très différents, avec par exemple des trains à l'arrivée pour lesquels on prédit plus de 90% de risque d'être en retard et pour d'autres moins de 5%, alors qu'en moyenne le taux de trains en retard est de 33% sur la base tronquée. Les motifs de retards sont stables et récurrents, leurs conséquences peuvent donc plus facilement être anticipées. De même pour les départs, les taux de concordance sont très élevés (entre 0.81 et 0.89) et les graphes montrent des probabilités bien calibrées et raisonnablement dispersées.

Les Transiliens circulent en zone dense et sont donc soumis à des effets très spécifiques liés à la saturation de l'infrastructure et l'affluence en gare. En effet, l'analyse des variables sélectionnées pour les arrivées et des différentes visualisation de données suggèrent que les distributions de retards s'expliquent bien par la densité, l'heure, les temps d'arrêt planifiés et l'origine du train. Les premières sont à la fois des indicateurs d'heure de pointe et de phénomènes de congestion qui empêchent de rattraper les retards et les propagent. Les temps d'arrêt planifiés insuffisants peuvent également être des facteurs de risque alors qu'au contraire un temps plus important permet de gagner en robustesse et de rattraper les retards.

Dans le cas des départs, les variables principales sont liées à la densité et à l'heure. On constate également que la variable la plus importante pour les forêts est la *marge*, ce qui est moins intuitif pour les trains au départ. On peut également voir que cette variable est utilisée dans l'arbre de probabilité construit en analyse préliminaire 4.7 dans une branche après avoir déjà séparé selon la densité en gare et en ne gardant que les TN, où un temps de trajet plus long que la normale est signe de meilleure ponctualité.

TER et IC : ces trains sont peu nombreux et assez hétérogènes, ce qui complique la modélisation. Ils peuvent éventuellement être intégrés à la base de données des Transilien car une partie des TER circulent sur la même infrastructure et avec le même matériel, cependant leur trafic est bien moins fréquent et ils parcourent des distances plus longues, ce qui fait que ces trains ne sont pas utilisés de la même manière par les usagers (moins de trajets domicile travail, pas de caractéristiques de trafic de masse). Ils ont des caractéristiques communes aux TGV puisque les TER et Intercité sortent du réseau d'Ile-de-France pour circuler à l'échelle de plusieurs régions sur un réseau ouvert et potentiellement plus instable, mais ne peuvent pas être intégrés à la base car les variables explicatives seraient très hétérogènes en raison des différences de vitesse et de matériel.

Comparaison : la différence de qualité entre les prédictions pour les TGV et celles pour les TN peut s'expliquer par les différences d'exploitation de ces types de trains. Les Transiliens sont des trains de zone dense, les circulations sont nombreuses et se répètent beaucoup, d'une heure à l'autre et d'un jour à l'autre, avec une liste limitée de dessertes assurées (une douzaine). Pour les TGV, il existe aussi un grand nombre de circulations se répétant, mais la variabilité des trajets et des horaires est plus importantes, avec beaucoup de dessertes rarement desservies ou densifiées temporairement pour des évènements. Par ailleurs, la zone dense fait que les retards se contrôlent plus mécaniquement pour les trains de banlieue que pour les TGV avec moins d'opportunités de rattraper les retards, et des

phénomènes d'aggravation plus prévisibles. Par exemple si un train arrive en retard en zone dense, il restera plus longtemps à quai pour absorber la quantité de passagers qui est arrivée durant son retard, et entrera ainsi dans un effet de propagation. Par ailleurs on observe très peu de retards importants sur les TN car ils peuvent être supprimés pour permettre un retour à la normale plus rapide, ce qui n'est pas le cas des trains à réservation (TGV, IC) ni des TER qui sont moins fréquents.

Les causes des retards faibles sont également moins connues pour les TGV et IC que pour les autres trains puisqu'ils sont moins soumis à l'influence des flux voyageurs qui sont prévisibles et dépendent du contexte temporel et de la desserte.

4.6.4 Pistes d'améliorations

Il existe deux principaux leviers d'amélioration des performances : changer le modèle ou enrichir la base de données.

Amélioration des modèles : les pistes d'améliorations ont été citées plus-haut, mais on peut les compléter en intégrant les recommandations suivantes :

- pour les modèles linéaires généralisés :
  - essayer de nouveaux encodages des données. En particulier, on constate que les deux modèles ne privilégient pas les mêmes variables, et donc captent une information différente. Par exemple, pour les TN à l'arrivée, certaines variables comme *SEVRES_RIVE_GAUCHE* qui est sélectionnée systématiquement en deuxième pour modéliser σ fait partie des variables les moins importantes pour les forêts aléatoires. De manière générale, les forêts semble exploiter fortement les variables de densité, d'heures et de marge mais beaucoup moins celles de mission (temps d'arrêt, numéro, origine). Au contraire, les GLM sélectionnent peu les variables de densité mais exploitent beaucoup plus les variables d'origines et de plage horaire. Il faudrait explorer plus en détail comment adapter les variables pour permettre aux modèles d'en extraire plus d'information. Il faut par exemple étudier les points de séparation les plus utilisés dans des arbres de probabilité pour les variables de densité ou voir si ces variables doivent être étudiée conditionnellement à une autre variable.
  - modifier la structure du modèle, comme avec une distribution à trois paramètres, ou avec des modèles en deux parties. Par ailleurs, la première partie est en général une régression logistique afin de prédire la probabilité d'observer une valeur nulle, mais on peut également utiliser une forêt aléatoire. Cette approche permet de s'adapter à des distributions continues ou discrètes non tronquées tout en tirant avantage de la plus grande précision des forêts aléatoires. Des tests antérieurs sur des modèles en deux parties (dont une partie des résultats peuvent être trouvés dans [63]) mettaient en évidence que le modèle binaire avait une très forte influence sur les performances globales en raison de la forte masse en zéro. Dans le cas où la seconde partie utilise une distribution continue, les capacités prédictives sont faibles pour la seconde partie (calibration et discrimination), cependant l'utilisation d'un bon modèle binaire en première partie permet de compenser fortement.
- pour les forêts aléatoires :
  - la piste principale est d'essayer d'exploiter l'ordre entre les valeurs de retards. Dans les modèles présentés ici, les retards sont traités comme des classes indépendantes (classification ordinale). D'autres alternatives ont été proposées dans le chapitre précédent, notamment la décomposition dichotomique où on calcule la probabilité d'avoir un retard inférieur ou supérieur à un seuil ou l'utilisation d'arbres de régression.

— comme pour les GLM l'encodage peut être revu. Les forêts aléatoires semblent ne pas capter autant l'information comprise dans l'origine des trains et leur mission. Plus de tests devraient également être effectués en intégrant plus de données brutes comme variables explicatives. Ici, la base est construite pour pouvoir être appliquée à la fois pour des GLM et des forêts aléatoires. L'information est donc agrégée en un ensemble de variables binaires et numériques exploitables par un modèle linéaire. Cependant pour des forêts aléatoires, on peut essayer de structurer moins l'information, quitte à avoir plus de variables inutiles. Une option est par exemple de créer une variable pour chaque gare du réseau et d'enregistrer la durée de l'arrêt dans la gare ou encore la densité du trafic, et une valeur négative par défaut quand le train ne fait que y passer sans arrêt ou quand il n'y passe pas du tout. L'information serait plus riche que dans l'état actuel des choses et l'algorithme de forêts aléatoires peut efficacement séparer les cas de figure et reconstituer les missions en fonction des valeurs de la variable. Cette option permettrait également de mettre en évidence les points critiques du réseau.

Au vu des expériences menées dans ce chapitre, on recommande l'utilisation de forêts aléatoires qui donnent systématiquement de meilleures performances que les modèles linéaires généralisés. Une approche mixte, c'est-à-dire une forêt aléatoire pour prédire la probabilité d'un retard négatif ou nul et un GLM pour la distribution strictement positive, devrait être testée pour des données ne rentrant pas dans le cas d'étude présenté (retards non tronqués ou mesures continues).

La séparation en différents types de train et sens de mouvement est nécessaire pour les approches linéaires car les données sont hétérogènes et plusieurs variables ont une relation très différente au retard selon le cas de figure. Cependant les forêts aléatoires sont en mesure d'identifier ces différences, et la séparation des données n'est plus forcément nécessaire. Les visualisations d'arbres de décision calculés sur les données complètes montrent tout de même que les séparations selon le type d'activité sont faites très haut dans l'arbre, les estimations ne s'amélioreront éventuellement pas, mais cela permettrait de simplifier la méthodologie en travaillant sur une base unique ou sur une base d'arrivée et une de départ. Cette approche requiert cependant d'utiliser une même valeur de troncature pour tous les trains sans quoi on risque d'obtenir des probabilités non nulles au delà du seuil de troncature pour certains trains.

Enfin, il est aussi possible de changer de méthode d'apprentissage. Comme on l'a vu au chapitre précédent, plusieurs autres algorithmes peuvent également prédire des probabilités. Les algorithmes XGBoost et k plus proches voisins avec bagging ont été testés sur les données à l'arrivée, mais les forêts aléatoires donnaient des performances au moins équivalentes et souvent supérieures dans le cas du boosting, et très supérieures pour les plus proches voisins [128]. Des tests plus approfondis devraient être faits en appliquant des réseaux de neurones ou des machines à support de vecteurs. Il semble cependant que la limite principale aux performances soit liée à l'insuffisance des données plutôt qu'aux échecs de modélisation.

Données complémentaires On peut pallier partiellement les limites présentées ci-dessus en enrichissant le modèle avec de nouvelles données. Étant donné que les prédictions doivent se faire en avance pour pouvoir avoir le temps d'adapter les planifications (au moins un jour et jusqu'à quelques semaines en avance), peu de données complémentaires ayant un impact peuvent s'ajouter la base déjà existante. Les bases suivantes peuvent être considérées :

— Travaux sur les voies : les pertes de capacité en ligne dues à des limitations de vitesse par exemple sont des facteurs d'instabilité du réseau. Quand ces travaux sont prévus de longue date, des suppléments de temps sont ajoutés au temps de trajet pour compenser ces pertes de capacité.

— Maintenance : concernant les trains au départ, un bon nombre de retards sont dus à une restitution tardive du matériel après sa maintenance. Cette information n'est cependant pas accessible,

et est très susceptible d'évoluer car les changements d'équilibres (utiliser une rame au lieu d'une autre pour effectuer un trajet) sont très fréquents en opérationnel.

— Données météorologique : de telles variables ont été utilisées dans un premier temps et on a pu constater qu'elles contribuaient à améliorer le modèle. Cet impact a été montré par plusieurs études statistiques, et est connu en opérationnel (les temps extrêmes comme les grands froids ou fortes chaleurs imposent une vitesse réduite et occasionnent des dysfonctionnement électriques, en temps de pluie les passagers ne se répartissent pas de la même manière sur les quais,...). Ces données n'étaient cependant pas compatible avec la modélisation mois par mois car les prévisions météorologiques ne sont pas disponibles autant de temps en avance. Si les prédictions sont effectuées à la semaine, l'intégration de ces données pourrait s'envisager.

— Composition du matériel : dans l'état actuel des choses, on ne connaît pas le nombre de rames qui composent les trains observés. Cette information n'est pas disponible dans Brehat, mais elle est connue pour l'organisation des opérations en gare. La récupération de la variable n'était pas envisageable ici en raison du format des données (un fichier par jour) mais cette piste devrait être explorée plus.

— Historique sur un trajet : l'ordre dans les observations et la temporalité pourrait être plus exploitées. Une approche envisageable serait d'ajouter une variable donnant les performances du train sur ses derniers trajets, par exemple la régularité au même horaire sur la semaine précédente. Une telle variable a cependant plus de sens pour un trajet fréquent, comme sur du Transilien ou des trajets grande ligne régulier que pour une desserte rare.

4.6.5 Adaptation à d'autres cas d'usage

4.6.5.1 Cas général

La méthodologie présentée ici et résumée dans la figure 4.19 est flexible et peut s'adapter à d'autres problèmes pour lesquels il est préférable de prédire une probabilité plutôt qu'une valeur ponctuelle. Les prédictions de probabilités sont intéressantes pour les cas où les prédictions vont servir à prendre des décisions, elles apportent alors une estimation de la certitude autour de la valeur prédite et la répartition des valeurs. Ces modèles sont utiles dans des cas comme le nôtre où la distribution est très asymétrique et où des prédictions ponctuelles renvoient le plus souvent la mode ou la moyenne, qui ne sont pas représentatives de la distribution des valeurs.

On recommande l'utilisation de modèles linéaires généralisés pour les cas où :

— la variable cible est continue ou à support infini ($\mathbb{N}$ ou restrictions de $\mathbb{R}$).

— il existe une distribution classique de la famille exponentielle qui adhère aux données.

On recommande l'utilisation de forêts aléatoires pour les cas où :

— la variable cible est binaire ou multi-catégorielle

— il existe des motifs non linéaires entre la cible et les variables explicatives

— si les variables explicatives ne sont pas sélectionnées ni transformées en amont de l'étude

— en éventuelle première partie d'un modèle linéaire généralisé en deux parties.

Les deux approches gagnent tout de même à être testées quand le format des données le permet étant donné qu'elles présentent toutes deux des avantages propres, et il a été montré dans le cas de variables binaires qu'aucune n'était systématiquement meilleure.

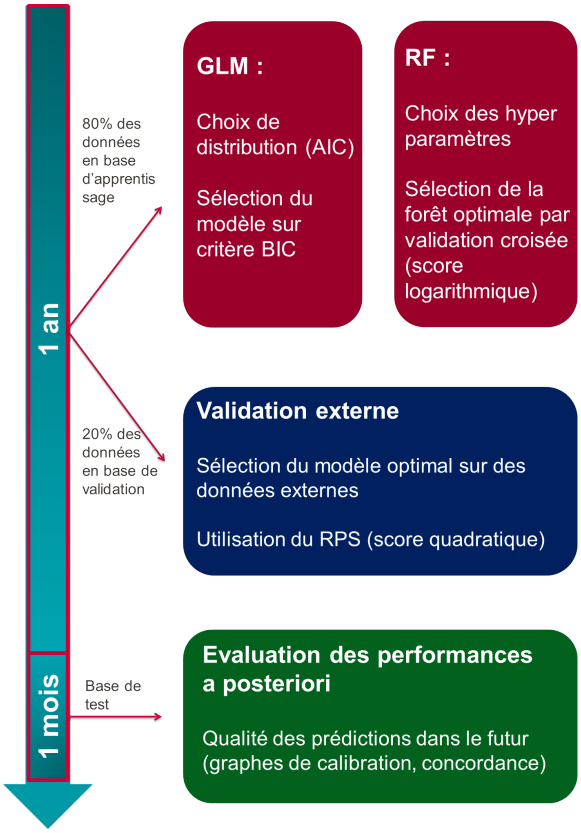

FIGURE 4.19 – Étapes de construction et évaluation du modèle final

4.6.5.2 Données ferroviaires

Cas d'étude de Montparnasse : cette méthodologie est adéquate pour les données de retards de trains étudiés en avance : on ne dispose pas d'assez d'informations pour prédire précisément le retard mais on arrive tout de même à distinguer différents profils de ponctualité. Le cas d'étude de la gare Montparnasse est riche car étant une gare terminale, les retards à l'arrivée y sont importants alors qu'au contraire les retards au départ y sont très faibles. Les bons résultats obtenus montrent que la méthode s'applique pour différents ordres de grandeur des retards. Les deux approches ont également été appliquées sur différents types de trains (TN, TER, TGV). Les résultats sont moins encourageants sur les TGV où on observe des erreurs de calibration et une concordance plus faible que sur les autres types de trains mais les modèles parviennent tout de même à prédire des probabilités de qualité acceptable.

Retards en temps réel : cette approche est moins pertinente dans le cas d'une prédiction en temps réel où le retard aux prochains arrêts dépend très fortement du retard courant, et les variables explicatives utilisées ici n'ont plus le même poids. Il est alors beaucoup plus aisé de prédire une valeur ponctuelle avec une forte certitude. L'algorithme de forêts aléatoires a été utilisé plusieurs fois comme un algorithme de régression (cf 2.4), et on pourrait envisager une approche alternative qui étudierait le contenu des feuilles terminales pour récupérer non pas une probabilité associée à chaque valeur, mais

les probabilités associées à quelques scenarios (le train maintient son retard, il le rattrape, l'empire un peu/fortement, etc.) qui viendraient compléter la valeur de retard prédite avec une mesure de la certitude associée.

Autres données de retards : les données traitées pour la gare Montparnasse couvrent des cas d'étude variés : les TGV, trains régionaux et trains de banlieue ont chacun leurs spécificités qui ont pu être prises en compte par la méthode utilisée ici. On peut imaginer l'appliquer à d'autres cas :

- trains commerciaux d'une gare passante : on peut supposer que le cas des trains à l'arrivée sera similaire à ceux étudiés pour la gare Montparnasse. Cependant, la gare Montparnasse étant terminale, les retards des trains à l'arrivée sont souvent rattrapés par une occupation en gare plus longue, ce qui conduit à une très bonne ponctualité au départ. Dans le cas d'une gare passante, les retards au départ sont hautement corrélés aux retards à l'arrivée, avec éventuellement quelques minutes rattrapées si le temps d'occupation prévu était un peu plus large que nécessaire, ou au contraire une amplification si trop de voyageurs arrivent pendant le retard, occasionnant un temps supplémentaire à l'embarquement.
- Correspondances : l'étude des retards peut être utilisée spécifiquement pour les correspondances afin de quantifier le risque que le temps d'échange soit insuffisant.
- Trains fret : ces trains sont connus pour avoir régulièrement des retards importants. Des estimations d'intervalle d'heure d'arrivée est intéressante pour les clients autant que pour les entreprises ferroviaires ou le gestionnaire d'infrastructure. L'estimation d'heure d'arrivée ou de temps de trajet est par ailleurs un problème connu du fret routier.

Durées d'évènements : la constitution d'une grille horaire repose sur l'estimation des temps nécessaires pour effectuer le trajet, comme les temps de trajets entre deux gares et les temps d'occupation des quais. Ces durées peuvent fortement varier selon le contexte, par exemple en cas d'affluence en gare plus de temps sera nécessaire pour permettre aux usagers de descendre ou monter. Elles sont en général fixées en avance selon le scenario, mais on pourrait envisager comme ici d'étudier la probabilité de dépassement de cette valeur prédéfinie et l'utiliser pour augmenter la robustesse des grilles.

Affluence : les données d'affluence (APC : *automatic passenger counting*) sont de plus en plus populaires car elles permettent de mieux estimer la demande et de fournir des informations aux usagers concernant la congestion de leur trajets. Le degré de certitude est important car il va conditionner des décisions.

4.7 Conclusion

Ce chapitre a présenté une méthodologie de traitement des données de retards de trains à l'échelle d'une grande gare afin de quantifier l'incertitude autour des horaires d'arrivée ou de départ selon le sens du mouvement. Les modèles sont évalués à la fois selon la fiabilité des distributions proposées, mais aussi selon la variété de probabilités prédites. En effet, l'objectif étant d'organiser les affectations de voies et d'itinéraires en gare de manière à anticiper les conflits liés à des retards prévisibles, il faut être en mesure de détecter là où les marges sont insuffisantes et là où elles sont le moins utiles.

Le prochain chapitre détaille une approche possible d'intégration de ce module d'estimation du risque de retard pour augmenter la robustesse des opérations en gare. Ces planifications se font de manière journalière, ce qui signifie que les modèles décrits ici sont appliqués pour les données du mois à venir puis pour une journée donnée, chaque train se verra affecté sa distribution de probabilité de retards estimée en amont.

Deuxième partie

Intégration des données de retard pour la robustesse en gare

Chapitre 5

Approches stochastiques pour l'affectation de quais en gare

Les gares sont des points de création et d'accumulation de retards dans le réseau en raison de la saturation de l'infrastructure et de l'interaction entre plusieurs sous-systèmes de la production ferroviaire (retours de maintenance, interface avec les voyageurs, etc.). Une planification des occupations de voies robuste aux petites perturbations dans les horaires est recherchée afin d'anticiper les phénomènes de propagation et contribuer ainsi à un meilleur fonctionnement global du service ferroviaire.

Les voies en gare sont affectées très en amont et adaptées au fur et à mesure pour ajouter de nouvelles circulations, modifier les horaires ou matériels, ou encore intégrer les contraintes et préférences des différents acteurs. Ce chapitre traite de ce problème d'adaptation sous incertitude des données d'entrée. L'incertitude sur les horaires des mouvements commerciaux est exprimée grâce aux probabilités de retards estimées dans le chapitre précédent, ce qui permet de quantifier la qualité des affectations. Étant donné qu'une solution initiale est disponible grâce à la construction en amont, une approche de recherche locale est mise en place afin de proposer des modifications ponctuelles permettant de gagner en robustesse.

Le vocabulaire utilisé dans la suite de ce chapitre est introduit ci-dessous.

Éléments d'infrastructure d'une gare : la gare est composée de différents types de voies :

- les voies à quai : elles sont positionnées dans le bâtiment voyageur et servent au stationnement des trains, notamment pour les arrêts commerciaux.
- les voies en ligne : elles marquent la sortie du périmètre de gare et ne font pas partie du problème de routage ici. L'utilisation de ces voies est liée à la provenance ou à la destination du train.
- les voies en avant gare : ces voies servent au routage des trains entre leur voie en ligne jusqu'à leur voie à quai.

Trains :

- On appelle **mouvement** une arrivée ou un départ d'une circulation en gare. A la différence du chapitre précédent, un **train** est composé de deux mouvements et d'une occupation de voie à quai entre les deux.
- Comme dans le chapitre précédent, les circulations commerciales qui transportent des passagers sont différenciées des mouvements dits techniques qui circulent à vide.

— les opérations en gare incluent des *coupes* et *accroches*, c'est-à-dire la séparation ou la liaison de deux rames. Ainsi dans le cas d'une coupe, deux trains arrivent ensemble sous le même numéro de train, puis sont séparés à quai pour former deux circulations indépendantes. Dans certains cas, un train est à la fois en coupe et en accroche sur le même quai.

La section 5.1 détaille le problème métier du point de vue de SNCF réseau, y compris l'explicitation des contraintes et la résolution actuelle. Le cadre de travail et les choix de modélisation adoptés sont ensuite introduits dans la section 5.2. Les algorithmes utilisés sont décrits dans la partie 5.3 et les résultats de l'expérimentation sont ensuite détaillés dans la partie 5.4. Enfin, les limites et perspectives de ces travaux sont discutées dans la section 5.5.

5.1 Modélisation du problème ferroviaire

On présente dans cette partie l'organisation pratique adoptée à la SNCF pour la circulation des trains en gare ainsi que la modélisation de ce problème pour sa résolution par des approches mathématiques.

5.1.1 Le cas de SNCF Réseau

5.1.1.1 Les graphiques d'occupation des voies

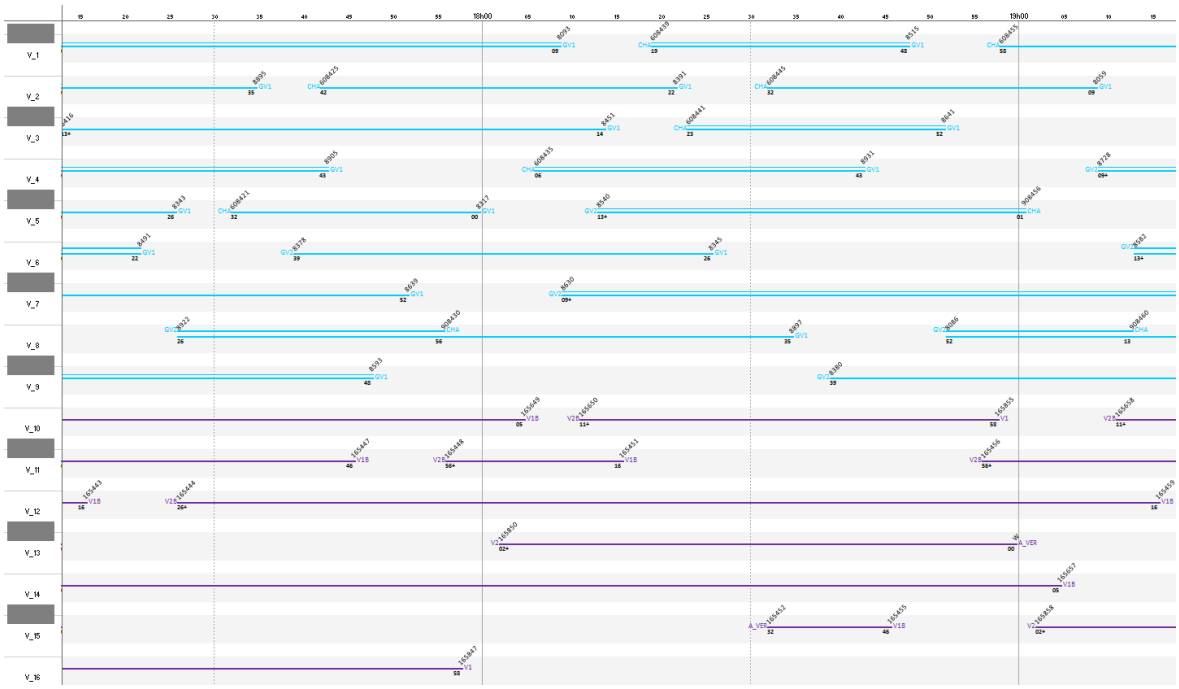

FIGURE 5.1 – Extrait d'un GOV de la gare Montparnasse

Un graphique d'occupation des voies, ou GOV, est un document contenant la planification des occupations de voies dans les gares. Un exemple est donné en figure 5.1, avec les voies à quai en

ordonnée et le temps en abscisse, et chaque trait tracé correspond à l'occupation de la voie par un train. Ces graphiques sont élaborés en trois phases [102] :

— Pré-construction : une ébauche de GOV est réalisée près de un an en avance afin d'anticiper la capacité résiduelle de la gare en fonction des premières versions du plan de transport.

— Conception : le GOV est affiné de quelques mois à quelques semaines avant les opérations.

— Adaptation : le GOV peut être adapté jusqu'à la veille à 17h des opérations pour faciliter au mieux le déroulement de la mise en service.

Ce graphique ne représente que l'occupation des voies en gares, cependant l'affectation doit également éviter les conflits entre les occupations de voies en avant gare.

5.1.1.2 Les règles de conception

Les GOV sont soumis à différentes contraintes, dont certaines sont très spécifiques à la gare et à la politique en oeuvre. Le paramétrage des contraintes dépend des normes de tracé imposées, c'est-à-dire les durées minimales à respecter entre les mouvements. Elles sont définies localement pour chaque gares et pour différents scenarios, selon par exemple le type de circulation, le sens ou les voies impliquées [10]. Ces contraintes se ramènent aux familles suivantes.

Les contraintes d'exploitation :

— Réoccupation : un temps minimal est nécessaire entre deux occupations d'une même voie à quai (cf figure 5.2).

— Voies imposées et interdites : tous les trains ne sont pas autorisés à stationner sur tous les quais.

— Espacements en ligne : cette contrainte concerne le temps nécessaire entre les trains sur les lignes. Elle ne dépend pas du GOV mais doit être vérifiée en conception pour éventuellement faire modifier les grilles horaires.

— Infrastructure : l'infrastructure affectée doit être compatible avec les rames reçues, par exemple la longueur des trains doit être inférieure à la longueur du quai assigné et des voies empruntées.

— Cisaillements : les contraintes de cisaillements imposent de respecter un temps minimal entre deux utilisations d'itinéraires qui empruntent une zone commune d'infrastructure (voie, aiguille,...). Ce temps, appelé aussi norme de cisaillement, est une contrainte de sécurité qui permet le dégagement du premier train pour que le passage du second sur son itinéraire soit possible, comme présenté sur la figure. 5.3 où le train sur l'itinéraire vert est bloqué pour laisser l'autre train circuler.

FIGURE 5.2 – Contraintes de réoccupation des quais

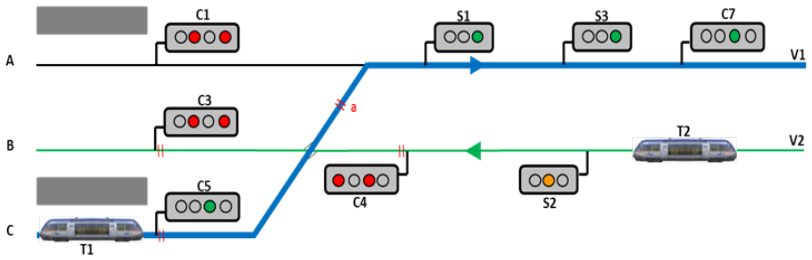

FIGURE 5.3 – Schéma de cisaillement d'itinéraires

Contraintes d'escales :

— Flux voyageurs : un temps minimal doit être respecté entre deux mouvements sur un même quai afin de limiter le nombre de passagers présents dessus au même moment, ou éviter que les voyageurs montent dans le mauvais train.

— Correspondances : des trains en correspondance sont placés sur le même quai ou proches pour faciliter les flux de voyageurs.

Contraintes commerciales : elles sont imposées par les entreprises ferroviaires, comme par exemple la contrainte d'utilisation d'un quai muni de bornes automatiques pour certains TGV ou les portiques de validation filtrant l'accès aux quais des Transiliens.

5.1.1.3 Gestion de la robustesse ferroviaire

La robustesse d'un système se caractérise de manière générale par sa capacité à résister aux imprévus. Il n'existe pas de définition unique de la robustesse ferroviaire d'un point de vue industriel, cependant elle est très souvent rattachée au respect du niveau de service annoncé. On considère par exemple souvent qu'un système robuste est un système assurant de bons chiffres de ponctualité et régularité, ou retournant rapidement à la normale.

Pour atteindre cet objectif, plusieurs politiques de robustesse se distinguent sur le terrain [5, 10] :

— Capacité résiduelle : on cherche à conserver en permanence des voies disponibles pour permettre la circulation de trains qui se rajouteraient, par exemple en cas de retard important. Cet aspect est particulièrement pris en compte en phase de pré-construction, avant que les horaires des trains ne soient figés. Si les contraintes de capacité résiduelle ne sont pas respectées, il peut arriver que des modifications soient apportées au plan de transport.

— Marges sur les normes de tracé : on va chercher un graphique où les normes de tracé (cisaillement, réoccupation) sont augmentées d'une marge permettant d'absorber de petites variations. Cette stratégie a par exemple été appliquée en gare de Montparnasse où les agents ont commencé à réaliser des GOV où on ajoutait, quand c'était possible, une marge égale au retard moyen afin d'assurer l'absorption de retards fréquents.

— Protection des trains au départ : pour limiter les apparitions et propagations des retards à l'échelle du réseau entier, les trains au départ sont priorisés et protégés. En effet, plusieurs études internes à la SNCF ont montré qu'avoir un retard même minime au départ était fortement corrélé au fait d'avoir un retard au terminus.

Ces stratégies sont parfois contradictoires : augmenter les espacements entre les trains en imposant des normes plus importantes réduit la capacité résiduelle, et au contraire, augmenter la capacité résiduelle conduit à trop utiliser certains itinéraires ou quais sur lesquels il y a peu de marge et où les conflits sont plus probables.

5.1.1.4 Adaptations en opérationnel

En opérationnel, un certain nombre d'aléas peuvent perturber les planifications et générer des retards. Les aléas nécessitant l'adaptation du GOV sont principalement les retards des trains en circulation, les changements d'équilibres demandés par les entreprises ferroviaires (un autre matériel roulant est utilisé pour faire circuler le train) et les dérangements d'installations. Pour intégrer ces aléas, les trains peuvent être réceptionnés sur un autre quai, emprunter un autre itinéraire, ou encore être retardés, par exemple pour laisser passer une autre circulation.

En particulier, les mouvements techniques sont très régulièrement retardés ou anticipés quand ils sont horairisés pour permettre un retour à la normale plus rapide. Dans beaucoup de gares, ces mouvements ne sont par ailleurs pas horairisés et leur gestion est purement opérationnelle.

5.1.2 OpenGOV

OpenGOV est un outil développé à la DGEX Solutions chez SNCF Réseau permettant de vérifier le respect des règles, visualiser et optimiser des graphiques d'occupation des voies.

Pour une journée donnée, cet outil prend en entrée les données suivantes :

— Caractéristiques des trains planifiés : horaires, types de train, longueurs, etc.

— Description de l'infrastructure : les gares sont modélisées de manière macroscopique en énumérant tous les itinéraires possibles d'arrivée et de départ. Chacun de ces itinéraires est constitué d'une voie en ligne (voie qui connecte le périmètre de gare aux principales lignes ferroviaires), plusieurs voies d'avant gare et une voie à quai.

— Description des contraintes : les normes de cisaillement sont données en fonction du type de circulation, de son sens et du niveau de cisaillement entre les itinéraires concernés, de même pour les normes de réoccupation, les voies interdites et dédiées.

— Règles d'optimisation : poids et pénalités associés à chaque scénario.

— Une solution courante : il s'agit d'un GOV chargé pour étude, qui peut ensuite être analysé, vérifié ou encore optimisé dans l'outil.

En particulier, les horaires des trains et le GOV courant proviennent de fichiers appelés *GROIX* qui sont les documents de travail des gares. Ces données sont extraites, puis centralisées avec les autres données dans un fichier unique par jour. A partir de ce fichier, OpenGOV permet de visualiser la solution courante, comme dans la figure 5.1, d'identifier les conflits d'infrastructure ou les non respects des normes en gare et en avant gare.

Le module d'optimisation d'OpenGOV a pour but d'améliorer la solution courante donnée en entrée. Il se base sur une modélisation classique par graphes de compatibilité, dont le principe sera détaillé dans la partie 5.1.4. Les variables représentent l'affectation d'une combinaison d'itinéraires aux mouvements qui composent un train. Les affectations simultanées entre deux mouvements sur des ressources conflictuelles sont interdites. Un autre niveau de contraintes est défini dans l'outil avec des contraintes molles qui déconseillent des itinéraires. Ces contraintes sont relâchées par une variable pénalisée en objectif. L'outil utilise le solveur Local Solver pour résoudre le modèle. Ce solveur permet l'optimisation d'objectifs dans un ordre lexicographique pour marquer les différents niveaux de priorité, avec tout d'abord les affectations d'un maximum de trains, la pénalité selon le type de train non placé, les pénalisation du premier niveau de contraintes molles de cisaillements puis du second niveau, etc. Ces deux niveaux sont définis par un système de normes dans les données d'entrée qui représentent des seuils d'espacements en deçà desquels les affectations sont considérées comme non robustes.

5.1.3 Cas d'étude de la gare Montparnasse

La gare de Paris Montparnasse est une gare terminale composée de 28 voies à quai. De manière générale, le type de circulation détermine la zone de gare à laquelle le train est affecté. Si l'exploitation peut varier selon les cas, le fonctionnement usuel est :

— les voies 1 à 9 accueillent les TGV et OuiGO

— les voies 10 à 17 sont réservées au Transiliens

— les voies 18 à 24 réceptionnent tous les types de trains (TGV, TN, TER, IC)

— les voies 25 à 28, qui correspondent à la gare de Vaugirard et reçoivent les Intercités, et réceptionnaient aussi les OuiGO jusqu'en fin 2018.

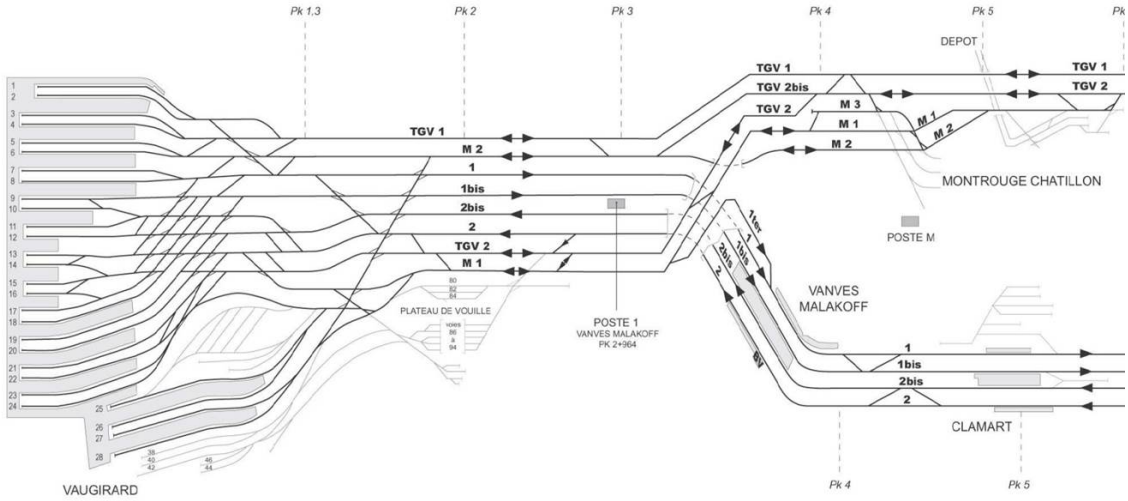

FIGURE 5.4 – Schéma de la gare Montparnasse

La gare est schématisée dans la figure 5.4 avec les voies à quai à gauche, l'arrivée des voies en ligne TGV en haut du schéma et TN/TER/IC en bas, et les voies en ligne qui connectent la gare aux deux technicentres de Montrouge et Chatillon et qui sont empruntées par les trains techniques. L'avant gare se situe entre le bâtiment voyageur et le point kilométrique 6 (indiqués *Pk i* sur le schéma). L'outil OpenGOV utilise une modélisation de l'infrastructure avec 228 itinéraires d'arrivée et 300 itinéraires de départ, chacun composé d'une voie à quai, une voie en ligne, et trois voies en avant-gare qui connectent les deux.

De nombreux mouvements techniques sont opérés en gare, par exemple pour réceptionner une rame venant du technicentre et qui est réutilisée pour un train commercial à l'origine de la gare. Ces circulations techniques sont bien moins contraintes que les circulations commerciales, et la voie de réception dépend surtout des contraintes du mouvement complémentaire.

5.1.4 Modélisation par un problème de graphe

Comme exposé dans le chapitre 2, il est classique de modéliser le problème d'affectation des voies comme un problème de graphe. Soit $G = (V, E)$ un graphe tel que chaque sommet représente un

triplet réalisable (t, i_a, i_d) avec t un train, i_a un itinéraire d'arrivée et i_d un itinéraire de départ. Le graphe de compatibilité G se construit en connectant toute paire de sommets dont les affectations sont réalisables en même temps (respect des contraintes). Par exemple, deux sommets correspondant au même quai pour des horaires ne respectant pas les contraintes de réoccupation ne pourront pas être reliés. Un GOV réalisable est donc une clique du graphe de faisabilité, c'est-à-dire un sous-ensemble de noeuds tous reliés entre eux, et dont le cardinal est égal au nombre de trains.

Le problème peut être vu comme un problème d'optimisation en introduisant des poids sur les noeuds (préférences sur les choix des voies à quai ou itinéraires) ou sur les arêtes (pénalités sur les successions de trains sur des itinéraires ou voies). Dans ce cas là, on cherche une solution de cardinalité égale au nombre de trains et de poids optimal.

La dimension de ces graphes est cependant très importante : par exemple dans le cas de la gare Montparnasse, près de 400 trains sont prévus chaque jour, ils sont autorisés à occuper en moyenne une dizaine de voies à quai parmi les 28 de la gare et chacune d'elle est reliée à la voie en ligne du train par plusieurs itinéraires d'arrivée et de départ. Par ailleurs le graphe obtenu est très dense étant donné que certains trains ne sont jamais en conflit (par exemple un train quittant la gare le matin est automatiquement connecté avec tous les noeuds correspondant à des trains arrivant après son départ de la gare). Des stratégies pour limiter autant que possible le temps de calcul malgré la taille du graphe sont abordées dans ce chapitre.

L'exemple simplifié ci-dessous illustre le cas d'une gare à deux quais, trois voies avant gare et deux voies en ligne où circulent trois trains. La gare est schématisée dans la figure 5.5. La modélisation se fait en trois parties, la description de l'infrastructure, le paramétrage des règles opérationnelles et la construction du graphe à partir des horaires des trains. L'infrastructure est décrite par l'énumération des différents itinéraires admissibles donnée dans le tableau 5.1.

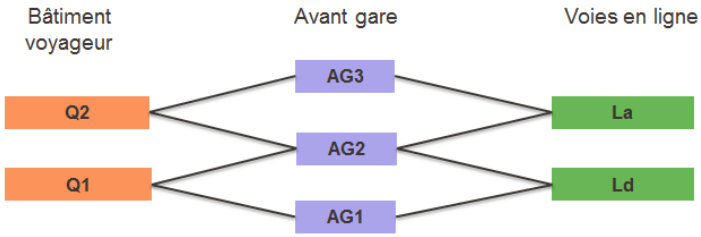

FIGURE 5.5 – Exemple simplifié : modélisation de la gare

nom	Itinéraire arrivée	nom	Itinéraire départ
a_1	La - AG2 - Q1	d_1	Ld - AG1 - Q1
a_2	La - AG2 - Q2	d_2	Ld - AG2 - Q1
a_3	La - AG3 - Q2	d_3	Ld - AG2 - Q2

TABLEAU 5.1 – Exemple simplifié : énumération des itinéraires

Les règles opérationnelles sont ensuite imposées, avec notamment les voies dédiées/interdites et les itinéraires admissibles selon le type de train, les normes de cisaillement et de réoccupation. Dans cet exemple on ne donne que deux contraintes simplifiées : la première triviale est de ne pas occuper un même quai en gare en même temps et la seconde de respecter les temps de cisaillements donnés dans le tableau 5.2.

Trois trains sont considérés : le train *rose* prévu à $t_{a,r} = 3$, repartant à $t_{d,r} = 5$, le train *orange* arrivant à $t_{a,o} = 6$ et repartant à $t_{d,o} = 10$ et le train *vert* arrivant à $t_{a,v} = 9$ et repartant à $t_{d,v} = 13$.

	a_1	a_2	a_3	d_1	d_2	d_3
a_1	3	2	1	0	0	0
a_2	2	3	3	0	0	0
a_3	1	3	3	0	0	0
d_1	1	0	0	3	3	1
d_2	2	2	0	3	3	2
d_3	2	2	1	1	2	3

TABLEAU 5.2 – Exemple simple : normes d'incompatibilité des itinéraires

Le graphe de compatibilité G est construit de la manière suivante. On ajoute un noeud par association admissible train/itinéraire d'arrivée/itinéraire de départ (dans la figure 5.6, on conserve le code couleur des noms des trains, le premier chiffre correspond à l'indice de l'itinéraire d'arrivée et le second à l'indice de l'itinéraire de départ). On place une arête entre deux noeuds si les contraintes d'incompatibilité sont respectées (cf 5.2) et s'ils n'occupent pas un quai en gare en même temps. Le graphe 5.6 représente donc les associations compatibles (sans pondérations).

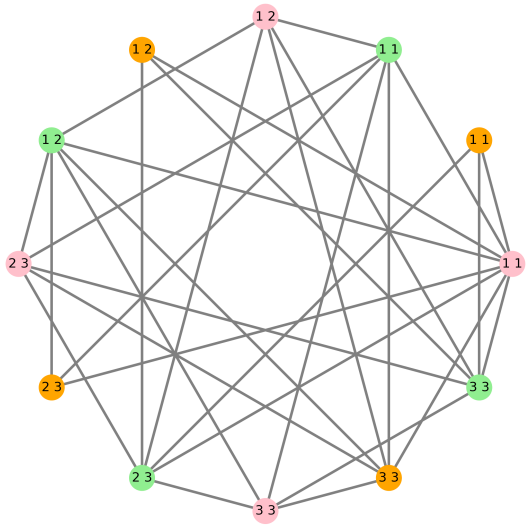

FIGURE 5.6 – Exemple simplifié : graphe de compatibilité

Une solution réalisable correspond à une clique de taille 3. Dans ce graphe on peut trouver 4 solutions réalisables mais de qualité variable. Par exemple deux solutions réalisables sont comparées dans la figure 5.7. La proposition de gauche est réalisable mais demande d'occuper le même quai par le train rose et le train orange avec une unité de temps de différence alors que l'autre option en laisse 4, ce qui est préférable en cas de retard. Au niveau de l'avant gare cependant, les solutions sont équivalentes.

Ajouter des pondérations aux arêtes et chercher une clique de cardinalité maximale et de poids minimal permet d'affiner les préférences pour choisir des solutions plus robustes.

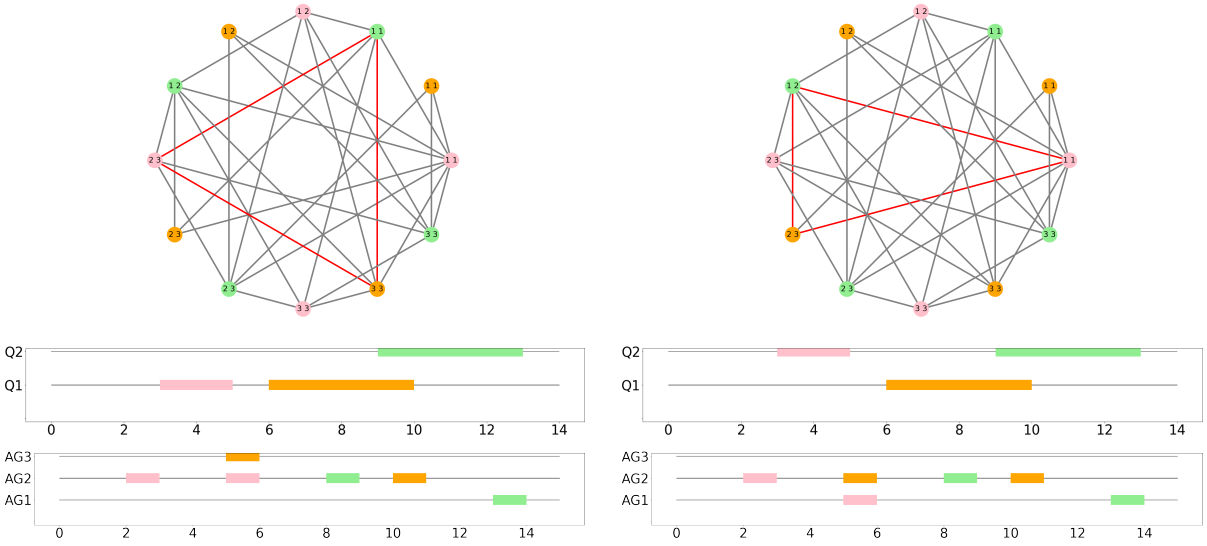

FIGURE 5.7 – Exemple simplifié : comparaison de deux solutions

5.1.5 Contribution

Ce chapitre présente une méthode d'adaptation des graphiques d'occupation des voies afin de gagner en robustesse face aux retards. Les deux principales contributions par rapport à ce qui a été fait ultérieurement, en particulier chez SNCF Réseau, sont l'utilisation de recherche locale pour la résolution de ce problème industriel et l'intégration du risque d'échec des contraintes pour améliorer la robustesse. Ce problème est vu comme un problème de théorie des graphes en cherchant une clique dans le graphe de compatibilité incertain.

La recherche locale est une méthodologie intéressante ici pour plusieurs raisons. La première est que la fonction objectif de ce problème n'est pas exacte, les pondérations utilisées ont pour but de guider la solution vers un état de robustesse globale en minimisant le nombre de conflits et le volume de retards générés, cependant ce volume est estimé approximativement. La seconde raison est que ce problème nécessite de traiter des instances de grande taille pour lesquelles la recherche de solution exacte pourrait être très chronophage. En particulier, on cherche à travailler ici sur la probabilité de conflit entre deux trains grâce aux résultats du chapitre précédent. Comme les distributions sont différentes pour chaque train, la probabilité de conflit doit être calculée pour chaque paire de sommets considérés simultanément, ce qui est très long à faire de manière exhaustive. Enfin, on dispose d'une solution initiale puisque nos travaux visent uniquement à adapter le GOV pour le rendre plus robuste, et non à le concevoir. Avec un algorithme de recherche locale, on va partir de cette solution initiale et explorer les voisinages successifs pour l'améliorer sans la perturber inutilement.

5.2 Optimisation orientée robustesse : cadre de travail

5.2.1 Cadre théorique

Cette partie rappelle plusieurs éléments théoriques concernant à la fois l'incertitude en recherche opérationnelle, les problèmes de cliques et la gestion de la dimension en recherche locale.

5.2.1.1 Incertitude sur les données

En pratique, les données d'entrées sont rarement fixes et parfaitement connues, et leurs variations peuvent avoir de fortes conséquences sur la qualité effective des solutions produites. Deux branches principales de l'optimisation combinatoire s'intéressent à la gestion de l'incertitude des paramètres dans la résolution de problèmes, l'optimisation stochastique et l'optimisation robuste [21, 25].

L'optimisation stochastique : ce domaine présuppose que la distribution des données est connue et l'utilise pour produire des solutions qui soient faisables avec une forte probabilité tout en optimisant l'espérance de l'objectif. L'espace des solutions n'est pas restreint mais des variables de recours sont éventuellement ajoutées pour permettre d'adapter la solution à l'aléa rencontré. La fonction objectif contient les coûts associés aux variables de décision pour résoudre le problème nominal et les coûts attendus générés par les variables de recours. Le cas le plus classique est celui des programmes en deux étapes (*two stages*) [25]. Ces formulations sont cependant très complexes, avec de nombreuses variables de recours à décider quand le nombre d'évènements aléatoires est grand. S'il est fini, on peut formuler un programme déterministe équivalent.

Un cas particulier de l'optimisation stochastique est à rappeler ici : les contraintes de chance. En pratique, on peut accepter que certaines contraintes sur les variables de premier ordre ne soient pas réalisables avec une probabilité bornée, ce qui peut s'exprimer avec une contrainte de ce type :

$$\mathbb{P}\left[A^i(\omega)x \leq b^i(\omega)\right] \geq 1 - \epsilon^i \tag{5.1}$$

Optimisation robuste : l'objectif est de produire une solution robuste à plusieurs évènements explicitement définis à l'aide d'un ensemble d'incertitude (non nécessairement à nature stochastique). La solution doit rester faisable face aux variations des paramètres d'entrée au sein de cet ensemble, sans possibilité d'ajustement avec des variables de recours. Une approche classique est de chercher la solution dont l'objectif est le meilleur dans le pire des cas, c'est-à-dire avec les variations des paramètres les plus pénalisantes pour l'objectif. Afin d'éviter un surconservatisme des solutions qui risquent d'être sous-optimales la plupart du temps pour se protéger contre le pire des cas qui est improbable, Bertsimas et Sim [21] ont proposé une modélisation par budgetisation de l'incertitude : les variations des paramètres sont bornées par un budget, évitant de prendre en compte des situations irréalistes où tous les paramètres évolueraient simultanément.

Approches intermédiaires : ces deux approches présentent des limites. L'optimisation stochastique est plus souple, s'adapte au risque ainsi qu'aux coûts de recours dans la prise de décision et optimise par rapport à un coût moyen. Cependant elle nécessite la connaissance des aléas possibles ainsi que leur distribution de probabilité, et le problème devient vite trop important pour être résolu efficacement. Au contraire, un programme d'optimisation robuste est plus facile à résoudre et s'adapte bien au problème nominal, cependant il repose sur une bonne description de l'ensemble d'incertitude et donne des solutions souvent trop conservatrices car on se protège contre le pire des cas, et non en moyenne. Selon l'enjeu des décisions, ce conservatisme peut être préférable.

Quelques alternatives mixtes utilisant des caractéristiques des deux approches ont été proposées. Un état de l'art plus complet peut être trouvé dans la revue de Gabrel, Moral et Thiele [76]. Deux méthodes qui ont également été appliquées dans un cadre de recherche opérationnelle ferroviaire et de robustesse aux retards peuvent être rappelées : la robustesse légère et la robustesse de récupération.

La robustesse légère (ou *light robustness*) présentée par Fischetti et Monaci [71] utilise le cadre classique d'optimisation robuste de Bertsimas et Sim [21] dans lequel la déterioration maximale de l'objectif qui est autorisée est bornée. Des variables de recours sont ajoutées au modèle pour permettre

à certaines contraintes de ne pas êtres respectées mais leur activation est pénalisée en objectif. Contrairement à un programme d'optimisation stochastique, ces variables ne représentent pas de stratégies de correction mais identifient les contraintes qui ne seraient plus faisables dans le pire des cas.

La robustesse de récupération (*recoverable robustness*) caractérise une solution qui reste faisable ou facilement réparable face à un ensemble d'évènements incertains [120]. Comme en optimisation stochastique, le problème nominal est optimisé en prenant en compte les solutions de recours malgré une information déterministe. Le modèle est cependant optimisé par rapport au pire des cas pour un ensemble d'aléas prédéterminé et borné.

5.2.1.2 Analyses prescriptives et approches data-driven

Plusieurs études récentes ont pris le parti d'exploiter les larges bases de données collectées à l'aide d'outils d'apprentissage statistique afin de pallier le manque d'information sur les données et de mieux aiguiller la prise de décision, ou à l'inverse d'appliquer des méthodes d'optimisation mathématique pour apporter de la valeur aux données analysées. Une classification de ces différentes recherches peut être trouvée dans l'article de Lepenioti et al [119]. Ils recensent les techniques d'analyse prescriptive, c'est-à-dire l'ensemble des méthodes utilisées pour recommander une décision optimale dans un contexte stochastique en se basant sur les résultats préliminaires obtenus par description ou apprentissage de données. On revient ici sur deux types de travaux qui rentrent dans cette logique.

Prescription prédictive : Bertsimas et Kallus [20] ont défini un cadre de travail d'optimisation stochastique basé sur l'utilisation de données. L'objectif est d'estimer la meilleure décision conditionnellement à des variables dites auxiliaires en exploitant des observations passées $(x_1, y_1), ..., (x_N, y_N)$. Le problème qu'on cherche à résoudre est de trouver la politique réduisant l'espérance des coûts conditionnellement à X : $z^*(x) \in \mathcal{Z}(x) = \text{argmin}_{z\in\mathcal{Z}} \mathbb{E}\left[c(z, Y)|X = x\right]$.

Dans ce cadre de travail, z^* est estimé en utilisant les observations passées similaires du point de vue des valeurs des variables auxiliaires par un système de poids calculés grâce à des méthodes d'apprentissage statistique. L'estimation se fait avec $\hat{z}_N^{local}(x) \in \text{argmin}_{z\in\mathcal{Z}} \sum_{i=1}^{N} w_{N,i}(x)c(z, y^i)$, où les poids $w_{N,i}(x)$ sont calculés avec des méthodes comme les forêts aléatoires, k plus proches voisins ou méthodes de noyaux. Par exemple, seuls les coûts des observations passées appartenant aux k plus proches voisins de x sont pris en compte pour estimer z^*. Dans le cas de forêts aléatoires, seules les observations qui appartiennent au moins une fois à un même noeud terminal de la forêt que x sont utilisées pour estimer les coûts, et leur contribution est pondérée par la fréquence relative de la valeur au sein du noeud, moyennée sur la forêt.

Les auteurs ne recommandent pas l'approche naïve consistant à exploiter les prédictions ponctuelles $\hat{m}(X)$ de Y pour mesurer les coûts. Choisir la décision z qui minimise les coûts $c(z, \hat{m}(x))$ ne prend pas en compte la variété des aléas possibles, ce qui donne des résultats rapidement sous-optimaux.

Ensemble d'incertitude : la composition de l'ensemble d'incertitude est primordiale : un ensemble incluant des évènements très peu probables va contraindre l'espace des solutions pour rien. Plusieurs approches récentes prennent le parti d'exploiter les données disponibles pour apprendre sur la distribution des aléas et permettre de réduire l'ensemble d'incertitude. Par exemple, Tulabandhula et Rudin [166] proposent d'utiliser du Machine Learning pour construire l'intervalle de variation de chaque paramètre incertain. Des estimations de quantiles de la loi conditionnellement aux variables explicatives sont utilisées comme bornes de ces intervalles. Bertsimas, Gupta et Kallus [19] identifient des caractéristiques de la distribution des données par des séries d'hypothèses et vérifications par des tests statistiques, et utilisent les connaissances rassemblées sur la distribution pour construire un ensemble d'incertitude de taille plus réduite que les méthodes traditionnelles et qui apporte des garanties en probabilité de faisabilité des contraintes.

5.2.1.3 Le problème de clique maximum et ses variantes

Soit $G = (V, E)$ un graphe. Le problème de clique maximum consiste à trouver un sous-graphe complet de G de cardinalité maximum. Ce problème est connu pour être NP complet. C'est un problème classique parmi les plus étudiés dans la littérature. Un état de l'art plus approfondi sur les problèmes de clique peut être trouvé dans les articles suivants [26, 177]

Ce problème peut se formuler très simplement par le programme suivant :

$$\begin{aligned} &\max \sum_{i=1}^{n} x_i \\ &x_i + x_j \leq 1, \forall (i,j) \in \bar{E} \\ &x_i \in \{0,1\}, \forall i = 1..n \end{aligned} \tag{5.2}$$

Le problème de clique maximum dans G est équivalent à la recherche d'un stable maximum dans le graphe complémentaire $\bar{G}$, c'est-à-dire un sous ensemble de sommets de V non reliés entre eux.

Une variante classique de ce problème consiste à associer des pondérations w_i aux noeuds $i \in V$ et à chercher une clique de poids maximum (*maximum weight clique problem*). On peut la formuler par :

$$\begin{aligned} &\max \sum_{i=1}^{n} w_i x_i \\ &x_i + x_j \leq 1, \forall (i,j) \in \bar{E} \\ &x_i \in \{0,1\}, \forall i = 1..n \end{aligned} \tag{5.3}$$

Les pondérations peuvent également être associées aux arêtes du graphe, et le poids de la solution est alors la somme des poids des arêtes comprises dans la clique (*edge-weighted clique problem* [125]). Une formulation quadratique de ce problème est :

$$\begin{aligned} &\max \sum_{(i,j) \in E} w_{i,j} x_i x_j \\ &x_i + x_j \leq 1, \forall (i,j) \in \bar{E} \\ &x_i \in \{0,1\}, \forall i = 1..n \end{aligned} \tag{5.4}$$

En pratique, il est courant que les données d'entrées ne soient pas fiables. Dans le cadre des problèmes de graphes, cela se caractérise par exemple par des arêtes non sûres. On parle de graphes incertains ou graphes probabilistes. Par exemple, le problème de clique probabiliste maximum vise à trouver un ensemble de sommets tel que la probabilité que ces sommets forment une clique soit supérieure à un seuil $\theta \in [0, 1]$ fixé [106, 134, 182]. Soit $\tilde{G} = (V, \tilde{E})$ un graphe aléatoire avec l'ensemble des arêtes, chacune étant associée à une probabilité p_e de succès. Les arêtes sont supposées indépendantes les unes des autres. Ce problème s'écrit sous la forme d'un programme avec contrainte de chance pour trouver $\mathcal{C}$ un ensemble de sommets tel que :

$$\begin{aligned} &\max |\mathcal{C}| \\ &\mathbb{P}\left[\mathcal{C} \text{ est une clique dans } \right] \geq \theta \end{aligned} \tag{5.5}$$

Le cadre classique du problème de clique maximum a été très étudié dans la littérature, les pistes de résolutions sont rappelées rapidement. Un état de l'art plus complet est disponible dans les articles

suivants [26, 92, 177]. Concernant la résolution exacte, les approches existantes utilisent en général un algorithme de Branch and Bound directement sur le problème de clique, ou sur ses variantes (coloriage, MaxSAT, etc) ou des algorithmes d'énumération. Les approches exactes nécessitent cependant des temps de calcul trop importants pour des graphes de grande taille.

La recherche locale est une alternative aux approches exactes. Son principe est de partir d'une solution initiale qui est modifiée localement par des opérations élémentaires. Ces opérations permettent de générer un ensemble de nouvelles solutions appelé *voisinage* à partir de la solution courante. On peut ainsi passer de voisinage en voisinage afin d'optimiser la solution. Cette technique de recherche permet une amélioration rapide et efficace, cependant il n'y a pas de garantie d'optimalité.

Dans le cas des cliques, la plupart des heuristiques proposées sont dédiées au problème de clique maximum. Wu et Hao [177] les classifient en deux catégories. La première consiste à chercher itérativement une clique de taille k, et à augmenter k à chaque fois qu'on en trouve une tout en parcourant des voisinages pour éventuellement trouver des solutions non réalisables qu'on va réparer par échanges de noeuds. L'autre approche définit comme voisinage toutes les cliques atteignables grâce à une opération élémentaire et cherche une solution de taille maximum en parcourant ces voisinages. Les voisinages les plus utilisés sont l'échange d'un noeud de la clique avec un nouveau noeud ($SWAP$), la suppression d'un noeud ($DROP$) et l'ajout d'un noeud dans la clique (ADD). D'autres techniques ont également été appliquées, comme par exemple une perturbation avec des ajouts forcés où un sommet est ajouté de force et les noeuds incompatibles sont supprimés de la solution [16], ou des voisinages de type k-opt où k modifications sont effectuées à la fois [16, 107]. Dans le cas de graphes incertains, les probabilités associées aux arêtes permettent de calculer la probabilité d'échec d'un sommet de la clique. Cette probabilité peut être utilisée pour guider la recherche locale en identifiant les noeuds à ajouter ou à enlever en priorité [182].

L'alternance de structures de voisinages est courante pour permettre d'atteindre un plus grand nombre de solutions candidates. Par exemple, de nombreuses études utilisent des *recherche de plateau* où des échanges de sommets sont opérés après des phases d'ajout. Ces échanges n'améliorent pas la qualité de la solution mais permettent de varier les voisinages et d'accéder à de nouvelles solutions [92].

5.2.1.4 Recherche locale sous contrainte de dimension

Cette partie recense des stratégies classiques de recherche locale pour appréhender des problèmes où la taille des données d'entrée est très importante, notamment pour les problèmes de cliques.

Grands graphes : les problèmes d'optimisation dans les grands graphes, et en particulier les problèmes de clique, ont gagné en importance ces dernières années. Pour la plupart, ces graphes ont un très grand nombre de sommets mais sont peu denses, comme pour des graphes de réseaux sociaux. La combinatoire ne permet pas d'appliquer les approches existantes, et la recherche locale s'impose là encore pour obtenir des temps de calcul acceptables. Cai et Lin [38] étudient la recherche heuristique de clique de poids maximum. La stratégie décrite consiste à alterner entre des phases de construction de la solution optimale et des phases de réduction du graphe. En particulier, un sommet v est supprimé du graphe si on constate que la borne supérieure du poids d'une clique contenant v est inférieure au poids de la clique courante. Wang et al [174] améliorent l'efficacité de la recherche par réduction des voisinages considérés en rejetant en avance les sommets non prometteurs et en choisissant aléatoirement k paires de sommets pour constituer le voisinage de la solution.

Parcours de voisinages importants : plusieurs études se sont penchées sur les problèmes de recherche locale où la taille du voisinage est très importante ou exponentielle en la taille des données, ce qui pose des problèmes de temps de calcul quand tout le voisinage doit être exploré à chaque itération. Cela part du constat qu'utiliser de grands voisinages permet d'atteindre de meilleures solutions, mais

cela requiert de l'explorer efficacement. Ahuja et al [13] proposent une revue des algorithmes utilisés dans ce cadre, appelé *very large-scale neighborhood search* (VLSN), qui est reprise par Pisinger et Ropke [149]. Ils identifient trois familles principales d'approches. La première stratégie est d'utiliser des méthodes à profondeur variable (*variable-depth methods*) qui explorent partiellement par heuristiques des voisinages complexes, composés à partir de séries de mouvements élémentaires successifs. La seconde stratégie utilise des méthodes de flux pour identifier les voisinages prometteurs. La troisième consiste à réduire le problème d'origine à des sous-problèmes qu'on peut résoudre en temps polynomial.

Les méthodes à profondeur variable s'adaptent bien aux problèmes de cliques où on explore le voisinage des solutions à distance k, c'est-à-dire obtenues par k ajouts, suppressions ou échanges de sommets de la clique initiale. Katayama et al [107] proposent un algorithme consistant à créer une série de solutions en ajoutant et supprimant des sommets selon une heuristique jusqu'à critère d'arrêt. La meilleure solution de la série est ensuite utilisée comme point de départ de la procédure qui est répétée.

5.2.2 Formulation

5.2.2.1 Robustesse ferroviaire :

La robustesse des affectations de voies en gare n'est pas bien définie, ce qui demande un travail préliminaire pour formuler le problème de manière adéquate. Comme cela a été vu dans le chapitre 2, la robustesse d'une solution se caractérise en général par sa capacité à limiter les retards créés en opérationnel et elle est souvent contrôlée en imposant des espacements suffisants entre les trains.

Les articles de référence ne posent cependant pas la question de la nature stochastique des horaires utilisés pour construire les contraintes, et la question de faisabilité de la solution est souvent ignorée ou contournée bien que fondamentale en pratique. En effet, la saturation des gares est telle que les solutions sont rarement faisables, avec le plus souvent des conflits en avant gare. Ces conflits mettent les agents sous pression en opérationnel car ils doivent adapter les solutions. Il existe plusieurs leviers d'action, par exemple avec les mouvements techniques qui sont très flexibles. Ces circulations sont en général en provenance ou à destination d'un centre de maintenance, en cas de conflit il est courant d'anticiper ou de retarder le mouvement pour le faire à un moment plus opportun.

On choisit ici d'intégrer ces deux aspects. Le non respect de contraintes est fortement pénalisé mais autorisé, et les pénalités sont prévues pour varier selon le degré de priorité des circulations impliquées. La nature stochastique des horaires est intégrée via des probabilités d'échec des contraintes de faisabilités.

Une solution est habituellement considérée comme robuste quand elle génère une quantité limitée de retards secondaires. L'estimation des retards créés en gare par une solution est cependant irréaliste à partir des seules distributions de retards car cela imposerait d'intégrer les règles de régulation dans la modélisation pour estimer les conséquences des interactions entre les trains, par exemple pour reconstituer les chaînes de propagation. On souhaite donc se concentrer ici sur la production de solutions robustes aux retards dans le sens où on minimise localement le risque de conflit et leur amplitude.

5.2.2.2 Structure du graphe de compatibilité

La structure classique d'un graphe de compatibilité pondéré est utilisée ici. Soit $G = (V, E)$ un tel graphe. Chaque sommet $v = (t, i_a, i_d) \in V$ correspond à une affectation d'un train t à une route complète (i_a, i_d). On part du principe que tout sommet généré est réalisable (les itinéraires partagent un quai que le train a le droit d'occuper, et sont reliés aux voies en ligne d'arrivée et de départ empruntées par le train), et que deux sommets correspondant à des occupations des voies en gare

incompatibles ne sont pas connectés entre eux. Chaque train est considéré comme placé dans un GOV réalisable si on arrive à trouver une clique de cardinal le nombre de train dans G. Comme expliqué précédemment, les affectations incompatibles en avant gare sont autorisées ici, c'est-à-dire que l'arête est tout de même générée dans le graphe mais associée à un poids strictement positif.

L'incertitude est liée aux horaires d'arrivée et de départ des trains, ce qui régit l'utilisation des ressources et compromet la faisabilité de la solution. En cas de retard, une arête présente dans le graphe pourrait ne plus être valable car les circulations entreraient alors en conflit. Grâce aux estimations de distributions de probabilités calculées dans le chapitre précédent, on est en mesure d'associer à chaque arête $e \in E$ une probabilité d'échec $p_e \in [0, 1]$.

On choisit de modéliser les règles de priorités entre les trains (trains au départ, trains techniques, etc) ainsi que les préférences de faisabilité et de robustesse grâce à un système de pondération des arêtes. Pour chaque $e = ((t_1, i_{1,a}, i_{1,d}), (t_2, i_{2,a}, i_{2,d})) \in E$, un poids w_e est associé afin de représenter les interactions de t_1 et t_2. Ces interactions peuvent être multiples, par exemple entre les deux arrivées, entre l'arrivée de t_1 et le départ de t_2, etc.

Un tel graphe a les particularités suivantes :

— Il est T-parti avec T le nombre de trains planifiés. En effet, le graphe se décompose donc en T groupes de sommets avec un groupe par train, au sein desquels les sommets sont tous disjoints entre eux. On peut l'observer sur la figure 5.6 où les trains sont représentés par des couleurs différentes.

— Il est de grande dimension : environ 400 trains circulent chaque jour, selon le type de circulation ils peuvent occuper environ une dizaine de voies différentes parmi les 28 de la gare, parfois plus, et en fonction de la voie en ligne empruntée et du quai, il y a parfois plusieurs itinéraires d'arrivée ou de départ envisageables. Selon l'instance on a entre 10000 et 20000 sommets environ.

— Beaucoup de trains ne se gênent jamais et sont donc toujours connectés, ce qui implique que le graphe est dense et qu'il contient des sous graphes k-partis complets.

— Selon la stratégie de pondération utilisée, beaucoup d'arêtes ont un poids nul car les circulations impliquées n'interagissent pas. C'est par exemple le cas pour les trains occupant une même voie à des moments distincts de la journée, ou n'occupant pas les mêmes quais sans risque de cisaillement en avant gare.

5.2.2.3 Positionnement théorique

Le cadre classique de la programmation mathématique, notamment robuste et stochastique, est peu adapté au cas présent. Les instances sont de grande dimension et la fonction objectif n'est pas définie clairement, mais approximée en fonction des objectifs industriels. Optimiser de manière exacte serait long et peu adéquat, ce qui dispense d'exprimer le problème par un programme linéaire robuste ou stochastique classique. D'autres arguments peuvent être avancés en défaveur de ces formulations.

L'optimisation robuste se concentre sur des problèmes d'optimisation où l'incertitude sur les paramètres d'entrée s'exprime par des ensembles de valeurs bornés et compacts, en général déconnectés de la distribution réelle de l'aléa. Dans notre cas, c'est la validité d'une contrainte qui est sujette à incertitude, et non les coefficients du programme, et cette incertitude est quantifiable grâce aux travaux statistiques réalisés en amont, ce qui rend l'utilisation de budget inappropriée.

La notion de recours, que ce soit dans un cadre d'optimisation stochastique ou de robustesse de récupération, est intéressante : les conflits sont nombreux, mais les règles de régulation en opérationnel, comme la propagation de retards, règles de priorité ou itinéraires de substitution, peuvent être modélisées. Étant donné la taille du graphe, le calcul des chaînes de propagation liées à chaque scenario de retard est cependant irréaliste. En effet, on ne peut considérer toutes les combinaisons de retards des différents trains en un temps raisonnable, et ce malgré la troncature des retards.

Les modèles de recherche de clique probabiliste maximum par contraintes de chances ne sont également pas adaptés ici : l'objectif est de trouver une clique faisable avec une garantie en probabilité, ici on impose sa cardinalité et on cherche à optimiser conjointement la faisabilité, les règles de priorité et le risque local de conflit. La probabilité que la solution proposée soit une clique est nulle d'emblée. Par ailleurs, de telles formulation du problème imposent que les échecs des arêtes soient indépendants entre eux, ce qui n'est pas le cas ici. En effet, un même retard peut faire échouer plusieurs arêtes.

L'approche qu'on propose ici est proche de la méthodologie de Bertsimas et Kallus [20] qui vise à estimer l'espérance du coût lié à la solution grâce aux observations du passé. En particulier, une des alternatives qu'ils proposent est d'utiliser des forêts aléatoires pour calculer cette quantité. Pour cela, ils pondèrent les coûts liés à chaque évènement incertain par sa fréquence relative au sein des noeuds terminaux associés à la réalisation des variables explicatives puis moyennée sur la forêt. C'est la même approche qui est utilisée dans cette thèse pour estimer la probabilité associée à chaque valeur de retard avec des forêts aléatoires. Une étape de validation de la qualité des estimations est cependant ajoutée dans ces travaux.

Le problème qu'on souhaite résoudre peut s'exprimer simplement à l'aide du programme ci-dessous. On note $G = (V, E)$ le graphe de faisabilité décrit ci dessous au sein duquel on cherche une clique de cardinalité maximum, et on note $\tilde{E} \subset E$ l'ensemble des arêtes du graphe soumises à une probabilité d'échec supérieure à un seuil ϵ (dans ces travaux on considère toutes les arêtes de probabilité d'échec non nulle et ceux où l'arête n'est pas faisable mais autorisée) .

$$
\begin{aligned}
\min \sum_{e \in \tilde{E}} & w_e z_e \\
& \sum_{v \in V : t \in v} x_v = 1, \forall t \in \mathcal{T} \\
& x_{v_1} + x_{v_2} \leq 1, \forall (v_1, v_2) \notin E \\
& x_{v_1} + x_{v_2} \leq 1 + z_e, \forall (v_1, v_2) \in \tilde{E} \\
& x_v \in \{0, 1\}, \forall v \in V \\
& z_e \in \{0, 1\}, \forall e \in \tilde{E}
\end{aligned}
\tag{5.6}
$$

Cette modélisation, comme dans un contexte de robustesse légère, utilise des variables de recours pour relâcher les contraintes qui contrôlent les affectations à risque de conflit liés aux retards. Ces contraintes peuvent être vues comme des contraintes de chance imposant que la probabilité de réussite de chaque arête $e = (v_1, v_2)$ soit supérieure à $1 - \epsilon$ si v_1 et v_2 sont dans la clique finale.

Les arêtes de $E \backslash \tilde{E}$ connectent des paires de trains n'interagissant pas, par exemple si les infrastructures sont compatibles, si les horaires sont suffisamment éloignés ou si le risque d'échec est considéré comme faible. Cette formulation ne diffère pas de celle utilisée classiquement pour gérer la robustesse des GOV, notamment dans OpenGOV, cependant le sous-graphe incertain $\tilde{G} = (V, \tilde{E})$ des conflits potentiels est créé différemment puisqu'il intègre ici le risque d'échec et non seulement des écarts déterministes aux normes recommandées.

5.2.2.4 Gestion de la dimension

Les méthodes de résolution exactes ont été mises de côté ici compte tenu de la taille des instances et du côté approximatif de la fonction objectif. Le problème est donc abordé avec des méthodes de recherche locale. Les problèmes de clique ont été beaucoup traités dans la littérature, avec de nombreuses approches différentes. Deux méthodes sont implémentées ici, une déterministe se basant sur des restrictions du problème afin de contrôler la taille du voisinage et de guider efficacement la

recherche, et une autre randomisée qui utilise des voisinages de tailles plus grandes mais qui sont explorés partiellement et aléatoirement.

La taille du graphe impacte tout de même le temps de calcul. Plusieurs stratégies ont été identifiées pour optimiser la solution de manière efficace :

— L'exploration de tous les voisinages n'est pas nécessaire, on peut s'attendre à ce que certaines parties de la solution n'aient pas besoin d'amélioration. Cela peut être le cas durant certaines heures creuses où il y a moins de ressources partagées, et donc moins de conflits potentiels mais également de plus grands voisinages à explorer sans gain conséquent. Une stratégie est donc de filtrer les solutions pour explorer en priorité les ajustements intéressants.

— On dispose d'un GOV initial qu'on veut adapter. On part donc de cette première solution et on construit le graphe au fur et à mesure que les voisinages sont explorés. Ainsi, les parties de l'espace des solutions qui ne sont pas explorées ne sont jamais générées dans le graphe, ce qui limite sa taille et réduit les calculs. En effet, chaque sommet ajouté au graphe nécessite un temps important car il faut étudier un à un tous les autres sommets représentants un conflit potentiel, évaluer si la paire est compatible, et enfin calculer le poids de l'arête et sa probabilité d'échec en fonction des probabilités individuelles des trains impliqués et des normes de tracés correspondant aux itinéraires/voies/types de trains en jeu. Plus le graphe est grand, plus l'ajout d'un sommet est long.

— De manière triviale, l'exploration du voisinage suit la structure du problème. En l'occurrence, un sommet de la clique ne peut être remplacé que par un sommet de sa partie dans la partition du graphe.

D'autres stratégies utilisées dans des cas similaires (cf 2.2 et 5.2.1) sont intéressantes mais ne peuvent pas être utilisées ici :

— Cai et Lin [38] procèdent par réduction de graphe en supprimant les sommets dont la borne supérieure de la solution en les incluant est inférieure à la valeur de la solution courante. Dans nos travaux, on pourrait considérer que la borne inférieure d'un sommet comme la plus petite valeur possible de la solution si on incluait ce sommet . Cependant tant que le graphe n'est pas construit entièrement, cette borne est approximative et ne permet pas de supprimer de sommets. On pourrait également considérer la réduction du graphe en supprimant les arêtes qui engendrent des solutions sous-optimales mais à moins que le poids seul de l'arête soit supérieur à la valeur de la solution courante, le calcul de la borne inférieure n'est pas possible.

— Dewilde et al [68] proposent de supprimer avant l'optimisation les sommets train/routes qui entreraient en conflit avec des trains n'ayant qu'une route possible. Ici, les sommets ne sont créés que s'ils peuvent être ajouté à la solution. Des cas qui ne sont jamais réalisables ne sont pas considérés.

5.2.3 Environnement et instances

5.2.3.1 Format de l'instance

Données de GOV : Lors de la phase de conception, puis d'adaptation, les affectation des voies en gare et avant gare peuvent être extraites dans un document appelé fichier GROIX. Ce fichier est ensuite téléchargé dans l'outil OpenGOV qui génère un fichier excel dans lequel on peut récupérer :

— la description de l'infrastructure (itinéraires d'arrivée et de départ, voies les composant, etc.)

— la liste des trains avec :

  — le type de circulation arrivée et départ (technique, TGV, etc.)

Voie	Train	Heure	Ori/Dest	time	Retard	Proba1	Proba2	Proba3	Proba4	Proba5	Proba6	Proba7	Proba8	Proba9	Proba10	Proba11	Proba12
7	8300	07:26:00	TO/PMP	266	0	0.593586	0.0849472	0.112381	0.0549486	0.0263257	0.0130369	0.0236407	0.0104348	0.0156177	0.00485121	0.00535324	0.0145624
5	8371	07:27:00	PMP/LR	266	0	0.929661	0.0201825	0.0077375	0.0152647	0.0139983	0.00146191	0.00682582	0.00312185	0.0014383	0.000210526	9.72222e-05	0
13	164460	07:28:30	RBT/PMP	268	2	0.349141	0.252564	0.149577	0.0833736	0.0575273	0.0304579	0.039167	0.0149768	0.0137887	0.0038343	0.00559234	0
11	164162	07:30:00	SVR/PMP	270	3	0.594544	0.123559	0.0550422	0.0811658	0.0882276	0.0356685	0.016596	0.000308006	0.00072009	0.00358834	0.000580772	0
10	164161	07:31:00	PMP/SVR	271	0	0.903736	0.0542713	0.0297391	0.00957148	0.000678025	0.00200435	0	0	0	0	0	0
16	164862	07:32:30	DX/PMP	272	10	0.285693	0.216717	0.124746	0.0602029	0.145475	0.0378017	0.0162021	0.0218225	0.0412586	0.0121121	0.0379699	0
4	8331	07:35:00	PMP/PS	274	0	0.947526	0.0346823	0.00607931	0.00394943	0.00347635	0.00072772	0.000239445	0.00111285	0.000911734	0.000708284	0.00058659	0
19	862460	07:35:30	NGR/PMP	275	2	0.244131	0.199863	0.177524	0.152165	0.0626808	0.0492058	0.0376304	0.0136245	0.0131206	0.00628076	0.00686795	0.00137354
19	862460	07:35:30	NGR/PMP	275	2	0.244131	0.199863	0.177524	0.152165	0.0626808	0.0492058	0.0376304	0.0136245	0.0131206	0.00628076	0.00686795	0.00137354
12	164662	07:37:00	MTE/PMP	277	4	0.309648	0.269187	0.184164	0.0964277	0.0738902	0.0214811	0.0294838	0.00446817	0.00580147	0.00348308	0.00196567	0
25	3411	07:38:00	PVA/GRV	277	0	0.787522	0.116837	0.0443593	0.00906413	0.0157027	0.00881197	0.00881865	0.00888455	0	0	0	0
8	8603	07:40:00	PMP/BT	279	0	0.80148	0.0826687	0.0247976	0.0347609	0.00517008	0.0100251	0.000488793	0.0189511	0.00106226	0.0190876	0.00150754	0
15	164661	07:39:00	PMP/MTE	279	0	0.752477	0.127665	0.0531328	0.0321616	0.0268078	0.00775617	0	0	0	0	0	0
22	862409	07:40:00	PMP/CH	280	0	0.901929	0.063012	0.0173003	0.00543413	0.00222793	0.00610324	0.00136578	0.00262757	0	0	0	0

FIGURE 5.8 – Données d'entrée

- les indices des itinéraires d'arrivée et de départ empruntés et le numéro de voie à quai occupée dans la solution initiale
- les trains avec lesquels ils sont en coupe/accroche

— la description des règles opérationnelles appliquées, dont les normes de réoccupation et les normes de cisaillement en fonction des itinéraires empruntés et des caractéristiques des circulations impliquées

Données de retard : pour la même journée, on récupère à partir du module présenté dans le chapitre précédent les distributions de retards pour chaque train commercial de la journée. Certains trains ne font pas partie de la base de données mais sont quand même enregistrés dans le fichier GROIX, par exemple en cas de suppression, et pour lesquels on applique une loi de probabilité par défaut égale à la fréquence relative des retards observés l'année précédente pour ce type et sens de circulation. Les trains techniques sont considérés comme à l'heure et n'ont donc pas de distribution associée. Un exemple est donné dans la figure 5.8 qui contient une fusion simplifiée d'un fichier de planification d'occupation des voies et des probabilités estimées ainsi que retards observés correspondants.

5.2.3.2 Règles de construction

Chaque noeud du graphe correspond à la combinaison d'un train, d'un itinéraire d'arrivée et d'un itinéraire de départ et n'est généré que s'il respecte les contraintes d'affectations. Les trains arrivés la veille, et qui sont donc déjà présents sur site à minuit, ont un quai fixe.

Pour chaque ajout d'un nouveau sommet au graphe, les autres sommets sont étudiés pour éventuellement ajouter une arête si les deux trains et voies à quai correspondants peuvent être affectés sans conflit d'occupation. En pratique, on vérifie les conditions suivantes avant de connecter deux sommets :

— ils doivent correspondre à des trains différents

— si les trains sont liés par une opération de coupe ou accroche, les noeuds doivent partager le même quai

— si les sommets correspondent au même quai mais que les trains ne sont pas en coupe ou accroche, les horaires ne doivent pas se chevaucher et être espacés d'au moins la norme de réoccupation.

5.2.3.3 Pondération des arêtes : quelques exemples

Une fois l'arête ajouté, elle peut être pondérée de plusieurs manières. Si le problème d'affectation des voies peut s'exprimer de manière générique, il existe de nombreuses subtilités et variantes pour répondre aux contraintes industrielles. Cette partie recense quelques approches possibles pour guider la solution grâce aux poids des arêtes. On peut les séparer en trois familles : les composantes de l'objectif liées à la faisabilité de la solution (conflits à quai ou en avant gare), celles liées à la robustesse aux retards (espacements et espérance du retard propagé) et enfin les pondérations et pénalités permettant de répondre à des besoins métiers spécifiques. Ces pénalités peuvent être soit additives soit multiplicatives. Les calculs des pondérations des arêtes effectivement appliqués dans ces travaux sont détaillé dans la sous-section 5.2.3.4.

Conflits à quai : dans ces travaux, deux noeuds ne sont pas reliés s'ils correspondent à un conflit à quai, à l'exception des conflits présents dans la solution d'origine. Les conflits à quais doivent être exclus autant que possible, mais en pratique on peut imaginer autoriser une flexibilité si le temps de conflit est inférieur à un seuil donné (par exemple s'il manque une minute pour respecter la norme), et créer l'arête correspondante en lui affectant une pénalité très forte.

Conflits en avant gare : les conflits en avant gare sont très fréquents, en partie car les normes de conceptions utilisées sont parfois surestimées et les agents en charge de la conception des GOV autorisent parfois des conflits sachant qu'en réalité la configuration est réalisable. Plusieurs travaux du domaine décident de ne pas intégrer les arêtes correspondant à des conflits en avant gare. Ici, on crée les arêtes à partir du moment où l'affectation à quai est réalisable, mais on pénalise en fonction du nombre de minutes de retards qui seront créées si les itinéraires sont incompatibles.

Espacements entre les trains : cette composante très classique vise à pénaliser les utilisations successives d'une ressource dans un intervalle de temps réduit. En pratique, cela revient à ajouter un poids sur une arête si la marge disponible entre les deux passages de trains sur un point (quai, cisaillement avant gare) est inférieure à un seuil, voire à faire dépendre ce poids de la marge elle-même. Il a été montré qu'une meilleure répartition des espacements entre les trains favorise la robustesse aux petits retards dans la mesure où ils sont mieux absorbés et moins propagés [46]. Cette stratégie est également utilisée dans le module d'optimisation de OpenGOV, et correspond à la stratégie industrielle de répartition de la capacité résiduelle.

Espérance du retard propagé : cette composante de l'objectif a pour but d'optimiser la robustesse en prenant en compte la nature stochastique des retards. En modélisant la distribution de probabilité de retards, il est possible d'estimer l'espérance du retard propagé d'un train à l'autre indépendamment du reste de la solution

Probabilité de conflit : cette quantité se calcule en évaluant la probabilité que les marges soient insuffisantes entre deux mouvements à partir des distributions de retards des circulations impliquées.

Minimisation du nombre de changements : la construction des graphiques d'occupation des voies par les gestionnaires d'infrastructure est un processus long, souvent de plus d'un an. Dans le cas d'une adaptation d'une solution pré-existante, il est parfois préférable qu'un nombre limité de modifications soit effectué, ou alors qu'une modification n'ait lieu que si le gain est conséquent. Cela permet de ne pas bousculer les habitudes des usagers et des agents et de garantir la stabilité de la

solution. On peut rajouter une pénalité forfaitaire sur les arêtes reliées à un sommet qui n'est pas dans la solution d'origine pour limiter les échanges à faible valeur ajoutée.

Poids selon le type de circulation : plusieurs trains sont partiellement constitués d'un mouvement technique. Les circulations techniques sont moins prioritaires car il y a beaucoup moins d'enjeux commerciaux sur la ponctualité ce qui les rend plus flexibles. Étant donné la congestion de la gare, il est parfois plus intéressant d'autoriser des conflits en avant gare ou à quai avec un train technique sachant qu'en cas de besoin il sera possible d'ajuster son horaire en temps réel pour respecter les différentes contraintes.

Poids selon le sens du mouvement : dans la politique actuelle de gestion opérationnelle des circulations, il est demandé de favoriser les départs à l'heure par rapport aux arrivées (programme H00). En effet, un retard même minime à l'origine a tendance à fragiliser la circulation et baisser ses chances d'être à l'heure au terminus. L'enjeu est donc plus important que pour les arrivées où le retard accumulé reste contrôlable. Il est ici possible d'ajouter ou de multiplier le poids de l'arête par une pénalité dans le cas où il y a un conflit ou un risque de conflit impactant un train au départ.

5.2.3.4 Choix de la pondération

Dans ces travaux on choisit de faire dépendre les poids des arêtes de deux types de pénalités, la faisabilité et le risque de conflit local, calculés pour chaque paire de mouvements. Le poids d'une arête est la somme des pénalités calculées par paires de mouvements en interaction, par exemple arrivée/arrivée ou arrivée/départ.

Pénalité de faisabilité : elle correspond au nombre de minutes manquantes pour respecter les contraintes de cisaillement imposées par les itinéraires des noeuds correspondants. La composante $fais$ est la somme des minutes perdues à chaque interaction :

$$fais_e = c_{AA,e} fais_{AA,e} + c_{AD,e} fais_{AD,e} + c_{DA,e} fais_{DA,e} + c_{DD,e} fais_{DD,e}$$

Pour une de ces interactions, si le train t_1 emprunte l'itinéraire i_1 à l'heure h_1, puis t_2 emprunte i_2 à h_2 avec $h_2 \geq h_1$ et si la norme de cisaillement entre i_1 et i_2 vaut n alors la contribution de cet interaction à $fais$ vaut $\min(0, n - (h_2 - h_1))$. Si les deux mouvements sont espacés de plus de n alors la contribution est nulle.

Les différentes contributions peuvent être pénalisées selon le sens du mouvement ou le type de circulation. En particulier il est classique que les circulations techniques soient concernées par les attributions non réalisables dans la mesure où leurs opérations sont plus flexibles. Ici les pénalités sont divisées par cinq si une des circulations est technique. En raison de la politique de protection des trains au départ à la SNCF, on fixe $c_{AD,e} = 2$ et $c_{DD,e} = 2$ pour éviter les cas où le second mouvement qui est impacté par un conflit est un train au départ. Si le second mouvement est un technique, ces pénalités valent 1.

Pénalité de risque de conflit : Cette quantité est calculée à partir des distributions de retards des mouvements qui composent les trains en interaction. Pour deux mouvements a et b de lois de retard $\mathcal{P}_a$ et $\mathcal{P}_b$ séparés par une marge m, la probabilité de conflit entre les deux mouvements se calcule avec la formule :

$$\begin{aligned} p_{a,b} &= \mathbb{P}\left[r_a - r_b > m\right] \\ &= \sum_{i=m+1}^{\max_a} \sum_{j=0}^{\min(\max_b, i-m)} \mathbb{P}\left[r_a = i\right] \mathbb{P}\left[r_b = j\right] \end{aligned} \tag{5.7}$$

avec r_a et r_b les variables aléatoires des retards respectifs des mouvements a et b supposées indépendantes, $\max_a$ et $\max_b$ les valeurs de troncature de ces variables. On notera que l'hypothèse d'indépendance des retards est utilisée. Cette hypothèse est pertinente quand les circulations en jeu sont hétérogènes, comme par exemple le départ d'un TER et l'arrivée d'un TGV ont a priori des retards indépendants, mais cette hypothèse est moins valable pour des trains fréquents circulant sur la même ligne, comme les TN. La troncature et l'étude des petits retards permet une meilleure indépendance des retards sur des lignes distinctes et peu fréquentes, comme les TGV ou IC/TER.

Ces probabilités sont ensuite agrégées en une pénalité de risque de conflit :

$$risq_e = c_{AA,e} risq_{AA,e} + c_{AD,e} risq_{AD,e} + c_{DA,e} risq_{DA,e} + c_{DD,e} risq_{DD,e}$$

avec $risq$ une fonction par paliers construite à partir de la probabilité qui permet de pénaliser plus fortement les hautes valeurs de risque.

On notera que cette quantité est une estimation locale du risque, c'est-à-dire que l'influence des retards des autres trains n'est pas prise en compte, seules les distributions de retards en dehors de l'avant gare sont utilisées. Par exemple, un train qui est en conflit de faisabilité avec une circulation le précédant subira une augmentation de son risque de retard qui n'est pas connue ici. De tels cas sont minimisés en évitant les enchaînements infaisables ou à haut risque afin de casser les chaînes de propagation, mais leurs conséquences ne peuvent être estimées simplement.

Pour les mouvements de trains doubles, c'est-à-dire avant une coupe ou après une accroche, la contribution de l'interaction est divisée par deux, voire par quatre pour ne pas compter plusieurs fois le même défaut de robustesse dans le graphe.

Exemple : un cas d'illustration est donné dans la figure 5.9 avec deux trains stationnant en gare sur deux quais différents. On suppose ici qu'il y a deux paires de mouvements en interaction, les autres étant trop lointaines pour occasionner un risque de conflit ou défaut de faisabilité. La première interaction est entre un TGV et un mouvement technique, avec 1 minute de défaut de faisabilité, et le second entre deux départs commerciaux. Si par exemple $p_{DD} = 0.3$, on a $fais = 0.2 \times 1$ et $risq = 2 \times (0.3 + 1)$, les pénalités supplémentaires étant liées à la présence d'un mouvement technique et à celle d'un train au départ ensuite, la valeur 1 dans la fonction $risq$ est liée à la pénalité par paliers visant à pénaliser plus fortement les risques élevés. On obtient alors un poids $w = 2.8$.

5.2.3.5 Construction du graphe initial

Le graphe initial est construit à partir de la solution qu'on souhaite adapter. Cette solution est considérée comme faisable, cependant plusieurs conflits en gares sont presque systématiquement détectés. Ces conflits sont ignorés dans la construction du graphe qui doit être complet ici, donc ces arêtes conflictuelles sont imposées avec une pénalité forfaitaire additionnelle valant 5. Le reste du graphe initial est construit en utilisant les fonctions de pénalités décrites précédemment.

5.2.3.6 Voisinages possibles

Les méthodes de recherche locale fonctionnent en effectuant des transformations élémentaires sur la solution courante. Ces transformations permettent de définir un voisinage de la solution courante qu'on va explorer pour améliorer l'objectif.

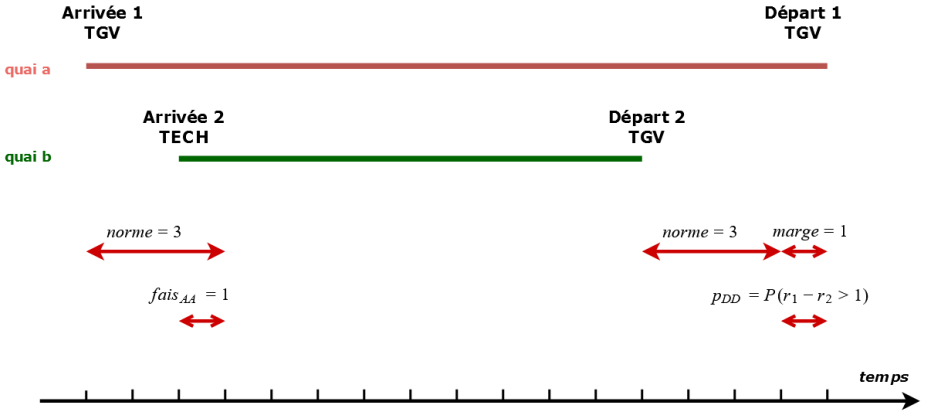

FIGURE 5.9 – Exemple calcul des pondérations

D'un point de vue ferroviaire, trois transformations élémentaires peuvent être considérées pour améliorer la qualité d'un GOV :

— changement d'itinéraire d'arrivée ou de départ

— changement de voie à quai

— modification des horaires des mouvements techniques

Dans une modélisation par graphe de compatibilité, les deux premiers voisinages reviennent à supprimer un noeud lié à un train de la clique pour en ajouter un autre. On ne considérera que ces deux voisinages pour la suite de ces travaux.

Les modifications des horaires de trains techniques sont très courantes, ce qui a également été observé dans le chapitre précédent (les données de retards des trains techniques ont une très grande variance, et des valeurs valeurs allant souvent de plusieurs heures d'avance à plusieurs heures de retards). En pratique, une fois qu'un train est prêt à la sortie du technicentre et que la gare a la capacité de le réceptionner à un moment favorable, l'horaire sera modifié. Ce type de modification concerne surtout la gestion opérationnelle des circulations, bien que certaines modifications puissent être anticipées. Deux alternatives sont possibles. La première consiste à modifier dynamiquement les propriétés des sommets. Elle n'est pas facilement compatible avec la structure de graphe car il faudrait recalculer tous les poids d'un noeud pour chaque modification d'horaire, et éventuellement supprimer ou ajouter des arêtes. Il est cependant envisageable de concevoir une heuristique préliminaire qui étudie les horaires de trains techniques et les ajuste en amont de l'étude pour résoudre des conflits de faisabilité ou assurer une marge minimale avec les autres circulations quand cela est possible. La seconde approche est de générer plusieurs noeuds identiques (même train, mêmes itinéraires) à la différence que les horaires sont différents. Les arêtes reliant ce noeud aux autres sommets du graphe sont alors calculées en tenant compte de ce nouvel horaire. Cette approche peut cependant considérablement augmenter la taille du graphe si l'amplitude de variation d'horaire autorisée est importante.

Deux types de voisinages sont bridés dans ces expérimentations. Le premier concerne les trains déjà présents sur site la veille, leur quai est considéré comme fixé ici. Le second est celui des trains en coupe ou accroche sur une voie. Les deux ou trois trains doivent être toujours sur le même quai, et les itinéraires doivent être les mêmes pour les mouvements liés. Changer cet aspect demande une implémentation différente qui n'est pas explorée ici.

5.3 Algorithmes utilisés

5.3.1 Introduction

Les algorithmes présentés ici exploitent la structure du graphe et recherchent par modifications locales successives des alternatives au GOV initial qui soient plus robustes, tout d'abord en résolvant certains conflits existants, mais également en utilisant les probabilités de retard pour éviter des situations présentant un risque de propagation.

5.3.1.1 Notations

T	Nombre de trains
L_T	liste des T trains planifiés dans la journée
$t \in L_T$	train, composé d'un mouvement d'arrivée et d'un mouvement de départ
$i_a \in I_a$,$i_d \in I_d$	Itinéraires d'arrivée et de départ
$G = (V, E)$	Graphe de compatibilité, avec V l'ensemble des sommets et E les arêtes
C	clique de G
n	sommet de G, composé d'un train et deux itinéraires

5.3.1.2 Fonctions de base

Fonctions de construction du graphe :

- $conflit\_quai(t_1, t_2)$: cette fonction renvoie VRAI quand les deux trains peuvent utiliser le même quai et FAUX sinon.
- $ajout\_noeud(G, t, i_a, i_d)$: cette fonction ajoute, s'il n'existe pas déjà, un noeud à G correspondant au train t et aux itinéraires (i_a, i_d) partageant un même quai admissible à t, et crée les arêtes pondérées avec les autres sommets de G. Elle renvoie le graphe G mis à jour et l'indice du noeud.
- $update(C, n)$: supprime le noeud courant occupé par le train de n dans la clique C et utilise n à la place.

Fonctions de score :

- $score(G, C)$ renvoie la somme des poids des arêtes du sous-graphe de G induit par la clique C.
- $score\_train(G, C, t)$, $score\_noeud(G, C, n)$: renvoient respectivement le score du train dans la solution (somme des arêtes connectées au sommet de t au sein du sous graphe formé par les sommets de la clique C et le score du noeud n s'il était inclus dans la clique C.
- $score\_nouvel\_iti(G, C, t, i)$: cette fonction calcule la contribution individuelle de l'itinéraire i s'il est utilisé dans C pour le train t. Elle est calculée en considérant un sommet fictif pour lequel n'y a pas d'autre itinéraire (train qui reste sur site) dans le calcul des poids des arêtes.

Fonctions d'exploration :

- $noeud\_courant(t, C)$ renvoie l'indice du noeud lié au train t dans la clique C.
- $quai(n)$ et $quai\_courant(t, C)$: la première renvoie le quai associé au noeud n et la seconde le quai utilisé par t dans la solution C.
- $Iti_a(n)$ et $Iti_d(n)$: renvoient respectivement les itinéraire d'arrivée et de départ du noeud n.
- $iti\_opt(G, C, t, q)$: renvoie les itinéraires d'arrivée i_a et de départ i_d utilisant la voie q et maximisant individuellement la fonction $score\_nouvel\_iti$.

— $noeuds\_alternatifs(G, C, t)$: renvoie tous les noeuds du graphe contenant t et pouvant être utilisés à la place du noeud courant dans la solution C.

— $quais\_admissibles(G, C, t)$: donne la liste des quais disponibles pour le train t dans la solution courante.

5.3.2 Algorithmes gloutons

5.3.2.1 Recherche d'itinéraire optimal

L'algorithme ci-dessous est un algorithme de liste qui étudie chaque train dans l'ordre de la liste pour éventuellement lui affecter un nouvel itinéraire maximisant le gain. Seuls les voisinages par changement d'itinéraires sont considérés car l'étude de chaque route possible pour tous les quais peut être très longue. C'est par ailleurs le voisinage le plus naturel d'un point de vue ferroviaire. Cet algorithme est surtout utilisé comme première amélioration de la solution initiale.

Ici, on utilise pour L la liste des trains triés par contribution décroissante à la solution courante, en calculant leur contribution comme la somme des poids des arêtes liées au train dans la clique.

Algorithme 1 Algorithme glouton d'optimisation des itinéraires

Données d'entrée : Solution initiale C_0, graphe initial G_0, nombre d'itérations autorisées $iter\_max$
procédure OPTIMISATION_ITINERAIRES(G_0, C_0)
  $C \leftarrow C_0$
  $G \leftarrow G_0$
  $ancien\_score \leftarrow score(G, C)$
  $score\_courant \leftarrow 0$
  $n\_iter \leftarrow 0$
  tant que $n\_iter \leq iter\_max$ **et** $score\_courant! = ancien\_score$ **faire**
    $ancien\_score \leftarrow score\_courant$
    Générer L
    pour $t \in L$ **faire**
      $s_t \leftarrow score\_train(G, C, t)$
      si $s_t > 0$ **alors**
        $q \leftarrow quai\_courant(t)$
        $i_a, i_d, s \leftarrow iti\_opt(G, C, t, q)$
        si $s < s_t$ **alors**
          $G, n \leftarrow ajout\_noeud(G, t, i_a, i_d)$
          $G, C \leftarrow update(C, n)$
        fin si
      fin si
    fin pour
  fin tant que
  Sortie : graphe G et solution C
fin procédure

5.3.2.2 Recherche parmi les voisinages déjà exploré

L'algorithme présenté précédemment nécessite un temps de calcul assez long puisque à chaque itération et pour chaque train il calcule le poids de tous les itinéraires admissibles. Par ailleurs, les mêmes noeuds sont souvent sélectionnés. On propose une alternative ici qui recherche uniquement

parmi les noeuds déjà générés. Le voisinage entier n'est donc pas étudié puisque certains itinéraires ou quais n'ont pas été testés, mais cette méthode permet de réparer rapidement une solution en explorant les noeuds alternatifs dont les poids sont déjà connus.

On utilise la fonction $noeuds\_alternatifs()$ qui prend en entrée un train t et renvoie tous les noeuds du graphe contenant t et pouvant être utilisés à la place du noeud courant dans la solution C.

Algorithme 2 Algorithme glouton dans le voisinage déjà explorés

Données d'entrée : Solution à optimiser C, graphe courant $G = (V, E)$, liste de trains L_T,
nombre d'itérations autorisées $iter\_max$
procédure OPTIMISATION(G, C)
    $n\_iter \leftarrow 0$
    tant que $n\_iter \leq iter\_max$ **et** $score\_courant \neq ancien\_score$ **faire**
        $ancien\_score \leftarrow score\_courant$
        Générer L
        pour $t \in L$ **faire**
            $s_t \leftarrow score\_train(G, sol, t)$
            si $s_t > 0$ **alors**
                $n_t \leftarrow noeud\_courant(t, C)$
                $V_t \leftarrow noeuds\_alternatifs(G, C, t)$
                pour $n' \in V_t$ **faire**
                    $C' \leftarrow C \cup \{n'\} \backslash \{n_t\}$
                    si $score(C') < score(C)$ **alors**
                        $C \leftarrow C'$
                        $n_t \leftarrow n'$
                    fin si
                fin pour
            fin si
        fin pour
    fin tant que
    Sortie : graphe G et solution C
fin procédure

5.3.2.3 Réduction de la distance à la solution initiale

On préfère souvent avoir une solution la plus proche possible de la solution initiale. Les algorithmes de recherche locale, en particulier les méthodes de voisinages variables, ont tendance à s'éloigner beaucoup de la solution initiale pour parcourir plus efficacement l'espace des solutions. Cependant, certaines modifications adoptées n'apportent pas d'amélioration. L'algorithme présenté ici vise à parcourir la solution afin de vérifier s'il est possible d'emprunter le quai ou les itinéraires prévus initialement sans baisse de qualité. Cet algorithme sera appliqué à la fin des métaheuristiques proposées ici.

Deux étapes sont réitérées jusqu'à convergence. La première passe en revue tous les trains pour lesquels il y a eu un changement de quai. S'il existe un noeud qui améliore ou maintient la solution en utilisant le quai d'origine, il est utilisé. Une fois tous les trains observés, le même raisonnement est appliqué aux itinéraires : pour chaque train, si utiliser l'itinéraire d'arrivée et/ou de départ initialement prévu ne dégrade pas la solution, la clique est mise à jour.

Algorithme 3 Algorithme de réduction de distance à la solution d'origine

Données d'entrée : Solution initiale C_0, solution courante C, graphe courant G
nombre d'itérations autorisées $iter\_max$
procédure REDUCTION_DISTANCE(G, C, G_0, C_0)
  $n\_iter \leftarrow 0$
  tant que $n\_iter \leq iter\_max$ **et** $score\_courant! = ancien\_score$ **faire**
    $ancien\_score \leftarrow score\_courant$
    pour $t \in L_T$ **faire**
      $n \leftarrow noeud\_courant(t, C)$
      $n_0 \leftarrow noeud\_courant(t, C_0)$
      $q_0 \leftarrow quai(n_0)$
      si $quai(n) \neq q_0$ **alors**
        $s_n \leftarrow score\_noeud(G, C, n)$
        $i_a, i_d \leftarrow iti\_opt\,(G, C, t, q_0)$
        $n_{t,i_a,i_d} \leftarrow ajout\_noeud(G, t, i_a, i_d)$
        si $score\_noeud(G, C, n_{t,i_a,i_d}) \leq s_n$ **alors**
          $C \leftarrow C \cup \{n_{t,i_a,i_d}\} \backslash \{n\}$
        fin si
      fin si
    fin pour
    pour $t \in L_T$ **faire**
      $n \leftarrow noeud\_courant(t, C)$
      $n_0 \leftarrow noeud\_courant(t, C_0)$
      $i_a, i_d \leftarrow Iti_a(n), Iti_d(n)$
      $i_{a,0}, i_{d,0} \leftarrow Iti_a(n_0), Iti_d(n_0)$
      si $i_a \neq i_{a,0}$ **alors**
        $s_n \leftarrow score\_noeud(G, C, n)$
        $n_{t,i_{a,0},i_d} \leftarrow ajout\_noeud(G, t, i_{a,0}, i_d)$
        si $score\_noeud(G, C, n_{t,i_{a,0},i_d}) \leq s_n$ **alors**
          $C \leftarrow C \cup \{n_{t,i_{a,0},i_d}\} \backslash \{n\}$
        fin si
      fin si
      si $i_d \neq i_{d,0}$ **alors**
        $s_n \leftarrow score\_noeud(G, C, n)$
        $n_{t,i_a,i_{d,0}} \leftarrow ajout\_noeud(G, t, i_a, i_{d,0})$
        si $score\_noeud(G, C, n_{t,i_{a,0},i_d}) \leq s_n$ **alors**
          $C \leftarrow C \cup \{n_{t,i_a,i_{d,0}}\} \backslash \{n\}$
        fin si
      fin si
    fin pour
  fin tant que
  Sortie : graphe G et solution C
fin procédure

5.3.3 Méthode Tabou

La méthode tabou est une métaheuristique consistant à interdire d'inverser une transformation élémentaire acceptée pendant plusieurs itérations. Des dégradations de la fonction objectif sont donc acceptées quand la solution correspond à un minimum local. La liste tabou des transformations interdites permet ainsi de ne pas visiter des solutions déjà rencontrées pendant une durée limitée.

On propose une recherche tabou où tout mouvement sur un train est interdit pendant la durée du tabou. La liste tabou de taille l_{max} contient donc les indice des trains non modifiables dans la solution. Les transformations élémentaires utilisées sont les échanges de noeuds uniquement afin d'assurer la faisabilité de la solution à tout instant.

Dans ce cas d'étude, la taille du voisinage est trop importante pour qu'il soit entièrement étudié à chaque itération, surtout que pour bon nombre de trains la solution initiale est satisfaisante et ne pose pas de problèmes de robustesse. Dans la méthode tabou proposée ici, la liste des échanges possibles est construite progressivement en intégrant peu à peu les trains en fonction de leur score courant. Le voisinage à explorer est initialisé avec uniquement les sommets admissibles liés aux l_{max} trains ayant le score individuel le plus important, puis toutes les Q itérations, le train avec le pire score de la solution courante est ajouté. Il s'agit d'une stratégie de filtrage de l'ensemble des voisins de la solution courante pour se concentrer sur les parties avec la plus grande marge d'amélioration.

L'algorithme présenté ici utilise une liste L_{vois} d'échanges possibles $v = (n_1, n_2, s_1, s_2)$, avec n_1 un noeud sélectionné dans la solution courante C, n_2 un noeud alternatif correspondant au même train, et $s_1 = score\_noeud(G, C, n_1)$ et $s_2 = score\_noeud(G, C, n_2)$. Chaque échange de la liste L_{vois} est réalisable.

Plusieurs fonctions dédiées sont utilisées ici :

- $ajout\_voisinage(G, t, C, L_{vois})$ qui ajoute au graphe G les sommets du trains t qui peuvent être ajoutés à C puis met à jour la liste L_{vois} des échanges en incluant les échanges de noeuds de t.
- $pire\_train(G, C, L)$ qui renvoie l'indice du train de la liste L dont le score est maximal dans C.
- fonctions $update(L_{tabou})$ qui supprime le premier élément de L_{tabou}, fonction $ajout\_tabou(L_{tabou}, t)$ qui ajoute t à la fin de la file L_{tabou}.
- $update\_voisinage(G, n_1, n_2, C, L_{vois})$ qui met à jour L_{vois} une fois qu'un échange est accepté. Plusieurs actions sont faites :
  - l'élément $v_{1,2} = (n_1, n_2, s_1, s_2)$ est remplacé par $v_{2,1} = (n_2, n_1, s_2, s_1)$
  - les éléments $v_{1,3} = (n_1, n_3, s_1, s_3)$ sont remplacés par $v_{2,3} = (n_2, n_3, s_2, s_3)$
  - pour un élément $v_{i,j} = (n_i, n_j, s_i, s_j)$ correspondant à un autre train que n_1 et n_2 :
    - si il n'y a pas d'arêtes entre n_2 et n_i ou n_j, on supprime $v_{i,j}$ de L_{vois}
    - si l'arête existe, on met à jour les scores : $s_i \leftarrow s_i - w_{1,i} + - w_{2,i}$ et $s_i \leftarrow s_i - w_{1,j} + - w_{2,j}$
  - Si il y a un changement de quai entre n_1 et n_2, on ajoute pour chaque train t présent dans L_{vois} les sommets utilisant le quai libéré et on ajoute les transformations correspondantes dans L_{vois}.

Algorithme 4 Algorithme de recherche tabou

Données d'entrée : Solution initiale C, graphe courant G, Q le nombre d'itérations avant l'ajout d'un nouveau train à L_{vois}, taille de la liste tabou l_max
$n\_iter \leftarrow 0$
$L_{tabou} \leftarrow [-1, ..., -1]$ de taille l_max
$L_{vois} \leftarrow []$
$L_{non\_parcouru} \leftarrow L_T$
pour i in $1..Q$ **faire**
    $t \leftarrow pire\_train(G, C, L_{non\_parcouru})$
    $L_{vois} \leftarrow ajout\_voisinage(G, t, C, L_{vois})$
    $L_{non\_parcouru} \leftarrow L_{non\_parcouru} \backslash t$
fin pour
tant que $n\_iter \leq iter\_max$ **faire**
    pour q in $1..Q$ **faire**
        $L_{tabou} \leftarrow update(L_{tabou})$
        $gain\_max \leftarrow$ int_max
        pour $v = (n_1, n_2, s_1, s_2) \in V$ **faire**
            si $s_2 - s_1 < gain\_max$ **et** $train(n_1) \notin L_{tabou}$ **alors**
                $swap \leftarrow (n_1, n_2)$
                $gain\_max \leftarrow s_2 - s_1$
            fin si
        fin pour
        $C \leftarrow C \cup swap[2] \backslash swap[1]$
        $V \leftarrow update\_voisinage(G, swap[1], swap[2], C, L_{vois}V)$
    fin pour
    $n\_iter \leftarrow n\_iter + 1$
    $t \leftarrow pire\_train(G, C, L_{non\_parcouru})$
    $L_{vois} \leftarrow ajout\_voisinage(G, t, C, L_{vois})$
    $L_{non\_parcouru} \leftarrow L_{non\_parcouru} \backslash t$
fin tant que
Sortie : graphe G et solution C

5.3.4 Méthodes par voisinages variables

La recherche par voisinages variables de base (*Basic Variable Neighborhood search* ou BVNS) est une métaheuristique consistant à perturber aléatoirement et réparer une solution en changeant de structure de voisinage régulièrement pour améliorer la solution [91]. Cette variante de la recherche par voisinage variable est particulièrement adaptée aux instances de grandes tailles où l'exploration du voisinage est coûteuse. Elle suit les étapes suivantes :

1. On détermine des structures de voisinages $\mathcal{V}_1, ..., \mathcal{V}_{k_{max}}$ et on initialise la solution courante $x = x_0$
2. Perturbation : on tire aléatoirement une solution x' dans $\mathcal{V}_k(x)$
3. Réparation : on applique un algorithme glouton à x'
4. Si on a amélioration, on conserve x' et on retourne à la structure de voisinage $\mathcal{V}_1$
5. Si on n'a pas d'amélioration : si $k = k_{max}$ alors $k = 0$, sinon $k = k + 1$

Une méthode de voisinage variable a été appliquée par Hansen et al [92] pour la résolution d'un problème de clique maximum. Les voisinages les plus classiques sont les solutions à distance k où k

est la taille de la différence symétrique entre la solution et les solutions du voisinage. Pour $k = 1$, le voisinage correspond à l'ajout ou à la suppression d'un sommet de la clique.

Ici, on considère trois types de voisinages :

— Échanges simples sur noeuds prioritaires : on génère aléatoirement une liste de trains en privilégiant ceux avec une contribution forte à la fonction objectif. Les trains sont étudiés un à un, on choisit un quai disponible au hasard, un itinéraire d'arrivée et un de départ (itinéraires aléatoires s'il n'y a pas de changement de quai, itinéraires optimaux en cas de nouveau quai), le sommet correspondant est échangé dans la clique avec le sommet du train courant. La liste est générée avec la fonction $liste\_prior(G, C)$ qui crée une liste de taille variable tirée aléatoirement entre 10 et 50. Pour chaque train de la journée, on choisit aléatoirement un nombre entre 1 et 8, si le score individuel du train est supérieur à ce seuil, il est sélectionné pour la liste. Si la liste obtenue est trop longue, des trains sont supprimés, sinon un tirage avec remise est effectué au sein de la liste pour atteindre la taille souhaitée.

— Échanges simples sur tous les trains : on génère une liste aléatoire avec la fonction $liste\_aléatoire()$ sans considération du score, puis on procède comme pour le premier voisinage. La taille de la liste est également décidée aléatoirement entre 15 et 50 et les trains sont tirés avec remise.

— Échanges doubles : afin de casser la structure de la solution on propose d'échanger les quais pour une paire de trains compatibles stationnant en même temps dans la gare. Les paires sont également choisies aléatoirement, et seuls les échanges neutres ou améliorant la solution sont conservés.

On utilise la fonction suivante pour perturber la solution courante :

Algorithme 5 Perturbation

procédure PERTURBER(G, C, k)
  si $k \leq 2$ **alors**
    si $k = 1$ **alors**
      $L \leftarrow liste\_prior(G, C)$
    fin si
    si $k = 2$ **alors**
      $L \leftarrow liste\_aléatoire()$
    fin si
    pour $t \in L$ **faire**
      $n \leftarrow noeud\_courant(t, C)$
      $q \leftarrow$ quai(n)
      q' choisi aléatoirement dans $quais\_admissibles(t)$
      si $q' \neq q$ **alors**
        $i_a, i_d \leftarrow iti\_opt(G, C, t', q)$
      sinon
        $i_a, i_d \leftarrow iti\_rand(G, C, t', q)$
      fin si
      $G \leftarrow ajout\_noeud(G, t, i_a, i_d)$
      $C \leftarrow C \cup \{n_{t,i_a,i_d}\} \backslash \{n\}$

    fin pour
  fin si
  si $k = 3$ **alors**
    pour $t_1 \in L_T$ **faire**
      $L_{candidats} = liste\_candidats(t_1, C)$
      si $L_{candidats} \neq \emptyset$ **alors**
        $t_2, q_2 \leftarrow rand(L_{candidats})$
        $n_1 \leftarrow noeud\_courant(t_1, C)$
        $n_2 \leftarrow noeud\_courant(t_2, C)$
        $i_{1,a}, i_{1,d} \leftarrow iti\_opt(G, C, t_1, q_2)$
        $G \leftarrow ajout\_noeud(G, t_1, i_{1,a}, i_{1,d})$
        $C' \leftarrow C \cup \{n_{t_1,i_{1,a},i_{1,d}}\} \backslash \{n_1\}$
        $i_{2,a}, i_{2,d} \leftarrow iti\_opt(G, C', t_2, q_1)$
        $G \leftarrow ajout\_noeud(G, t_2, i_{2,a}, i_{2,d})$
        $C' \leftarrow C' \cup \{n_{t_2,i_{2,a},i_{2,d}}\} \backslash \{n_2\}$
        si $score(G, C') \leq score(G, C)$ **alors**
          $C \leftarrow C'$
        fin si
      fin si
    fin pour
  fin si
  renvoyer G, C
fin procédure

La réparation est faite en utilisant l'heuristique de recherche parmi les noeuds déjà explorés, c'est-à-

dire sans création de nouveaux noeuds. Sur le même principe que la recherche de plateau, des solutions de qualité équivalente à la solution courante sont choisies avec une probabilité égale à 1/4. On espère ainsi faire évoluer la solution régulièrement pour atteindre de nouvelles configurations.

Algorithme 6 Algorithme de recherche par voisinage variable

Données d'entrée : Solution initiale C, graphe courant G, temps de calcul max t_{max}, $rand(0,1)$ une fonction de tirage aléatoire d'un réel entre 0 et 1
$n\_iter \leftarrow 0$
$k \leftarrow 1$
tant que $time \leq t_{max}$ **faire**
    $score\_courant \leftarrow score(G, C)$
    $G, C' \leftarrow$ PERTURBER(G, C, k)
    $G, C' \leftarrow$ OPTIMISATION(G, C')
    si $score(G, C') < score\_courant$ **ou** $(score(G, C') = score\_courant$ **et** $rand(0,1) \leq 0.25)$ **alors**
        $score\_courant \leftarrow score(G, C')$
        $C \leftarrow C'$
        $k \leftarrow 0$
    sinon si $k = 3$ **alors**
        $k \leftarrow 1$
    sinon
        $k \leftarrow k + 1$
    fin si
fin tant que

Sortie : graphe G et solution C

5.4 Expérimentations

5.4.1 Introduction

On applique dans cette section les algorithmes présentés ci-dessus. Pour une date d donnée pour laquelle les affectations de voies sont optimisées, la méthodologie suivante est appliquée :

1. Les données du GOV sont chargées (liste L_T des trains, solution initiale, description de l'infrastructure de la gare, description des contraintes à prendre en compte). Ces fichiers d'entrée sont générés par l'outil OpenGOV.
2. Les probabilités de retards des T trains prévus sur la journée d sont estimées à l'aide d'un modèle entraîné sur l'année précédente (mois de d exclu) en utilisant l'approche par forêts aléatoires.
3. Le graphe initial G_0 est construit à partir du fichier GROIX
4. Les algorithmes suivants sont appliqués parallèlement sur le graphe initial :
    — algorithme glouton seul en étudiant les trains par ordre de contribution décroissante
    — algorithme glouton, recherche tabou, réduction de distance. On fixe la taille de la liste tabou à 15, et Q le nombre d'itérations avant l'ajout d'un nouveau train dans la liste d'échanges de noeuds possible est fixé à 5
    — algorithme glouton, VNS, réduction de distance, moyenné sur 10 itérations

5. Les algorithmes sont comparés avec la valeur de la fonction objectif, et leurs performances sont également évaluées en appliquant les retards réellement observés aux solutions. Des résultats complémentaires sur la taille du graphe sont donnés en annexe.

5.4.2 Instances

Les distributions de probabilité estimées sur les ensembles de test couvrent la période de juillet 2018 à mars 2019. Étant donné la grande redondance des planifications au sein d'une même période (saison) et selon les jours de la semaine, toutes les dates ne seront pas testées, mais on propose un échantillonage visant à avoir plusieurs types de jours pris à différentes périodes (vacances, été, hiver, ...). Les dates sélectionnées, ainsi que la description de l'instance, le nombre de trains et le nombre de mouvements commerciaux sont données dans le tableau 5.3.

id	Jour	Mois	Vacances	Trains	Mvts comm.
0307	Mardi	Juillet	Oui	388	556
1407	Samedi	Juillet	Oui	264	398
2507	Mercredi	Juillet	Oui	336	486
2208	Mercredi	Août	Oui	342	497
2608	Dimanche	Août	Oui	259	390
0409	Mardi	Septembre	Non	387	563
1409	Vendredi	Septembre	Non	391	575
2209	Samedi	Septembre	Non	241	345
0410	Jeudi	Octobre	Non	381	557
0810	Lundi	Octobre	Non	401	572
1810	Jeudi	Octobre	Non	385	567
2910	Lundi	Octobre	Oui	393	566
0911	Vendredi	Novembre	Non	395	583
1611	Vendredi	Novembre	Non	398	588
2511	Dimanche	Novembre	Non	247	387
0512	Mercredi	Décembre	Non	374	547
0912	Dimanche	Décembre	Non	247	391
1812	Mardi	Décembre	Non	380	559
2712	Jeudi	Décembre	Oui	341	507
0501	Samedi	Janvier	Oui	245	384
1501	Mardi	Janvier	Non	379	556
2401	Jeudi	Janvier	Non	381	567
0602	Mercredi	Février	Non	385	567
1102	Lundi	Février	Oui	385	566
2402	Dimanche	Février	Oui	249	383
0802	Vendredi	Mars	Non	389	585
1803	Lundi	Mars	Non	383	566

Tableau 5.3 – Description des instances utilisées

5.4.3 Scores

Les expérimentations sont menées en Python, et le graphe est construit avec la librairie *NetworkX* [88] qui le génère comme un dictionnaire de noeuds et d'arêtes.

Les scores obtenus en appliquant les algorithmes sur les différentes instances sont donnés dans le tableau 5.4, avec le temps de calcul en secondes de l'algorithme glouton. Les algorithmes Tabou et

VNS sont arrêtés après 10 minutes de calcul et les scores VNS sont moyennés sur 10 itérations.

id	Initial	Glouton	Temps	Tabou	VNS
0307	690.9	360.9	147	307.1	**286.6**
1407	150.8	73.2	35	**42.2**	43.4
2507	274.0	138.8	84	90.9	**89.9**
2208	303.5	172.8	92	122.6	**117.7**
2608	175.7	146.9	28	130.0	**109.5**
0409	770.4	369.4	148	**258.8**	257.5
1409	811.2	471.1	146	396.0	**387.6**
2209	174.1	86.1	37	**42.2**	42.7
0410	724.6	375.3	145	264.7	**260.1**
0810	728.4	394.7	142	299.7	**279.7**
1810	760.5	400.9	146	295.5	**269.0**
2910	738.9	397.3	136	323.4	**302.0**
0911	811.4	501.5	145	402.5	**390.8**
1611	692.6	386.8	148	322.3	**316.4**
2511	146.8	85.8	35	60.5	**50.9**
0512	751.2	450.9	138	313.8	**304.9**
0912	158.4	118.6	36	**73.9**	83.3
1812	798.2	490.7	135	323.7	**302.1**
2712	406.8	295.7	96	170.6	**136.5**
0501	150.3	82.8	38	**27.2**	29.7
1501	691.8	384.9	130	255.1	**251.5**
2401	779.5	417.1	139	291.5	**283.5**
0602	670.7	341.8	142	248.1	**242.6**
1102	722.7	369.8	133	230.2	**226.4**
2402	112.3	52.8	33	**40.3**	42.3
0803	659.2	376.9	128	284.4	**258.3**
1803	688.7	341.0	140	236.1	**229.7**

TABLEAU 5.4 – Scores obtenus

On constate une brutale amélioration de la fonction objectif grâce à l'algorithme glouton, puis une baisse moins importante en appliquant les deux algorithmes VNS et tabou. Les scores des deux métaheuristiques sont très proches sur toutes les instances, avec de légèrement meilleures performances pour la méthode par voisinages variables.

Le tableau E.1 donné en annexe E donne la taille des graphes obtenus à la suite de ces expérimentations. Les graphes VNS sont toujours de taille inférieure à ceux par recherche tabou pour des performances similaires. Plus de solutions sont visitées et accessibles par l'algorithme de liste tabou, cependant la recherche est plus efficace pour le VNS. On peut trouver plusieurs raisons à cela, la première étant que dans le cadre de l'algorithme tabou, tous les noeuds alternatifs admissibles pour un train sont générés alors que potentiellement aucun ne présente d'intérêt. C'est par exemple le cas si le poids est du à un conflit à l'arrivée, les noeuds avec des itinéraires alternatifs au départ ont une valeur ajoutée moindre. Ensuite, la recherche tabou n'étudie pas tous les trains puisqu'ils sont ajoutés au fur et à mesure dans l'ordre de leur contribution dans la solution courante. Certains trains ayant une contribution nulle ou très faible ne sont donc jamais modifiés, ce qui n'est pas le cas dans

l'approche VNS où ils peuvent être modifiés aux $2^{\text{ème}}$ et $3^{\text{ème}}$ structures de voisinage. Même si les modifier n'améliore pas directement la solution, un ajustement peut libérer une voie ou un itinéraire stratégique. Enfin, les échanges de voies ne sont pas possibles dans la méthode tabou. En étudiant les changements opérés, on a observé qu'ils concernent souvent des trains à contribution faible ou nulle, et que les échanges n'influencent pas l'objectif immédiatement, cependant ils sont souvent suivi d'une amélioration après dans la phase de recherche locale ou après un nouveau passage dans la première structure de voisinage.

A titre informatif, deux graphes ont été générés entièrement, c'est-à-dire avec pour chaque train l'ensemble des associations quais/itinéraires admissibles. Pour l'instance du 25/11 qui compte 247 trains, ce graphe final a 10 173 sommets et est construit en environ 7h30, et le graphe de l'instance du 04/10 avec 381 trains contient à la fin 18 643 sommets pour 34h de construction. D'après le tableau E.1 donné en annexe, les graphes générés par les heuristiques ne contiennent que 800 à 1200 sommets. Le temps nécessaire à l'ajout d'un nouveau sommet et au calcul des poids de ses arêtes croît très fortement avec la taille du graphe, les ajouter dans un ordre stratégique permet d'améliorer rapidement la solution sans perdre de temps à générer les données.

5.4.4 Retards réels

On dispose ici des retards réels rencontrés par les trains pour les journées optimisées. Les performances en conditions réelles des adaptations proposées peuvent être étudiées. Les graphiques ci-dessous mesurent le nombre de conflits et leur amplitude cumulée en minutes, soit le nombre de minutes manquantes pour faire les enchaînements de mouvements prévus en respectant les contraintes. Ces conflits sont séparés en deux types : les conflits qui impliquent au moins un mouvement technique et les conflits entre deux circulations commerciales. Pour plus de visibilité les départs ne sont pas distingués des arrivées, bien que le modèle optimise plus leur placement. On ajoute par ailleurs les solutions optimisées par l'outil OpenGOV pour comparaison.

Ces résultats sont cependant très partiels, ils ne prennent pas en compte d'autres aléas opérationnels comme les suppressions de trains ou les retards techniques, et n'intègrent que les retards inférieurs aux seuils de troncature. Par ailleurs, un même retard important, par exemple d'un TGV avec 20 minutes, va générer plusieurs conflits et de nombreuses minutes de retard avec cette métrique alors que dans la réalité le train retardé serait replanifié sur un autre itinéraire avec un impact moindre. De plus, un nombre réduit d'instance est testé.

Les résultats sont présentés en figure 5.10 et 5.11, avec en surbrillance les valeurs pour les conflits commerciaux et en plus clair les conflits avec des mouvements techniques. Les croix sur le graphique d'amplitude 5.10 représentent la somme de minutes de retards créées en raison de conflits de faisabilité.

Le tableau 5.5 donne le pourcentage moyen d'amélioration de la solution initiale en volume de retards généré. La première ligne prend en compte tous les conflits et la seconde ne compte que les conflits commerciaux. Les conflits impliquant des trains techniques ne sont pas très représentatifs dans la mesure où leur retard est inconnu.

	Glouton	Tabou	VNS	OpenGOV
Tous conflits	34 %	41 %	40 %	36 %
Juste conflits commerciaux	40 %	52 %	54 %	28 %

Tableau 5.5 – Amélioration moyenne en amplitude de retards

On constate une baisse systématique du nombre de conflits et de leur amplitude entre la solution d'origine et les solutions optimisées. Cette forte amélioration est principalement due à une efficace diminution des conflits de faisabilité grâce à l'algorithme glouton. Concernant les solutions optimisées,

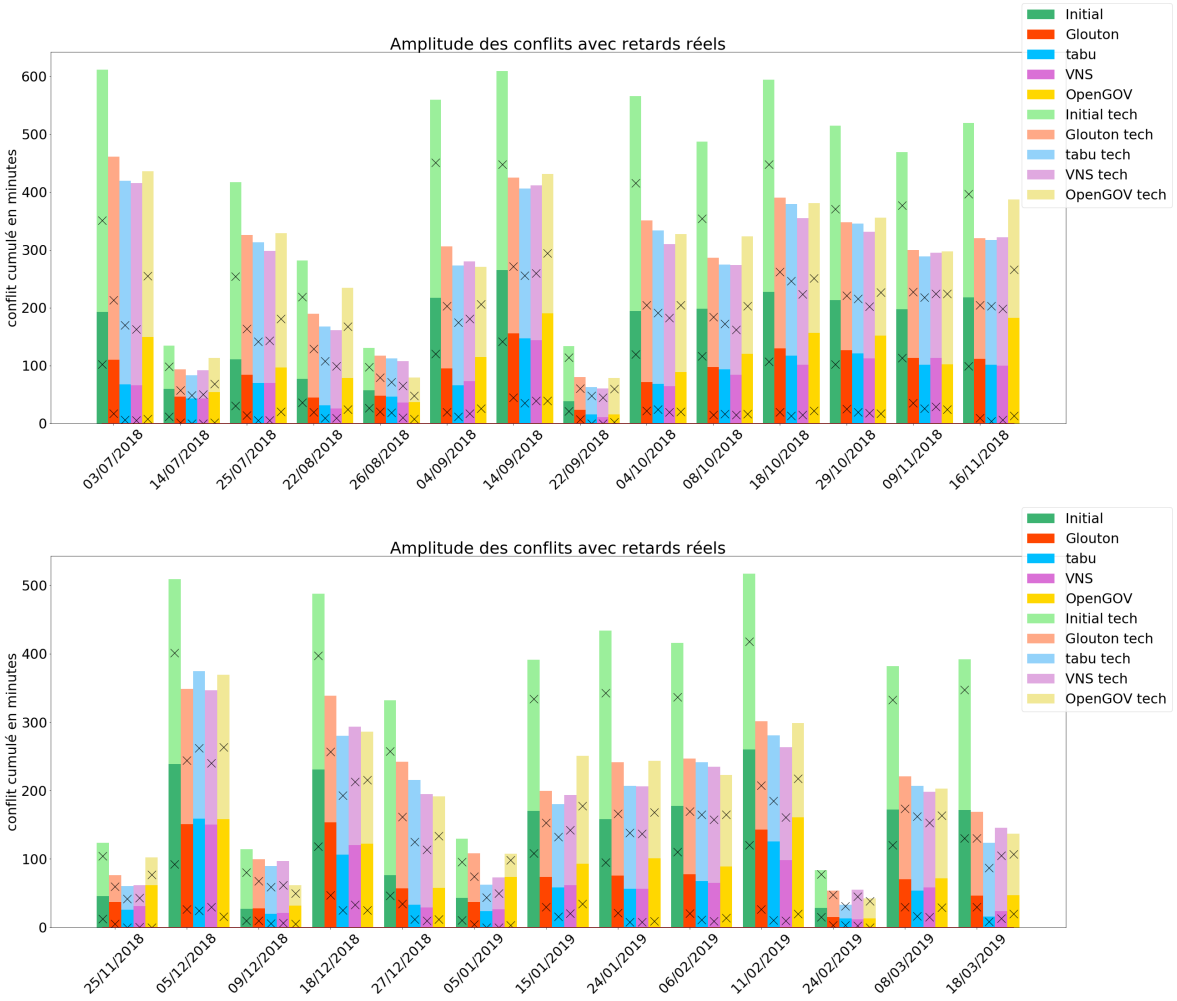

FIGURE 5.10 – Amplitude des conflits

il y a également une amélioration entre l'algorithme glouton et les métaheuristiques, cependant elle n'est plus systématique, avec par exemple la recherche tabou pour l'instance du 05/12/2018 où l'amplitude cumulée de conflits est légèrement supérieure à celle de l'algorithme glouton. Comme le laissait présager les scores, les méthodes tabou et VNS donnent des performances relativement équivalentes.

Les solutions optimisées par l'outil OpenGOV sont en général légèrement moins bonnes que les solutions optimisées par nos algorithmes sur la plupart des instances. En moyenne les écarts de performance sont plus importants car il y a quelques instances où les solutions d'OpenGOV créent beaucoup plus de conflits que toutes les autres solutions optimisées, comme par exemple pour les dates du 08/10/18, 16/11/18 ou 05/01/18. Plus d'attention est portée sur la faisabilité et la robustesse des mouvements techniques dans OpenGOV que dans nos algorithmes, ce qui peut justifier en partie les disparités dans les résultats.

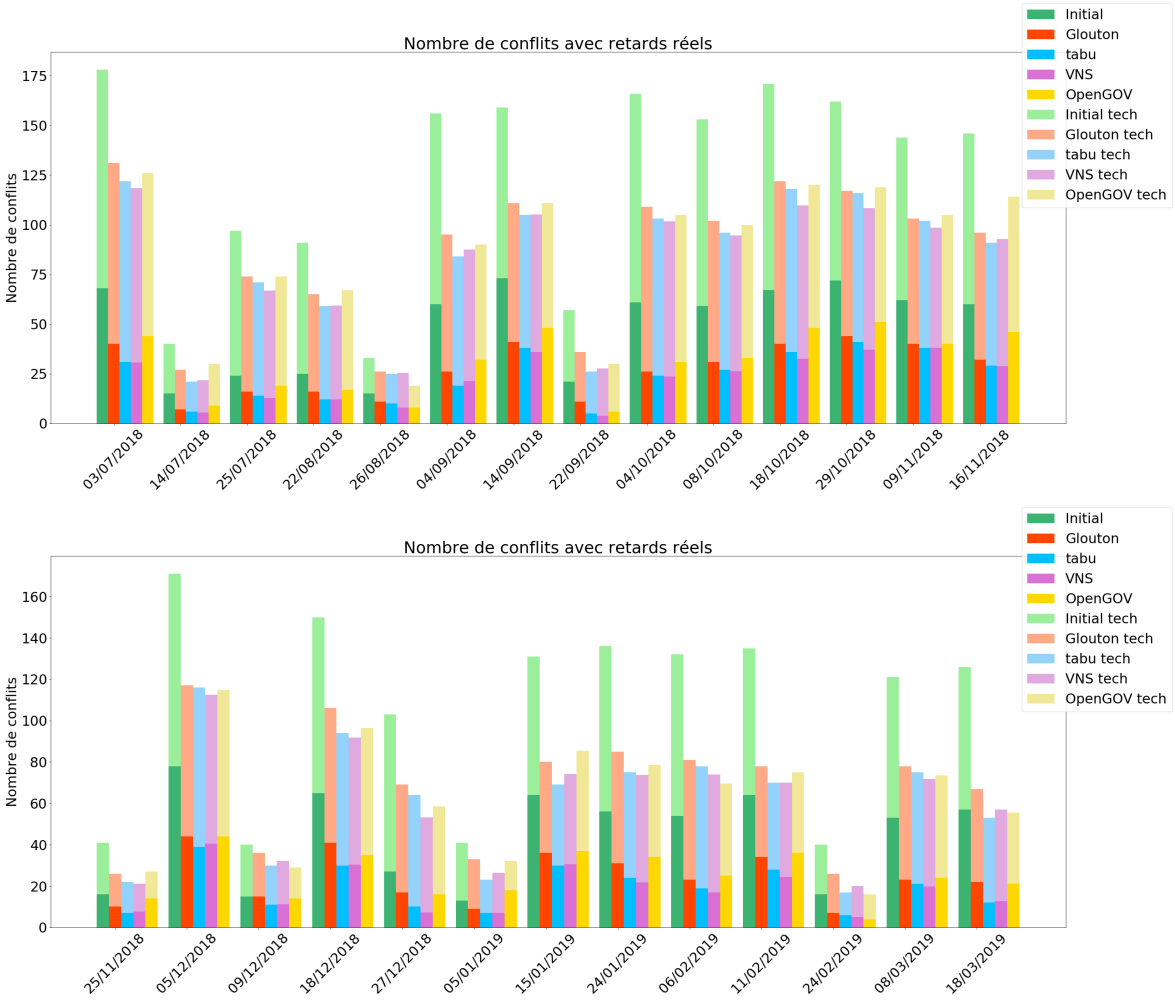

FIGURE 5.11 – Compteur de conflits

5.5 Discussion

Ce chapitre a présenté une première approche d'intégration d'analyses des données passées pour appuyer la prise de décision en gare. Cette section résume les points d'amélioration de la méthodologie et les limites identifiées.

5.5.1 Perspectives d'amélioration

5.5.1.1 Intégration de nouvelles contraintes

L'approche de modélisation et de résolution peut être revue pour permettre de mieux intégrer les contraintes opérationnelles et permettre une recherche plus efficace :

— Voies à quai et itinéraires fictifs : l'introduction de sommets fictifs permettrait d'opérer des échanges de voie à quai entre des trains en cas de conflit. On peut imaginer retirer un des trains

de la solution afin d'optimiser le reste des circulations puis le replacer ensuite. Cette méthode requiert plus d'attention quant à la faisabilité de la solution produite puisqu'on doit être en mesure de replacer le train.

— Trains en coupe/accroche : les algorithmes présentés ici ne permettent pas la modification des affectations de quais ou d'itinéraires pour les mouvements liés, à l'exception des itinéraires pour le mouvement simple avant l'accroche ou après la coupe. Ajouter cette flexibilité permettrait d'explorer d'autres solutions.

— Modification des horaires techniques : comme cela a été mentionné, les mouvements techniques sont planifiés mais leurs horaires sont peu respectés. En cas de conflit la priorité est donnée aux circulations commerciales. Les solutions produites pourraient être améliorées en proposant des adaptations locales des horaires des mouvements techniques afin de générer moins de conflit. On peut par exemple construire une heuristique qui étudierait les conflits impliquant un mouvement technique, et chercherait à décaler le mouvement dans une fenêtre de temps donnée pour en limiter l'impact, ou encore générer de nouveaux sommets avec la même infrastructure mais des horaires différents. Ces modifications d'horaires pourraient aussi être appliquées aléatoirement dans le cadre d'une recherche à voisinage variable.

5.5.1.2 Optimisation du temps de calcul

Le temps de calcul est une contrainte industrielle importante. Ici, les phases de recherches sont bloquées à 10 minutes, auxquelles il faut rajouter la durée de l'algorithme glouton appliqué au préalable ainsi que l'algorithme de réduction de distance, ce qui peut aller jusqu'à 13 minutes de calcul environ. Ces temps sont acceptables, mais pour permettre un usage industriel, il est préférable que l'optimisation ne dépasse pas les quelques minutes pour permettre par exemple de tester différents cas de figure (changement d'horaire, quai fixé, ...). Plusieurs pistes pour améliorer les performances ont été identifiées :

— Revoir l'encodage de la solution : dans ces travaux, seule la structure de graphe de la librairie Python *NetworkX* a été essayée, avec en parallèle une base de données qui contenait les informations liées à chaque noeud (identifiant, indice du train, itinéraires, horaires,...). D'autres encodages devraient être essayés afin d'optimiser au mieux le parcours du graphe.

— Ne pas générer certaines arêtes connues : de nombreux trains ne rentrent jamais en conflit, par exemple s'ils circulent sur des plages horaires distinctes. Les sommets de ces trains sont donc toujours connectés avec des arêtes de poids nul. On peut donc garder en mémoire ces paires de train sans les générer dans le graphe. Cette stratégie a été essayée et a permis d'améliorer les temps de calcul d'environ 15 à 20% lors des tests, cependant elle n'a pas été explorée plus dans ces travaux car en contrepartie on perd la structure de clique qui permet de contrôler facilement la faisabilité des solutions. L'autre option est de travailler sur deux graphes, un qui est un graphe d'incompatibilité égal au complémentaire du graphe utilisé ici sans pondération, et un graphe formé des mêmes sommets et ne contenant que les arêtes de poids non nul de G. On chercherait alors un stable de taille $|T|$ du premier graphe de poids minimal dans le second.

— Travailler sur des sous-graphes : il est possible de se concentrer sur des parties de la planification, par exemple sur une plage horaire donnée ou une zone de la gare en considérant les autres planifications fixées. Dans ce cas on peut réduire le graphe aux seules circulations potentiellement impactée par les modifications apportées, et éventuellement paralléliser des recherches disjointes, puis fusionner après les graphes générés. Cette approche peut être sous-optimale mais est intéressante en début d'optimisation pour rapidement résoudre des situations non robustes et identifier de bons sommets alternatifs.

— Améliorer le prétraitement des données pour réduire les listes de voies et itinéraires envisageables, par exemple en excluant de la recherche certains éléments d'infrastructure incompatibles avec des affectations non modifiables (trains en coupe/accroche, trains déjà présents sur site)

— Ajuster les valeurs de troncature : les valeurs maximales des retards considérés peuvent être modifiées, en particulier pour les TGV à l'arrivée pour lesquels elle est de 20 minutes. Les retards de 20 minutes font déjà partie des perturbations les plus importantes qui ne peuvent pas être absorbées par un GOV et pourrait être revue à la baisse. L'argument contre cette réduction est d'ordre statistique car un retard de 20 minutes pour un TGV n'est pas une valeur extrême (environ 6% des trains ont un retard supérieur). Concernant l'optimisation de la robustesse, les valeurs de troncature ont une influence sur la taille du graphe d'incertitude $\tilde{G}$ puisque plus les domaines des distributions de probabilité sont grands plus il y a de conflits potentiels, et donc d'arêtes avec un poids non nul.

5.5.2 Limites

5.5.2.1 Approximation de la robustesse

La définition de la robustesse impose encore des difficultés. Durant les opérations, deux aspects sont importants dans la solution : elle doit être robuste aux petits aléas, en particulier en plaçant les trains de façon à avoir des espacements pour absorber les petites perturbations, et elle doit permettre une réorganisation facile, par exemple grâce à des itinéraires ou voies libres pour accueillir des circulations en cas de conflit. Ces deux aspects sont contradictoires puisque avoir des marges entre les trains consomme plus de capacité, et au contraire, pour avoir des voies disponibles il faut densifier l'utilisation du reste de l'infrastructure et donc réduire les espacements. Dans ces travaux on se concentre sur le premier aspect de la robustesse.

La fonction objectif a été construite afin de limiter les conflits de faisabilité en pénalisant leur amplitude cumulée, et afin de réduire les occurrences de conflits liés aux retards en pénalisant leur probabilité. Il n'y a cependant pas de fonction objectif à suivre qui soit clairement définie, et de futurs travaux devraient se concentrer sur cela. En particulier, il faudrait évaluer l'impact des différentes politiques d'optimisation et de construction de la fonction objectif sur les performances réelles.

En reprenant la distinction entre robustesse individuelle et collective [110], l'optimisation actuelle se fait individuellement, c'est-à-dire qu'on cherche à réduire l'impact individuel de chaque train sur ce qui l'entoure. L'idéal serait de travailler dans un cadre de robustesse collective où les interactions entre l'ensemble des circulations et le déroulé du trafic est optimisé.

5.5.2.2 Mesure de l'incertitude

Trains techniques : l'incertitude sur les horaires des mouvements techniques est en partie stochastique, car ils peuvent être très aléatoires, par exemple en cas de retard de maintenance. Cependant une partie des retards techniques est liée à des prises de décision. Ces trains sont souvent retardés ou avancés afin de faciliter le déroulement des opérations. Pour ces raison leur incertitude est difficilement quantifiable. Cela crée un déséquilibre dans ce cas d'étude puisque l'information sur les aléas est partielle. On peut pallier cette difficulté avec une pénalité forfaitaire pour les trains techniques mais les conséquences sont difficilement estimables. Ici seuls les conflits de faisabilité avec une circulation technique sont pénalisés, l'ajout de pénalités pour les interactions non robustes augmenterait significativement le nombre d'arêtes non nulles car de telles interactions sont nombreuses.

En pratique les mouvements commerciaux seront presque systématiquement prioritaires face aux mouvements techniques en cas de conflit de courte durée, comme ceux dont on cherche à se protéger.

Cependant certaines circulations techniques ont une plus grande priorité, par exemple dans le cas où la rame est attendue pour un départ dans un temps court.

Biais des trains au départ : les distribution de retards des trains au départ sont évaluées en utilisant les historiques de retards à quai. Cependant certains de ces retards peuvent être dus à des conflits dans le GOV pour lesquels le train est maintenu à quai, bien qu'il ait été en mesure de partir à l'heure. Cela signifie qu'il y a une légère surestimation du risque de retard au départ. Les trains à l'arrivée sont beaucoup moins concernés étant donné que les valeurs sont relevées sur la fin des voies en ligne, avant la zone de routage.

Décision et incertitude : une limite connue de ce type de problème est que la prise de décision et l'incertitude ne sont pas indépendantes [20]. Dans le problème présenté par exemple, les probabilités de retards au départ sont calculées en utilisant les données historiques basées sur les performances des anciennes planifications, mais une modification des habitudes de routage en gare peut affecter ces risques. Ce serait le cas si on utilisait une mesure d'incertitude pour les trains techniques alors que leur retard est fortement lié aux décisions prises en gare. C'est également vrai à plus grande échelle, si un tel système est utilisé dans plusieurs gares, les données d'apprentissage des trains à l'arrivée en bout de ligne ne seront plus forcément valides si les motifs de propagation changent dans les gares intermédiaires.

5.5.2.3 Cas d'étude

La gare de Paris-Montparnasse est une grande gare terminale, ce qui rend les problèmes d'affectation de voies difficiles et la robustesse plus complexe à contrôler. Le fait que la gare soit terminale influence l'approche à deux niveaux. Tout d'abord elle compte de nombreuses circulations techniques en raison des mouvements depuis ou vers les technicentres : les trains ayant fini leur maintenance sont utilisés pour un nouveau trajet à l'origine de la gare Montparnasse, ou à l'inverse après leur terminus de nombreux trains vont effectuer leur maintenance. Le second aspect concerne l'indépendance des distributions de retards : étant donné que les trains restent à quai plus longtemps, les retards à l'arrivée et au départ peuvent être considérés comme indépendants car absorbé par l'occupation de voie. Le premier aspect pose des difficultés déjà mentionnées, le second facilite la modélisation car les effets de propagation lors de l'occupation des voies peuvent être ignorés.

L'exploitation de la gare ne permet pas de tirer parti au mieux de l'estimation de l'incertitude. En effet, les différents types de circulation sont affectés à des zones distinctes de la gare et se croisent peu. Or, comme on l'a vu dans le chapitre précédent, c'est dans l'hétérogénéité des circulations que les risques de conflits varient le plus car certains motifs ne sont pas partagés de la même manière selon le type de train. Par exemple les heures de pointes ou les WE n'ont pas la même influence sur la distribution des TGV et des TN, mais dans le cas d'un GOV ces deux motifs seront partagés par les trains en conflits. Par exemple plusieurs TGV dans le même sens de circulation partageront plusieurs variables explicatives en commun, comme les variables temporelles et la densité en gare, et auront donc potentiellement un risque proche. Cela ne permet donc pas d'arbitrer aussi efficacement entre les affectations possibles que si les risques étaient variés.

5.5.3 Approches alternatives orientées données

On a fait le choix dans ces travaux de proposer un modèle séquentiel avec une première partie qui quantifie l'incertitude autour des données et une seconde qui optimise la prise de décision en fonction de cette incertitude par recherche locale. D'autres approches pourraient être menées.

5.5.3.1 Simulation stochastique

Une première idée serait d'utiliser les distributions estimées pour générer des retards dans un modèle de simulation stochastique. En effet, la simulation permet d'évaluer plus précisément les performances réelles d'une planification et d'intégrer des règles opérationnelles de gestion de conflit (propagation, changement de quai, etc.). Alors que dans les modèles présentés ici seuls les conflit entre paires de circulations sont évalués, une méthodologie de simulation stochastique permet de représenter plus précisément les interactions indirectes entre les trains, comme par exemple les chaînes de propagation.

On peut donc plus facilement évaluer la robustesse d'une solution en générant des retards conformes aux motifs réels, et utiliser les résultats moyennés sur un grand nombre de simulations pour proposer des ajustements dans la solution. La question de la calibration des perturbations en simulation est un sujet souvent abordé, mais dans le cadre du ferroviaire très peu d'études ont exploités les données réelles pour calibrer les retards simulés, et de manière générale l'estimation de distributions de probabilité utilise très rarement ou très peu des variables explicatives pour affiner les profils de retards (cf 2.3.2.1 et 2.4.1.1). Les probabilités estimées ici permettent de pallier ce manque.

Une telle méthodologie de simulation stochastique a été implémentée en utilisant les données de ces travaux et les résultats peuvent être trouvés dans l'article [62]. Seule la méthodologie d'évaluation de la robustesse par simulation stochastique est présentée dans l'article, la piste d'optimisation des GOV à partir de ces résultats n'a pas été explorée plus.

5.5.3.2 Robustesse de récupération

Le cadre de travail de robustesse de récupération consiste à relâcher certaines contraintes à condition qu'on puisse les réparer facilement avec un coût borné [120]. Une adaptation au problème d'affectation de voies en gares avec l'utilisation de propagation comme récupération a été proposé par Caprara et al [42]. Le problème d'occupation des voies se décompose en deux sous-problèmes : le problème nominal, qui consiste à affecter des voies à quai et itinéraires aux trains, et un sous-problème de propagation des retards, qui minimise la borne supérieure des retards générés par la solution du problème nominal pour un ensemble prédéfini d'aléas sur des évènements (arrivée, occupation des voies, départ).

On utilise les notations suivantes :

T	Nombre de trains
$K_{t,t'}$	ensemble de cliques de sommets incompatibles entre t et t' deux trains distincts
S	ensemble de perturbations, un scénario $s \in S$ associe à chaque évènement e un retard $\delta_{e,s}$ initial pouvant être nul
D	variable de retard propagé (somme des retards primaires et secondaires de chaque train)
N	$N = (\{a_t, t \in T\} \cup \{q_t, t \in T\} \cup \{d_t, t \in T\}, A(N))$ réseau de propagation des retards. Les sommets représentent les évènements planifiés (arrivées, occupations de quai et départs), et les arêtes représentent les retards propagés
x_P	variables représentant les affectations *itinéraires/train*
$d_{e,s}$	variables modélisant le retard de l'évènement e dans le scénario s

Le modèle proposé est le suivant :

$$min \sum_{t \in T} \sum_{t(P)=t} c_P x_P + D$$

$$\sum_{t(P)=t} x_P = 1, \qquad \forall t \in T \tag{5.8a}$$

$$\sum_{P \in K} x_P \leq 1, \qquad \forall (t,t') \in T^2, K \in K_{t,t'} \tag{5.8b}$$

$$D \geq \sum_{t \in T} \left(d_{a_t,s} + d_{q_t,s} + d_{d_t,s} \right), \qquad \forall t \in T, \forall s \in S \tag{5.8c}$$

$$d_{a_t,s} \geq \delta_{a_t,s}, \qquad \forall t \in T, \forall s \in S \tag{5.8d}$$

$$d_{q_t,s} \geq \delta_{q_t,s} + d_{a_t,s}, \qquad \forall t \in T, \forall s \in S \tag{5.8e}$$

$$d_{d_t,s} \geq \delta_{d_t,s}, \qquad \forall t \in T, \forall s \in S \tag{5.8f}$$

$$d_{q_t,s} \geq \delta_{q_t,s} + d_{a_t,s}, \qquad \forall t \in T, \forall s \in S \tag{5.8g}$$

$$d_{e_t,s} \geq d_{e_{t'},s} - f(e_t, e_{t'}), \qquad \forall (e_t, e_{t'}) \in A(N), \forall s \in S \tag{5.8h}$$

$$f(e_t, e_{t'}) = \sum_{t(P)=t} \sum_{t(P')=t'} \alpha_{P,P'} x_P x_{P'}, \qquad \forall (e_t, e_{t'}) \in A(N) \tag{5.8i}$$

La première contrainte est une contrainte d'affectation, la seconde d'incompatibilité de cliques, les cinq suivantes sont des contraintes de propagation de retards (en suivant les différents cas dépendant de l'événement considéré, et la dernière contrainte lie les deux sous problèmes puisque la quantité de retards absorbée entre deux événements dépend des affectations choisies.

Les auteurs recommandent de construire S en budgétisant : $S = \{\delta : \sum_{\delta_e} \delta_e \leq \Delta\}$, cependant dans notre cas on dispose d'une estimation de l'incertitude sur les évènements. Il n'est pas possible de construire un ensemble S exhaustif, cela demanderait d'évaluer toutes les combinaisons possibles de retards ce qui ajoute un nombre de variables de propagation exponentiel en la taille des données. Une alternative serait de construire S aléatoirement de la même manière que pour générer les perturbations d'une simulation stochastique. Les probabilités estimées peuvent servir à générer K scenarios complets avec des retards aléatoires plausibles pour chaque train, possiblement nuls, sans besoin de budgetiser les aléas. On peut adapter la modélisation pour optimiser le retard propagé en moyenne plutôt que dans le pire des cas en introduisant des variables D_s qui remplacent D dans la troisième contrainte et dont on minimise la somme par exemple.

Cette modélisation peut cependant poser des difficultés computationnelles. Plus K est grand plus un grand nombre de scenarios sont pris en compte pour la robustesse cependant cela augmente fortement le nombre de variables et de contraintes.

5.6 Conclusion

La quantification et l'intégration de l'incertitude sont des points centraux en optimisation combinatoire La disponibilité de bases de données de plus en plus nombreuses et complètes combinée à la démocratisation de l'apprentissage statistique ouvrent le champs à de nouveaux travaux prometteurs dans la gestion de l'incertitude en aide à la décision en apportant une information complémentaire aux problèmes.

Ces travaux rentrent dans cette dynamique, et ce chapitre a pu exposer un exemple d'intégration des estimations de distribution de retards dans les processus d'adaptation des planifications en gare. Le cadre de résolution du problème nominal est déjà connu et est bien traité, que ce soit d'un point

de vue théorique par résolution de clique maximum, ou en pratique dans le cas des opérations en gare (cf 2.2).

L'intégration de la robustesse est cependant plus problématique. D'un point de vue ferroviaire, la caractérisation de ce que devrait être une solution robuste n'est pas claire. On propose ici de définir qualitativement la robustesse des planifications d'occupation de voies selon deux axes : elles doivent limiter la quantité de conflits en opérationnel et minimiser la masse de retards créés par propagation. Le premier est important car il assure un travail opérationnel réduit, sachant que les conflits sont inévitables mais peuvent être pris en charge au cas par cas. Le second axe évalue la capacité de retour à la normale sans action supplémentaire.

On ne peut cependant pas quantifier ces aspects directement. De manière générale, seules les conséquences locales des décisions d'un train sur l'autre peuvent être estimées. Les conséquences globales sur les effets de trafic sont bien plus complexes et irréalistes à mettre en place dans une méthodologie déjà contrainte par les temps de calcul. On a donc proposé de construire une fonction objectif limitant le nombre de conflits directs par faisabilité, et les conflits indirects en se basant sur le risque de retard.

Plusieurs heuristiques et métaheuristiques ont donc été mises en places pour permettre l'adaptation de la planification d'origine selon cet objectif. Les résultats montrent une amélioration moyenne importante de la quantité de conflits générés quand ces planifications adaptées sont confrontées aux retards réels des jours correspondants.

Chapitre 6

Conclusions

Ces travaux ont étudié le potentiel des archives de retards pour améliorer la robustesse des affectations de voies en gare. On a construit pour cela une méthodologie complète reposant sur deux étapes coordonnées : la première d'analyse de données visant à affiner la connaissance sur les paramètres incertains du problème par de l'apprentissage statistique, et la seconde d'aide à la décision pour résoudre le problème en intégrant cette incertitude.

L'application à un problème ferroviaire a imposé plusieurs contraintes industrielles qui ont été prises en compte dans la méthodologie, à savoir la qualité des données, l'hétérogénéité des instances, avec des trains de types différents, des objectifs d'optimisation multiples, une définition imprécise de la robustesse, une forte congestion de l'infrastructure et enfin des contraintes de temps de calcul et de pertinence industrielle.

6.1 Contributions

6.1.1 Aspects scientifiques

Ce travail se place dans un cadre d'analyse prescriptive, où des recommandations pour une décision optimale sont recherchées en se basant sur les résultats d'une analyse prédictive. La principale difficulté pour ce cas d'étude consistait à structurer la phase de prédiction pour répondre correctement aux enjeux du problème : les données disponibles contiennent peu d'informations et une prédiction ponctuelle classique du retard n'est pas envisageable. Par ailleurs l'objectif général est de proposer une solution au problème nominal qui se comporte bien malgré la possibilité d'aléas : on cherche donc une description de ces aléas, mais cette description ne doit pas se substituer aux valeurs d'origine.

Étant donné la structure du problème d'affectation des voies en gare, qui peut être vu comme une recherche de clique de taille maximum dans un graphe de compatibilité des circulations, l'approche imaginée pour la gestion de la robustesse s'est concentrée sur l'évaluation de la probabilité d'échec des arêtes impliquées. La robustesse est en effet gérée de cette façon dans le cadre déterministe, où des affectations sont jugées non robustes si l'espacement entre les trains sur des ressources conflictuelles est insuffisant. Ici, on a souhaité juger l'insuffisance de l'espacement en fonction des risques de retard des trains impliqués.

Deux techniques de modélisation ont été proposées pour estimer les probabilités de retards de trains conditionnellement à un ensemble de variables explicatives. La première utilise des modèles linéaires généralisés. Ces modèles permettent de décrire la distribution globale de la variable cible en représentant les paramètres d'une loi de référence par des modèles linéaires. La seconde technique

modifie la sortie de forêts aléatoires de manière à ce qu'elles renvoient une distribution et non une valeur ponctuelle. L'idée générale est d'utiliser chaque arbre pour séparer la base en sous-ensembles de données au sein desquels la distribution de valeurs est récupérée. Ces distributions sont ensuite moyennées sur les arbres de la forêt pour construire des distributions de probabilités individuelles. Dans le cas où les données de retards de trains sont des entiers positifs bornés, la variable cible est vue comme une variable catégorielle par la forêt.

L'évaluation des modèles est cependant complexe : les distributions sont potentiellement toutes différentes, et il n'est pas possible de valider ou exclure une distribution en se fondant sur une unique valeur observée. On a proposé pour cela une méthode de sélection et d'évaluation. La sélection se base sur un score de distance entre la distribution estimée et les valeurs observées, ce qui permet d'identifier efficacement le meilleur modèle parmi plusieurs en compétition mais ne donne pas d'évaluation de la qualité des prédictions. Dans cette thèse, on définit la qualité d'un modèle d'estimation de distributions conditionnelles selon sa discrimination et sa calibration. La discrimination consiste à être en mesure d'identifier correctement et de classer les individus selon leur risque, et est calculée ici à partir de la statistique c. La calibration signifie qu'en moyenne les probabilités estimées sont concordantes avec les taux d'évènements observés. Plusieurs méthodes sont possibles, mais nous recommandons ici l'utilisation de graphiques de calibration. Les tests d'Hosmer-Lemeshow et l'étude des résidus de quantiles reposent sur des tests statistiques qui sont peu stables et difficilement interprétables, rejettent rapidement des déviations acceptables pour les cas pratiques et dépendent trop fortement de la taille de l'échantillon. Même si les graphes de calibration n'offrent pas de garantie statistique, ils permettent de visualiser facilement les déviations et de laisser soin au décideur si elles sont acceptables pour le cas d'étude.

A notre sens, la calibration et la discrimination doivent être attestées conjointement pour assurer une prise de décision bien informée. Si le modèle statistique n'est pas discriminant, il ne produit pas d'estimations variées, et l'information apportée ne permettra donc pas d'arbitrer entre les scénarios puisqu'ils seront tous équivalents. La calibration est également nécessaire. Dans notre cas, plusieurs modèles différents sont entraînés selon le type et le sens de mouvement ; il faut que les modèles soient calibrés individuellement pour assurer la conservation des propriétés discriminantes des modèles seuls quand les prédictions sont utilisées ensembles. Ensuite la calibration assure d'avoir une juste évaluation du risque dans la prise de décision. Par exemple dans notre cas on souhaite ordonnancer les circulations de façon à avoir des marges en adéquation avec les taux de retards.

Une fois le modèle final sélectionné et validé, il est appliqué pour estimer les probabilités de retard de tous les trains commerciaux de la journée dont on souhaite optimiser les placements en gare. Ces distributions sont utilisées pour calculer les probabilités que les marges entre chaque paire de trains soient insuffisantes, et pondérer à partir de cela les arêtes du graphe de compatibilité. Étant donné la taille de la gare et le nombre de trains en circulation, le nombre de sommets du graphe est grand, jusqu'à 20 000 en excluant les affectations interdites. Le calcul même des pondérations prend un temps très important car les paires de trains doivent être considérées une à une. Des stratégies d'optimisation ont été proposées pour compenser ces difficultés liées à la taille des données et à la complexité de l'information à intégrer. En particulier, plusieurs heuristiques et métaheuristiques ont été développées sur le principe d'un graphe qu'on construit au fur et à mesure de l'exploration, permettant ainsi de ne pas calculer les poids liés à des solutions qui ne seront jamais considérées.

6.1.2 Aspects ferroviaires

Les données de retards ont fait l'objet d'un intérêt croissant ces dernières années. Les premiers travaux se sont concentrés sur leur distribution, et ont souvent conclu en faveur de la loi exponentielle grâce à des tests statistiques du type Kolmogorov-Smirnov. Ces expériences ont cependant été entreprises sur des bases de données de petite taille, en général sans prise en compte du contexte, et avec

une faible couverture temporelle ne permettant pas d'intégrer les différentes périodicités. D'autres travaux plus récents ont trouvé des distributions plus adéquates, telles que les lois lognormale ou Weibull, cependant la loi exponentielle est encore très largement utilisée pour modéliser les retards dans la prise de décision ou la simulation ferroviaire.

Ces travaux ont permis d'enrichir les connaissances actuelles concernant la modélisation des retards. Bien que ce soit encore peu fait dans la littérature, il nous semble incontournable d'inclure les éléments de contexte dans l'estimation de la distribution de retards. L'instabilité des circulations est fortement influencée par la densité du trafic, les marges planifiées, etc. et certaines dessertes sont plus sensibles que d'autres aux perturbations. On a pu observer de grandes variations de risque au sein d'un même type de circulation, comme par exemple des Transiliens à l'arrivée avec plus de 90% de chance d'observer un retard quand ce taux n'est que de 10% pour d'autres. Par ailleurs, ces travaux ne soutiennent pas l'idée d'utiliser une distribution paramétrique pour modéliser les retards, malgré une bonne adhérence a priori. On a pu constater que les forêts aléatoires détectaient des motifs nouveaux dans les données, avec des distributions moins lisses.

Cette thèse s'inscrit également dans une dynamique nouvelle d'application de Machine Learning aux données ferroviaires. Peu d'études se sont consacrées à la modélisation hors-ligne des retards, la majorité des travaux portant sur de la prédiction d'affluence ou de retards en temps réel. L'analyse en phase d'adaptation a montré qu'il était possible d'identifier des tendances dans les retards, en particulier dans ceux des Transiliens pour lesquels les performances sont les meilleures. On est en mesure d'identifier automatiquement les trains à haut risque de retard de ceux pour lesquels il est plus faible, et de quantifier avec fiabilité ce risque sous forme d'une distribution de probabilité dont on sait évaluer la qualité. Les prédictions pour les TGV sont moins bonnes, ce qu'on peut expliquer par leur exploitation moins dense, les distances parcourues et une plus grande insensibilité aux flux voyageurs.

Les retards sont un problème majeurs pour les acteurs ferroviaires, à la fois car la ponctualité est un indicateur clé de qualité de service pour les voyageurs, mais aussi car ils mettent la production sous pression en raison de la forte congestion du réseau. Si la planification des ressources est par nature complexe à produire, il est nécessaire d'y intégrer les possibilités d'aléas pour assurer un bon déroulement des opérations. La modélisation des retards pour la prise de décision est un sujet encore nouveau mais qui prend de l'importance, par exemple avec des modèles d'évaluation de la robustesse par simulation. Cette thèse propose une nouvelle application d'un modèle de probabilités individuelles de retards pour les opérations en gare, permettant de juger la robustesse des enchaînements de circulations sur des ressources critiques grâce à des observations passées. Si le modèle ne prend pas encore en compte les effets de trafic de manière globale, cette approche a tout de même permis une forte réduction des conflits réels en avant gare.

6.2 Limites rencontrées

On recense ici les limites qu'on a pu identifier lors de ces recherches. Certaines limitations propres à la modélisation statistique ou à l'optimisation de la robustesse sont rappelées brièvement mais ne seront pas explicitées plus en détail. On s'intéresse surtout ici aux limitations de la méthodologie dans son ensemble.

6.2.1 Évaluation de la qualité

La question de l'évaluation de la qualité des solutions produites est centrale ici, à la fois pour la validation des modèles statistiques et pour la comparaison d'algorithmes d'aide à la décision. Dans les

deux cas, on dispose des données réelles observées et on souhaite attester de la bonne représentation du risque, que ce soit par des probabilités individuelles ou par l'identification de conflits.

Concernant l'évaluation de probabilités conditionnelles, on a pu constater un vide méthodologique. Le problème était relativement bien traité pour le cas de données binaires, mais les autres supports de données, notamment discrets, ont beaucoup moins attiré l'attention. Si dans le cas binaire la comparaison entre une observation valant 0 ou 1 et la prédiction qui est un nombre décimal est déjà difficile à faire rigoureusement, le cas des variables non binaires est encore plus complexe et moins intuitif. On a proposé une méthodologie de validation basée sur des généralisations aux données de comptage de ces approches dédiées au cas binaire.

La robustesse des GOV est difficile à mesurer, et les indicateurs proposés ne la représentent que partiellement. En effet, son évaluation dépend de la bonne représentation de l'aléa et des interdépendances entre les circulations en cas de perturbations. Ces travaux ont apporté en connaissances sur l'aléa, cependant le second point est encore à éclaircir. Dans l'idéal, la modélisation du trafic sous perturbations devrait être collective et intégrer des éléments simples et réalistes de replanification, comme par exemple les options de reroutage de train, de changement d'ordre, et de propagation des retards. Dans ces travaux, on ne considère que la propagation du retard d'un train sur les circulations qui le suivent immédiatement. Cette politique casse indirectement les chaînes mais ne représente pas de manière réaliste la capacité d'adaptation de la planification. La modélisation collective est cependant complexe et nécessiterait un module d'optimisation à elle seule, ce qui n'est pas réaliste. La mesure de la robustesse aux retards est donc contrainte à être approximative.

6.2.2 Prédictabilité des retards

Dans ces travaux, l'étude est concentrée sur l'analyse de petites perturbations, dont on a fait l'hypothèse qu'elles étaient partiellement dues à des instabilités prévisibles du réseau. Si les retards primaires sont imprévisibles, tant dans leur amplitude que sur l'heure ou le lieu où ils vont survenir, il y a des contextes favorisant la propagation et d'autres qui au contraire permettent une absorption rapide. C'est indirectement ces contextes qu'on cherche à identifier pour estimer les probabilités de retards, ce qui pose plusieurs difficultés.

Tout d'abord on a pu identifier un problème de stabilité des prédictions au cours du temps. Même si des motifs sont clairement identifiables, les performances baissent entre les données de validation et celles de test qui sont postérieures à l'apprentissage. Les modèles construits sont donc sensibles aux variations dans le réseau ou dans les planifications, ce qui ne peut pas toujours se prévoir. L'impact de ces imprécisions sur la prise de décision est cependant compliqué à évaluer. Étant donné les graphes de calibration et des taux de concordance sur les données de test, les probabilités estimées restent cohérentes.

La qualité des données altère également le potentiel d'apprentissage. L'information disponible est limitée et repose sur une description macroscopique du trafic le long des lignes et dans les gares parcourues. Une description plus précise, par exemple par les espacements aux différents trains rencontrés ou les occupations de voies prévues, pourrait faire gagner en précision le modèle, mais ces données ne sont pas forcément accessibles et sont susceptibles d'évoluer pendant la phase d'adaptation. La précision de mesure peut aussi bruiter les prédictions. Ici les retards sont mesurés en minute, ce qui permet de gommer les variations liées au mauvais l'alignement des capteurs. Cependant comme les trains étudiés ont une forte ponctualité et qu'on n'étudie que des petites valeurs, plusieurs motifs n'apparaissent pas en raison d'une mesure trop grossière.

6.2.3 Question de la troncature

Une forte hypothèse de modélisation a été appliquée ici avec la troncature des retards les plus importants. A notre sens, cette troncature est nécessaire pour plusieurs raisons, mais elle crée également un cadre à la méthodologie dont on ne peut pas forcément s'abstraire.

La motivation principale de cette troncature est d'isoler de la base de données les retards vus comme anormaux : les retards de plusieurs dizaines de minutes sont rares et ont des causes imprévisibles qui ne reflètent pas l'instabilité récurrente des opérations qu'on cherche à identifier. Ensuite, cette troncature permet d'adapter le support des distributions estimées aux besoins en terme de robustesse.

Borner les retards présente d'autres avantages dont on a pu tirer profit dans ces travaux. Tout d'abord cela a permis d'utiliser des forêts aléatoires pour représenter les retards dans la mesure où ils pouvaient être modélisés comme des variables catégorielles ordonnées. Ensuite, cela donne plus de contrôle de la complexité de l'adaptation des GOV. Les arêtes des graphes utilisés ici sont pondérées à partir des probabilités de conflits de leurs sommets. Sans la troncature, des circulations très espacées dans le temps se verraient associer une pondération non nulle, ce qui augmenterait le temps de calcul sans gain particulier, car en pratique, si un train a un retard conséquent et génère des conflits en avant gare, des mesures de régulation seront appliquées.

Enfin, l'hypothèse d'indépendance des circulations a été appliquée pour la modélisation des retards et pour les calcul des probabilités de conflits en avant gare. On peut facilement justifier cette hypothèse dans le cas de trains de types distincts ou circulant à des jours différents, mais elle est plus difficile à affirmer dans le cas contraire. La troncature des retards permet partiellement d'assurer cette indépendance si la valeur de troncature est inférieure à la fréquence sur la ligne.Une étude plus approfondie des profils de trains en conflit potentiel dans la partie d'optimisation serait à entreprendre pour évaluer la validité de cette hypothèse, notamment selon les seuils de troncature choisis.

Le choix de la valeur de troncature est complexe à figer. On a pris le parti ici de choisir des seuils représentant la limite des retards rares selon le type et le sens de circulation choisis. Néanmoins, un cadre harmonieux avec une même valeur de troncature permettrait de construire une unique forêt aléatoire, ou encore d'évaluer ensemble les distributions estimées, ce qui n'est pas possible ici pour la discrimination car le retard moyen est trop fortement dépendant de la valeur de troncature.

6.2.4 Cas de la gare de Montparnasse

L'utilisation de la méthodologie sur le cas d'étude de la gare de Paris Montparnasse donne des résultats prometteurs, tant sur l'estimation de risque de retards que pour la réduction de conflits ; cependant on peut identifier plusieurs points limitant les performances.

Le point principal est la très forte présence de mouvements techniques dans la gare en raison des arrivées et départs de technicentres à proximité. Ces mouvements ont une faible priorité, difficilement quantifiable pour la conception des GOV, et une très forte variabilité de leurs retards. Dans de nombreux cas, ces mouvements techniques sont avancés ou retardés durant les opérations pour faciliter les circulations en gare. On ne peut donc pas les expliquer conditionnellement au contexte comme cela est fait pour les circulations commerciales. Cette disparité d'information complique l'adaptation des occupations des voies, puisqu'on cherche à agencer avec précision des trains pour lesquels on dispose d'une estimation fiable de leur distribution de retards, et d'autres pour lesquels elle est inconnue. Ici un facteur de priorité permet d'arbitrer, mais une homogénéité des données d'entrées serait préférable.

L'exploitation de la gare limite également les capacités d'amélioration de la robustesse. En effet, toute l'infrastructure de la gare n'est pas disponible à tous les trains, et les capacités d'évolution dépendent fortement du type de circulation. Par exemple les Transiliens et les TGV se croisent très peu puisqu'ils ont des voies en ligne et des voies à quai toujours différentes. Ainsi, les choix d'affectation de voies et itinéraires sont principalement faits entre des circulations partageant beaucoup de carac-

téristiques communes (type de train, temporalité, densité en gare, etc.). Des trains ayant plusieurs variables explicatives proches peuvent avoir des fonctions de risque similaires, que ce soit par GLM ou par forêts aléatoires. Les différents noeuds alternatifs à une association train/itinéraires dans le graphe de compatibilité peuvent avoir des poids équivalents, ne permettant pas d'améliorer la robustesse.

6.3 Perspectives

6.3.1 Pistes d'amélioration

Concernant le module d'estimation des distributions de retards, nous considérons que les limites proviennent plus des données que d'une faiblesse de modélisation. On recommande simplement d'essayer d'exploiter l'ordre entre les valeurs de retards dans l'algorithme de forêts aléatoires ou de construire des modèles linéaires généralisés en deux parties avec une forêt aléatoire pour estimer la probabilité d'un retard nul. Cette seconde option serait surtout à explorer pour le cas où les données ne sont pas tronquées ou à support continu. Dans le cas de retards continus, il est possible d'appliquer la méthodologie de Bertsimas et Kallus [20] qui est proche de la notre et utilise des forêts. Elle peut s'adapter aux données continues mais n'intègre pas de module d'évaluation des distributions.

De nouvelles données devraient être ajoutées à la base pour gagner en pouvoir prédictif, comme par exemple des informations sur les travaux planifiés sur les voies ou une description plus précise des opérations. En particulier, il serait intéressant de coupler les données aux GOV théoriques qui contiennent par exemple les informations sur les retours de maintenance et les espacements en gare, ou encore la durée d'occupation du quai qui serait pertinente pour les trains au départ. Certaines de ces données sont cependant susceptibles d'évoluer, mais peuvent être mises à jour au fur et à mesure pour recalculer les GOV. De même pour les données météorologiques qui peuvent être estimées quelques jours en avance avec une relative précision.

Le plus gros point d'amélioration concerne l'adaptation des GOV en fonction du risque. Les algorithmes proposés ici reposent sur une minimisation de l'impact individuel de chaque train sur les trains voisins. Une amélioration conséquente consisterait à modéliser les impacts collectifs afin de briser les chaînes de propagation des retards. Cet impact collectif peut se mesurer par la quantité de retards secondaires générés par des scénarios, soit dans le cadre de simulation stochastique, soit dans un contexte d'optimisation face à des aléas explicitement définis dans un ensemble d'évènements.

6.3.2 Autres cas d'études

Choix de la gare : les mouvements techniques ont posé plusieurs difficultés dans ces travaux : l'incertitude autour de leurs horaires ne peut pas être quantifiée, elle dépend trop fortement du déroulement des opérations et des décisions prises par les agents, et leur circulation est moins prioritaire que les circulations techniques. On recommande de privilégier des cas d'étude où les mouvements commerciaux seraient très majoritaires.

Les modèles prédictifs construits pour les retards des circulations commerciales sont particulièrement performants pour les Transiliens et un peu moins bons pour les TGV. On conseille tout de même de choisir des gares ayant un trafic hétérogène : comme cela a été décrit dans les limites du cas de la gare Montparnasse, il est préférable d'appliquer la méthodologie dans un cas où les conflits impliquent des circulations très différentes. En effet, une variable telle que le jour, la plage horaire ou la densité en gare, peut caractériser des profils de retards très variés pour les différents types de trains. On pourrait alors tirer parti d'une plus grande ponctualité de certains trains pour en protéger d'autres selon le contexte. Le périmètre des gares est donc favorable pour cela, mais on suppose que la

méthodologie serait plus profitable pour le cas d'une gare où les éléments d'infrastructure sont moins disjoints selon les types et sens de circulation.

La gare Montparnasse est une gare terminale, cependant il est tout à fait possible d'appliquer cette méthodologie à une gare passante. La différence principale est que les retards au départ sont fortement liés aux retards à l'arrivée, tant dans leur distributions de probabilité que dans la propagation des retards car l'occupation à quai ne permet plus d'absorber les retards à l'arrivée. Dans l'évaluation de la robustesse il faudrait considérer qu'un conflit de faisabilité peut pénaliser doublement un même train à son arrivée et à son départ. Une seconde différence est que les gares passantes sont bien moins concernées par les circulations techniques car les technicentres se situent près des gares terminales.

Autres problèmes ferroviaires : cette méthodologie peut s'appliquer à d'autres problèmes où l'information sur les potentielles perturbations est nécessaire pour la prise de décisions. On recommande tout de même d'évaluer les points suivants :

- le problème qu'on cherche à résoudre doit contenir une incertitude sur les données d'entrées qui puisse être intégrée sous forme d'une distribution de probabilité. Pour certains problèmes une prédiction ponctuelle est à privilégier. C'est par exemple le cas de la gestion des circulations en temps réel : la robustesse n'est plus un enjeu, l'incertitude sur les retards est plus faible car l'information est actualisée fréquemment et les perturbations ne sont plus dispersées autour de la valeur nominale mais autour d'une nouvelle valeur de retard. Dans le cas où on cherche simplement à définir les bornes de variations des paramètres, l'estimation de probabilités peut être utile mais des modèles de régression de quantiles peuvent être suffisants. A notre avis le format de la distribution de probabilité est particulièrement adapté pour les cas où la valeur nominale est évidente, comme ici car les trains arrivent très largement à l'heure, mais où on ne sait pas quantifier la dispersion des valeurs, et on n'est quoi qu'il arrive pas en mesure de protéger les circulations de manière systématique en raison de la congestion.
- une étude préliminaire des données va permettre d'évaluer si les méthodes proposées ici sont adaptées. Pour pouvoir utiliser un GLM, les perturbations doivent pouvoir être représentées par une ou plusieurs distributions de référence. Pour utiliser des forêts aléatoires, le support de la variable cible doit être discret et fini. Il est toujours possible de discrétiser et tronquer les données si besoin mais cela s'accompagne d'une perte d'informations. Dans le cas de données continues et non modélisables par des GLM, on recommande d'appliquer des méthodes d'estimation de probabilités non-paramétriques de type estimateurs de noyaux pour modéliser une distribution continue à partir des fréquences renvoyées par les noeuds terminaux, ou la méthodologie d'analyse prescriptive de Bertsimas et Kallus.
- l'incertitude doit pouvoir s'intégrer judicieusement pour répondre aux enjeux du problème. La description de la distribution complète permet par exemple de mettre en place un cadre de simulation calibrée ou de pénaliser des décisions en objectif.

6.3.3 Pertinence d'une industrialisation

La méthodologie proposée ici exploite une large quantité de données, avec des types de trains et des mouvements différents observés sur une longue période, afin de proposer une mesure de l'incertitude sur les horaires. On peut se demander dans quelle mesure une industrialisation est possible.

La première brique permettant d'estimer les probabilités de retards peut éventuellement être allégée. On a pu voir dans ces expérimentations que les modèles par forêts aléatoires produisaient systématiquement des distributions meilleures que les modèles linéaires généralisés. On peut se contenter de construire des forêts aléatoires, qui sont aussi plus simples à mettre en place que les GLM pour lesquels des sélections de variables sont nécessaires. Par ailleurs, si une unique valeur de troncature

est utilisée, il est possible de ne construire qu'une seule forêt car l'algorithme sépare par lui-même les données pour reconstituer des profils de retard homogènes.

Les contraintes de temps de calcul sont moins importantes pour la phase d'analyse statistique. Les bases de données peuvent être mises à jour et traitées de manière automatique à intervalles réguliers, et les modèles peuvent être construits pendant la nuit, surtout les forêts aléatoires qui ne prennent que quelques minutes à être calculées. Ce n'est cependant pas le cas des algorithmes d'adaptation de GOV qui sont potentiellement directement utilisés par des agents, par exemple pour tester des configurations. On a montré dans ces travaux qu'une amélioration rapide des conflits en gare était possible et plusieurs stratégies ont été mises en place pour optimiser au mieux malgré la taille des instances, cependant des efforts complémentaires devraient être consacrés à l'amélioration du code en cas d'industrialisation.

La question de la confiance dans les données doit être également être mise au clair avant une industrialisation. La méthodologie conçue fonctionne en boîte noire et est peu interprétable, ce qui peut rendre difficile son acceptation, car on n'est pas toujours en mesure de justifier les estimations des modèles. Par ailleurs, la question de la confiance des données se pose également en raison du problème de stabilité au cours du temps. Rien ne garantit la validité des modèles dans le futur, en particulier en cas de changements majeurs sur le réseau. Des analyses pratiques menées a posteriori ont mis en évidence que les modèles fonctionnaient correctement quand ils étaient appliqués à des données futures, mais le risque que le modèle se trompe doit être accepté.

Enfin, des questions d'équité doivent se poser avant l'industrialisation d'une telle approche. Dans un contexte d'ouverture à la concurrence, le gestionnaire d'infrastructure doit garantir un accès au réseau égalitaire entre les différentes entreprises ferroviaires. Affecter plus de capacité aux trains qui ont un fort risque de retard peut être vu comme une récompense aux mauvais élèves.

6.3.4 Synthèse

L'objectif de développement de modèles apprenants a été rempli dans le sens où les données du passé ont été exploitées pour ajouter de la connaissance pour la résolution du problème d'affectation des voies. Ces sujets sont fortement d'actualité, tant dans le domaine des mathématiques appliquées, où de nombreuses recherches portent sur la coordination de l'apprentissage statistique et de l'aide à la décision, que dans le domaine ferroviaire où les données de retards sont de plus en plus analysées pour permettre de mieux identifier et anticiper les aspects critiques de la production. De futures recherches pourront enrichir la méthodologie, étudier de nouveaux cas d'étude ou préparer une potentielle industrialisation.

Troisième partie

Annexes

Annexe A

La gare Montparnasse

A.1 Trafic de la gare Montparnasse

Concernant les trains commerciaux en circulation en gare de Paris Montparnasse, les circulations commerciales sont :

- Les trains TGV : ils desservent les régions Bretagne, Pays de la Loire, Centre Val de Loire et Nouvelle Aquitaine, ils correspondent à environ 36% du traffic
- Les trains régionaux (15% des circulations) :
  - TER Centre-Val de Loire vers Le Mans, Chartres et Nogent-le-Rotrou
  - TER Normandie et anciens Intercités vers Argentan et Granville
- Les trains Transilien de la ligne N. Ils assurent la desserte vers Mantes-la-Jolie, Dreux et Rambouillet, en intégrant notamment les noeuds de Versailles et Sèvres et reprsentent 49% des trains commerciaux.

Ligne	Tranche	Divers
Argentan	16500-16549	Train régional (Granville)
Le Mans	16700-16800	Train régional
Le Mans	862400 à 862599	TER Paris Chartres Nogent Le Mans
Granville	3400-3499	Grande ligne en réservation
Granville	13240 à 13279	Corail IC sans réservation
St Malo	8000 à 8099	TGV Rennes St Malo
La Rochelle	8300 à 8399	TGV Tours Poitiers La Rochelle
Arcachon	8400 à 8499	TGV Arcachon Bordeaux Angouleme
Après Bordeaux	8500 à 8599	TGV après Bordeaux sauf Arcachon
Brest	8600 à 8699	TGV Rennes St Brieuc Lannion Brest
Quimper	8700 à 8799	TGV Rennes Vannes Lorient Quimper
Nantes	8800 à 8899	TGV Le Mans Nantes
Après Nantes	8900 à 8999	TGV après Nantes
Sèvres	164100 à 164199 et 165100 à 165199	TN
Versailles	164200 à 164299 et 165200 à 165299	TN
La Verrière	164300 à 164399 et 165300 à 165399	TN
Rambouillet	164400 à 164499 et 165400 à 165499	TN
Plaisir Grignon	164500 à 164599 et 165500 à 165599	TN
Mantes La Jolie	164600 à 164699 et 165600 à 165699	TN
Houdan	164800 à 164899 et 165800 à 165899	TN Houdan Dreux
Montfort	164900 à 164999 et 165900 à 165999	TN
Adaptation	163950 à 163999	TN Adaptation du service horaire

Tableau A.1 – Numérotation des circulations à Paris Montparnasse

TGV : Les lignes TGV desservant Paris Montparnasse sont notées dans la carte A.1. Les lignes principales passent par Rennes, Nantes et Bordeaux. Les trains circulent sur des voies grandes vitesse dédiées entre Paris et Rennes et entre Paris et Bordeaux, ce qui assure une circulation homogène à haute vitesse.

FIGURE A.1 – Carte du réseau TGV Atlantique

TN : Les Transiliens correspondent à du trafic en zone dense. Ils circulent de manière très régulière et transportent des masses importantes de passagers. Ils sont sensibles à l'affluence en gare, dans la mesure où des temps d'occupation en gare trop importants peuvent générer des retards qui se propagent vite entre les circulations. En contrepartie, ces circulations offrent une flexibilité importante en opérationnel car elles peuvent être supprimées ou voir leur desserte modifiée pour permettre un

retour à la normale plus rapide. Les distances parcourues sont assez courtes puisqu'ils restent en Ile-de-France, mais ils assurent de nombreux arrêts.

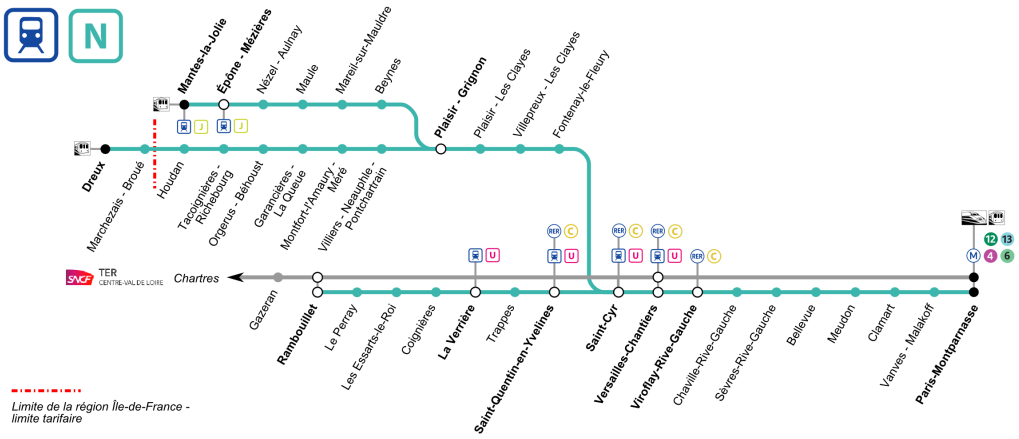

FIGURE A.2 – Schéma de la ligne N

TER et Intercité : les TER Centre et Normandie (ancien Intercités) circulent à assez basse fréquence, mais ont des caractéristiques communes aux transiliens, par exemple sur leur rôle dans les trajets professionnels. La figure A.3 schématise les arrêts desservis par la ligne TER Normandie.

FIGURE A.3 – Schéma de la ligne TER Normandie et anciens intercités

A.2 Infrastructure de la gare

A.2.1 Les points remarquables

Les points remarquables ou PR du secteur de Paris Montparnasse qui ont été utilisés pour cette étude sont les suivants :

— Paris Montparnasse BV ou bâtiment voyageur : il s'agit des PR au niveau des quais, l'ensemble des trains s'arrêtant à Paris Montparnasse sont enregistrés

— Vaugirard BV : les trains s'arrêtant ou partant de la gare de Vaugirard sont enregistrés

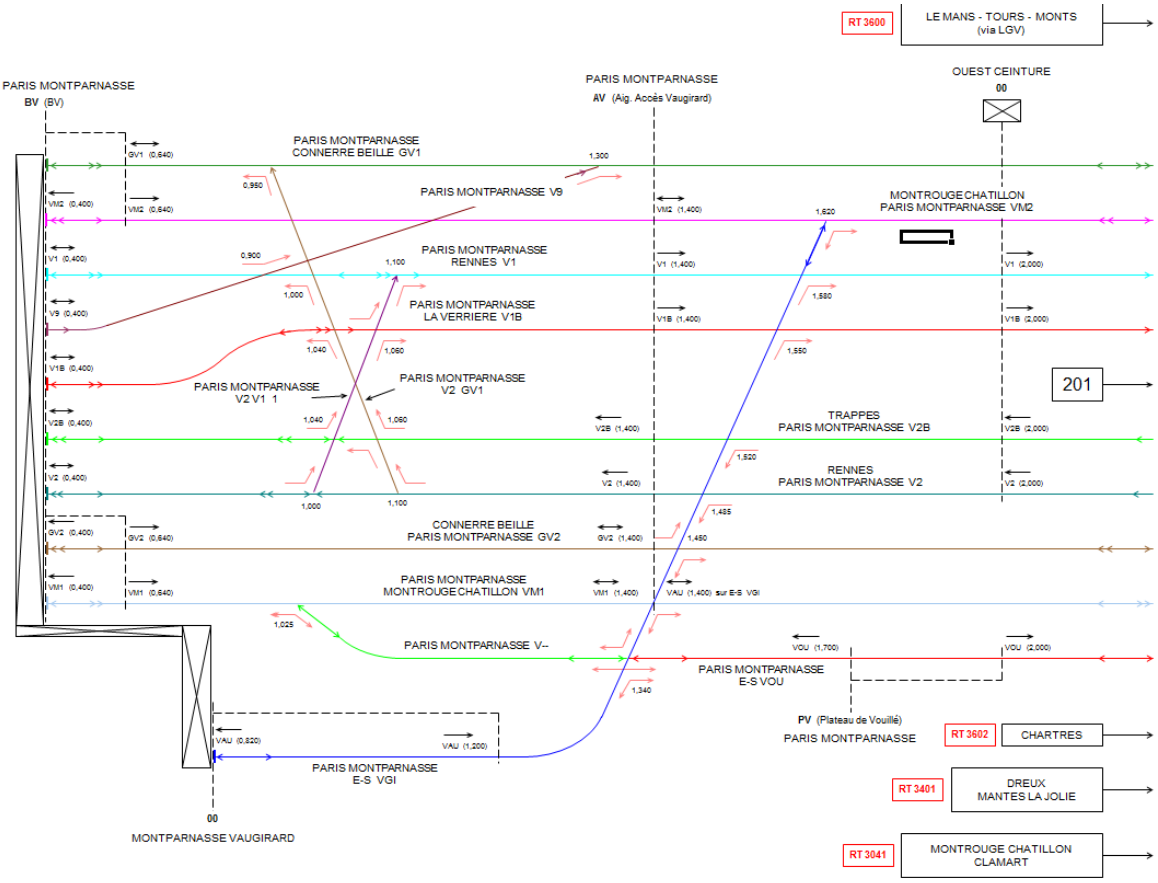

FIGURE A.4 – Plan des PR : début de la zone d'avant gare

— Paris Montparnasse AV ou accès Vaugirard : ce PR est situé au milieu de la zone d'avant gare et coupe une majorité des circulations environ 2 minutes avant leur arrivée et à 1,4km des quais de Montparnasse

— Paris Montparnasse VO ou accès Vouillé : ce PR est à 3 minutes du baptiment voyageur, soit 2,5km, et correspond aux voies empruntées par les Transiliens, Intercités et TER.

— Vanves-Malakoff BV : ce PR est le premier avant le secteur de la gare Montparnasse et l'ensemble des circulations Transiliens, Intercités et TER y passent à environ 4 minutes, soit 3,7 km de Montparnasse BV.

— Montrouge-Châtillon ES :il intercepte toutes les circulations TGV environ 5 minutes avant leur arrivée à quai

Des schémas localisant les PR peuvent être trouvés sur les figures A.4,A.5,A.6. Les temps de parcours moyen donnés ci-dessus correspondent à une estimation en conditions nominales.

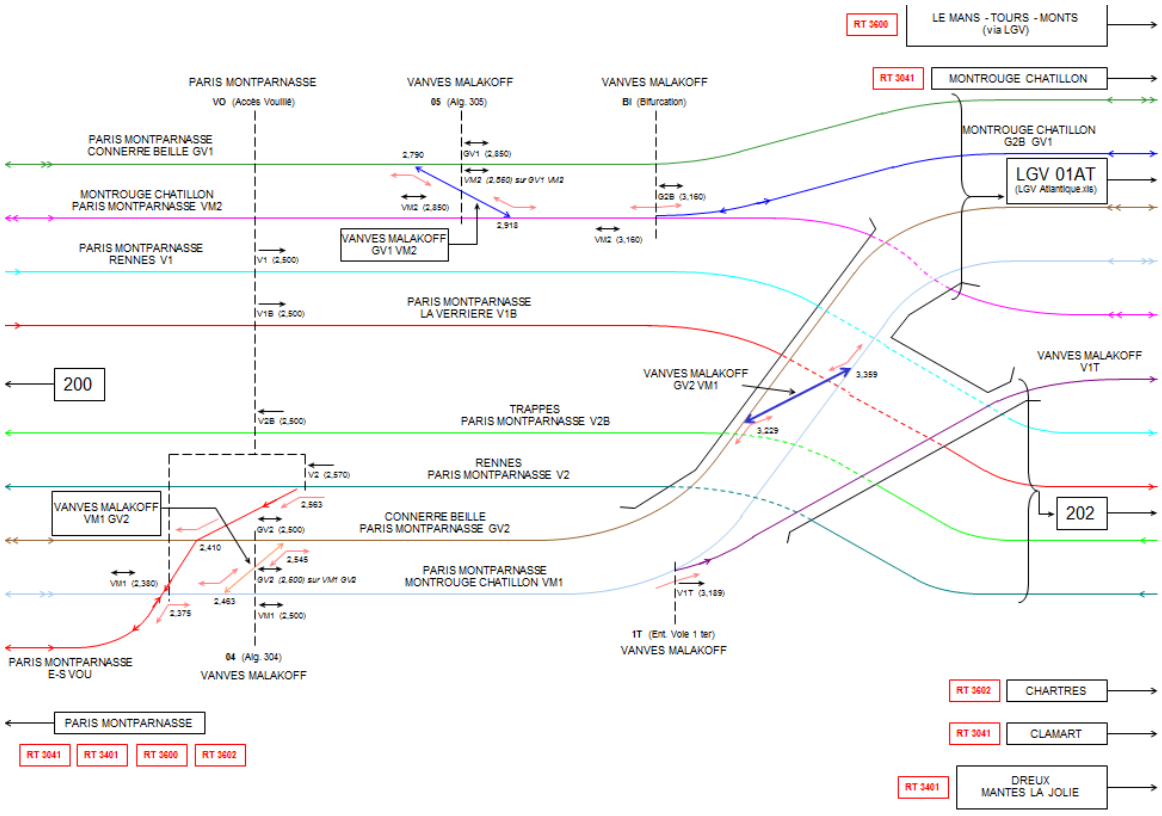

FIGURE A.5 – Plan des PR : fin de la zone d'avant gare côté TER et TN

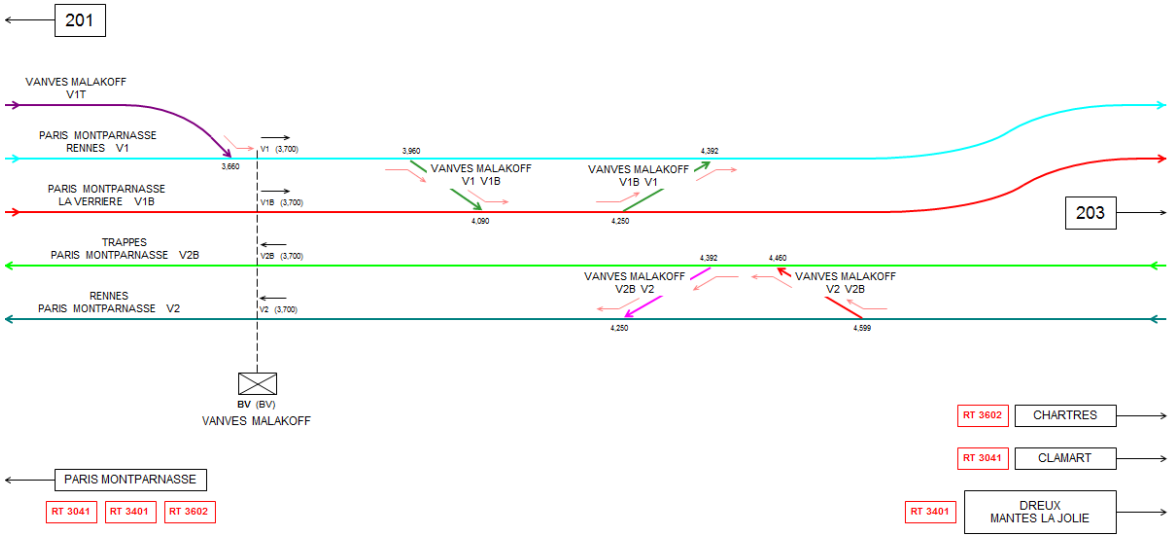

FIGURE A.6 – Plan des PR : début des voies en ligne côté TER et TN

A.3 Enjeux de la robustesse

Les données de retards recueillies au niveau des quais et en sortie de la gare permettent d'avoir une idée des variations des retards dans le périmètre de gare.

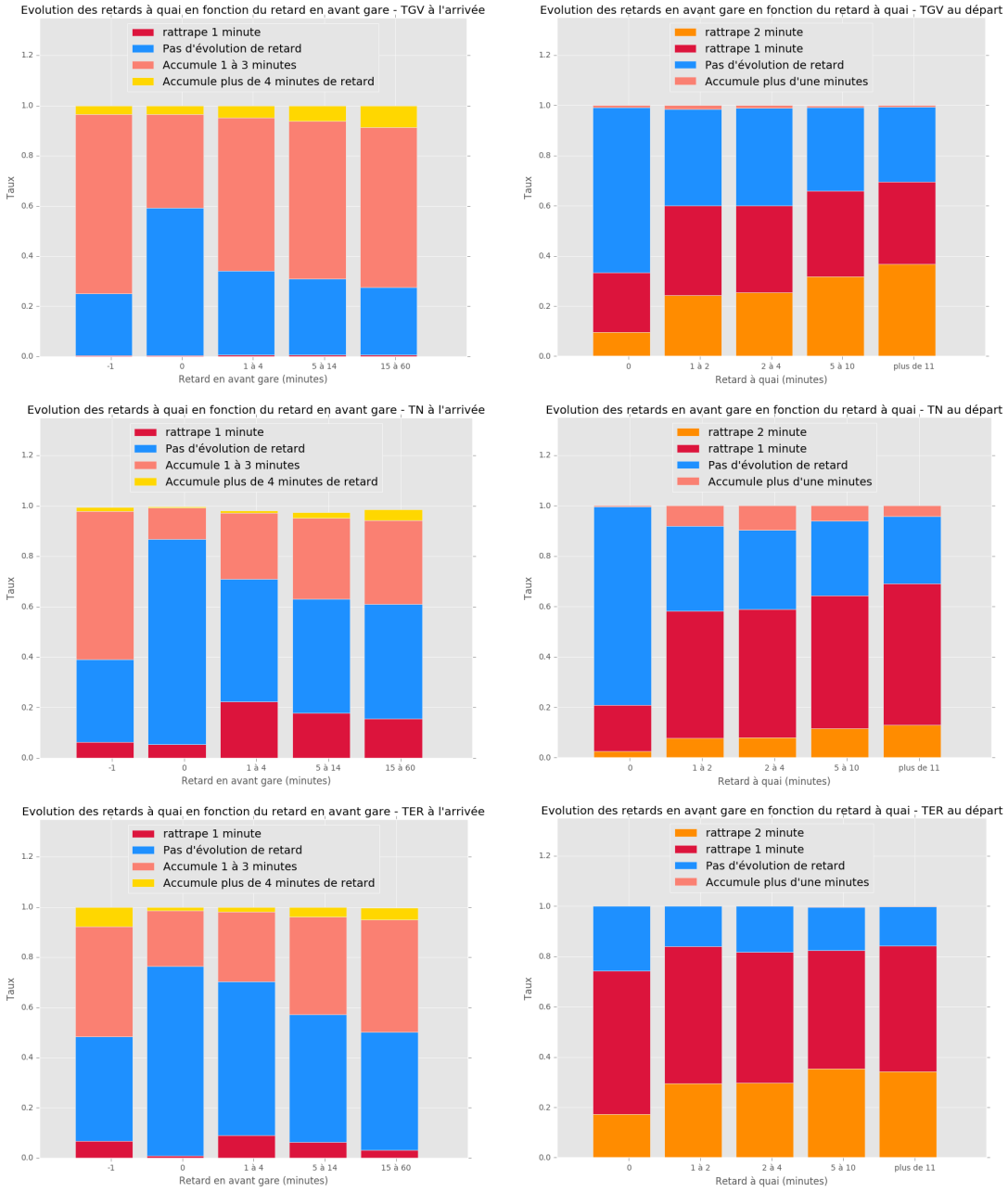

FIGURE A.7 – Propagation des retards entre le bâtiment voyageur et l'avant gare

Annexe B

Documentation de la base de données

B.1 Détails des différentes variables

On présente ici les variables utilisées pour les différents modèles et analyses de données. Certaines variables venant des données brutes ne sont plus utilisées ou alors transformées (comme la date ou le numéro de train) et ne seront pas rappelées.

B.1.1 Variables d'écart

Ecart : cette variable correspond au retard en minutes qui est modélisé dans ces travaux. Pour les trains au départ de la gare, on utilise directement les écarts horaires enregistrés au bâtiment voyageur des gares de Paris Montparnasse et Paris Vaugirard. Pour les trains à l'arrivée, il faut utiliser les données de balises placées en avant gare, cependant il n'y a pas de point remarquable commun à l'ensemble des circulations, et quelques cas de données manquantes ou absurdes peuvent être observés sur ces PR (beaucoup moins pour les données en gare). Pour les TGV, on utilisera la valeur de la balise placée à Châtillon-Malakoff (voir détail des PR en A.2.1). Pour les autres trains (TER Centre et Normandie, Transiliens), on utilisera la balise du bâtiment voyageur de Vanves-Malakoff.

B.1.2 Variables de type de circulation

Structure : Il s'agit de l'entreprise ferroviaire responsable des trains. La base de données contient alors une variable binaire par type de circulation, à savoir TGV, TN, TER Centre et TER Normandie (dont anciens intercités).

Série : Cette colonne fait partie des données brutes extraites de Brehat et correspond au matériel prévu pour le train. On aura alors une variable binaire par type de matériel.

B.1.3 Variables temporelles

Heure (resp Heure_Depart) : c'est l'horaire prévu d'arrivée (resp. de départ) de la gare Montparnasse. La variable Heure_Depart représente l'heure de départ à l'origine pour le set de retards à l'arrivée. Toutes les variables d'heure sont converties en minutes et translatées pour valoir 0 à 3h du

matin (avant l'ouverture de la gare) et 1439 à 2h59. Cela permet d'avoir une continuité de l'heure passé minuit.

Jour : jours de la semaine numérotés entre 1 et 7, du lundi au dimanche

Vacances : variable binaire valant 1 pour les jours de vacances scolaires à Paris

Nuit (resp Dep_Nuit) : période avant 7h et après 20h30

Matin (resp Dep_Matin) : période entre 7h et 10h

Journee (resp_Dep_Journee) : période entre 10h et 16h

Soir (resp Dep_Soir) : entre 16h et 20h30

grands_departs : trajet entre le vendredi soir et le dimanche soir

trajets_pro : trajet le matin ou le soir en semaine

WE : variable binaire valant 1 le samedi et le dimanche

B.1.4 Variables de mission

Origine : nom de la gare d'origine, chaque origine ayant un taux d'occurrence suffisant sera encodée par une variable binaire. Les valeurs possibles sont :

- Lignes TGV : ANgers-St-Laud, Bordeaux St Jean, Brest, Hendaye, La Rochelle Ville, Les Sables d'Olonnes, Le Croisic, Nantes, Poitiers, Quimper, Rennes, St Brieuc, St Malo, St Nazaire, St Pierre des Corps, Tarbes, Toulouse et Tours
- Lignes TN : Dreux, La Verriere, Mantes la Jolie, Plaisir Grignon, Rambouillet et Sevres Rive Gauche
- Lignes TER Centre : Chartres, La Villette St Prest, Le Mans, Nogent le Rotrou et Rambouillet
- Lignes Intercité (TER Normandie) : Angentan et Granville

trajet_direct : vaut 1 dans le cas où aucun arrêt n'est desservi

NbArrets : nombre de gares avec arrêt planifié dont l'origine

nbTraversees : nombre de gares traversées sans arrêt

temps_arret : somme du nombre de minutes d'arrêt planifiées dans les différentes gares desservies

temps_arret_min : valeur minimale du temps d'arrêt planifié en gare

temps_arret_max : valeur maximale du temps d'arrêt planifié en gare

Duree : temps de trajet en minutes

marge : écart du temps de trajet du train avec la médiane du temps de trajet sur cette desserte. La valeur est fixée à 0 s'il n'y a pas assez d'observations pour conclure.

B.1.5 Variables de densité

premier_train_ligne (resp dernier_train_ligne) : variable binaire si le train est le premier (resp dernier) de la journée sur sa desserte

densitePrec20 : nombre de trains arrivant ou partant de Montparnasse dans les 20 minutes précédant le mouvement

densiteSuiv5 : nombre de trains arrivant ou partant de Montparnasse dans les 5 minutes suivant le mouvement

densite20_5_Origine : nombre de trains arrivant ou partant de la gare d'origine dans les 5 minutes suivant le départ du train

densite20_5_Arret (resp densite20_5_Traversee) : moyenne pondérée du nombre de trains arrivant ou partant de la gare d'origine dans les 5 minutes suivant l'arrêt (resp passage) du train dans les différentes gares du parcours

B.2 Visualisations

Les corrélations entre les variables sont montrées dans les figures B.1 et B.2. On constate que la variable *Ecart* ne montre pas ou peu de corrélation directe aux variables utilisées.

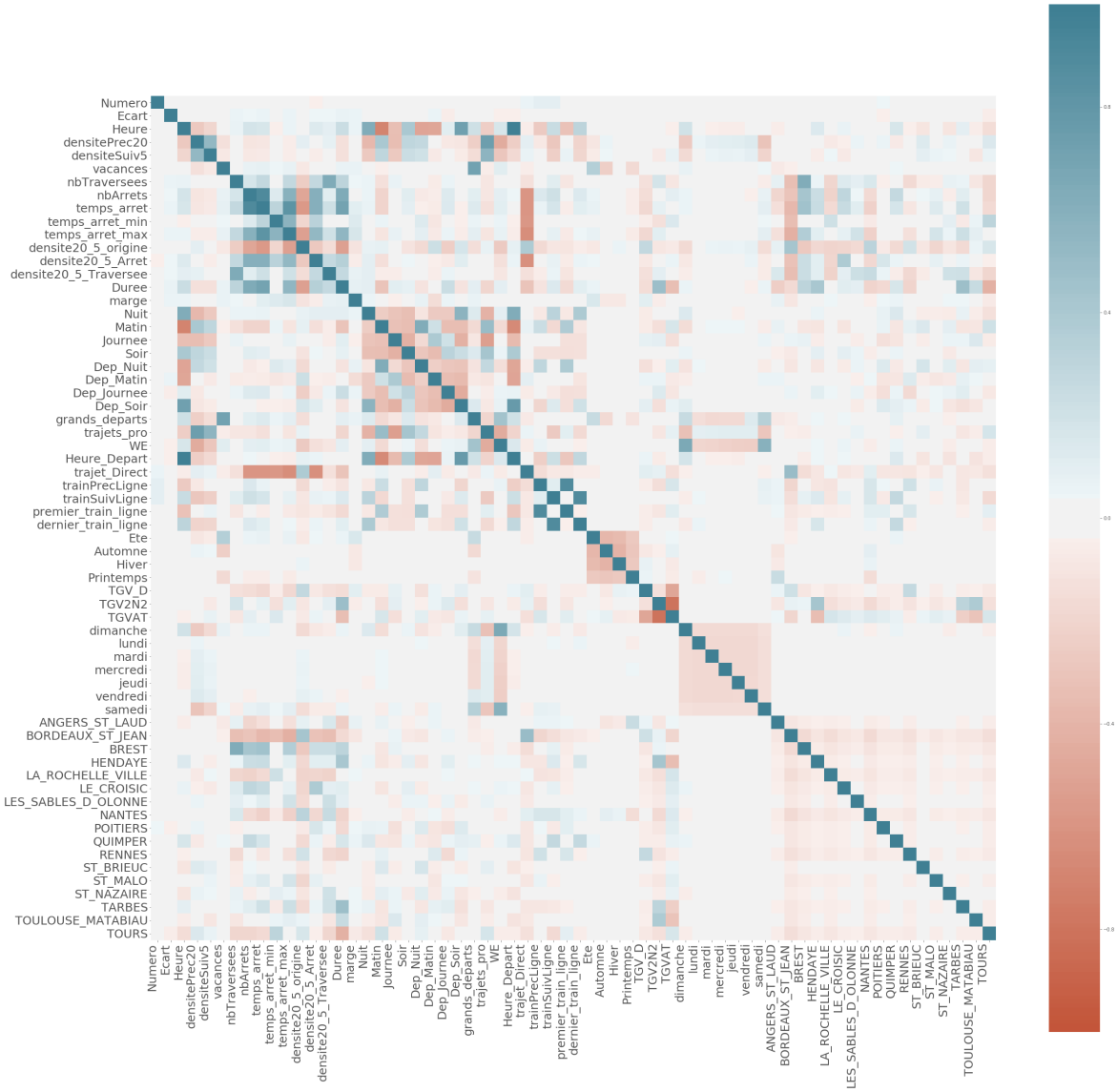

FIGURE B.1 – Corrélation des variables - TGV arrivée

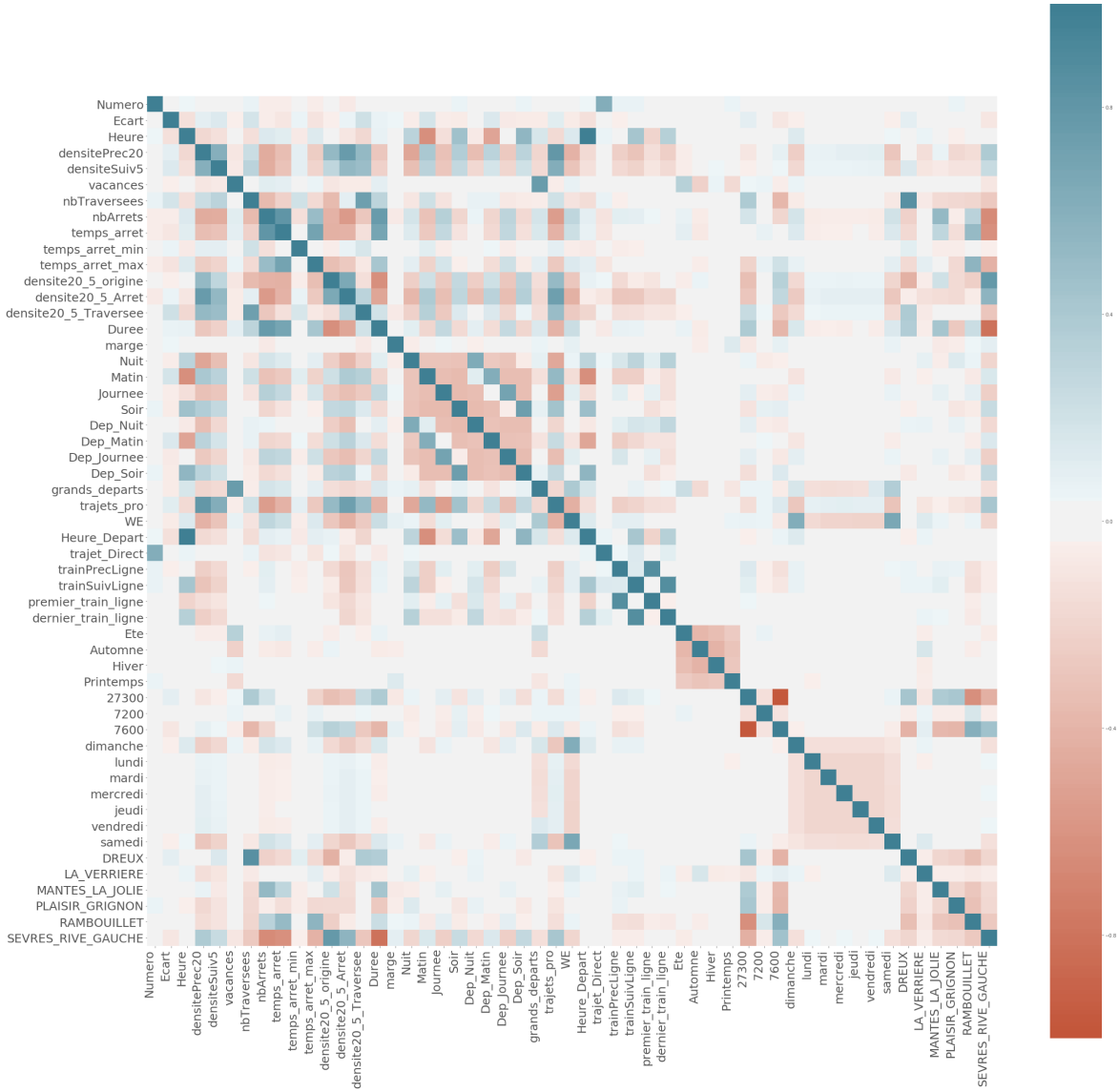

FIGURE B.2 – Corrélation des variables - TN arrivée

Annexe C

Modélisation par GLM

C.1 Évaluation préliminaire des stratégies

Quelques tests ont été effectués pour évaluer la relation entre la complexité du modèle (distribution choisie, ensemble de variables explicatives, nombre d'observations), le temps d'entraînement du modèle et sa qualité. Les tests sont effectués sur les données de trains TN à l'arrivée pour la période d'octobre (entraînement de septembre 2017 à septembre 2018, test sur octobre 2018). Les expériences suivantes sont conduites :

- Exp1 : distribution NBI, tous les features (57 auxquels sont ajoutés 5 features d'interaction entre l'origine du train et le nombre d'arrêts)
- Exp2 : distribution NBI, selection de 35 features (suppression des variables de jour et d'interaction)
- Exp3 : distribution NBI, selection de 17 features (suppression des variables de jour, plage horaire et d'interaction, tri dans les variables redondantes)
- Exp4 : distribution géométrique, tous les features
- Exp5 : distribution ZIPIG (Poisson inverse gaussien à inflation de zéros), même sélection de features que Exp3
- Exp6 : distribution ZIPIG (Poisson inverse gaussien à inflation de zéros), même sélection de features que Exp2

A chaque fois le nombre d'observations n_{sample} utilisées est variable et les temps de calculs sont moyennés sur 50 itérations avec sélection aléatoire avec remise de n_{sample} observations dans le set d'origine contenant 23510 trains. Le nombre d'itérations maximal dans l'algorithme est fixé à 100 et les temps de calculs sont donnés en secondes dans le tableau C.1.

On précise que lors du calcul, plusieurs itérations ont mené à des échecs quand les bases de données étaient trop réduites (a priori car certaines variables utilisées étaient constantes sur le set) et à des situations où le modèle ne converge pas, principalement pour les expérimentations utilisant la loi ZIPIG, et ce dès 500 observations.

La très forte dépendance du modèle en le nombre de paramètres de la distribution vient du fait que l'optimisation des coefficients pour chacun des paramètres est faite séquentiellement : l'algorithme optimise les coefficients pour μ en considérant les autres coefficients fixés, puis de même pour σ et ainsi de suite. Cette étape est renouvelée tant que la déviance globale ne converge pas.

n_{sample}	Exp1	Exp2	Exp3	Exp 4	Exp5	Exp6
Distribution	*NBI*	*NBI*	*NBI*	*GEOM*	*ZIPIG*	*ZIPIG*
$n_{features}$	*62×2*	*35×2*	*17×2*	*62*	*17×3*	*35×3*
100	0.1	0.4	0.5	0.1	2.9	3.8
500	1.0	1.5	0.9	0.5	5.9	10.0
1000	1.8	3.6	2.0	0.5	13.2	23.9
2500	19.6	7.8	2.9	0.8	42.2	66.1
5000	19.9	13.0	6.7	0.9	75.3	131.8
7500	36.0	17.6	7.0	1.2	116.1	211.6
10000	66.0	32.3	10.1	2.3	150.8	228.6
15000	75.1	44.9	17.5	1.7	274.5	323.4
20000	81.4	68.7	26.3	2.1	351.7	576.4
30000	148.8	85.4	40.9	2.4	547.5	828.3
50000	120.6	175.1	74.9	3.4	1073.8	1259.5

TABLEAU C.1 – Temps de calcul des modèles GAMLSS - données TN arrivée

	Exp1	Exp2	Exp3	Exp 4	Exp5	Exp6
RPS train	0.0568	0.0569	0.0588	0.0580	0.0603	0.0562
RPS validation	0.0560	0.0556	0.0577	0.0570	0.0591	0.0550

TABLEAU C.2 – Variation de RPS selon le modèle GAMLSS - TN Arrivée

Le tableau C.2 présente le RPS obtenu en construisant selon les différents cas testés lors de la construction du modèle sur le set de données complet. Le score est mesuré sur le set d'apprentissage et le set de validation.

C.2 Exemples de modèles détaillés

Cette partie donne le détail de tous les modèles pour la période 5 après sélection de variables. En particulier, on donne les temps de calculs nécessaires ainsi que la pénalité du critère BIC qui est égale au logarithme de la taille de l'échantillon. Les coefficients estimés pour chaque variable ainsi que leur significativité sont indiqués.

TGV Arrivée :

```
pen =  9.81334400268316  , iterations max =  30
temps de calcul  196.711044184367  minutes
*****************************************************************
Family:  c("NBItr", "right truncated Negative Binomial type I")
-----------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.69235    0.07126   9.717  < 2e-16 ***
Dep_Matin           0.29847    0.08063   3.702 0.000215 ***
Duree               0.39292    0.04501   8.730  < 2e-16 ***
WE                 -0.44203    0.08851  -4.994 5.96e-07 ***
ANGERS_ST_LAUD      0.65951    0.11428   5.771 8.02e-09 ***
Automne             0.25846    0.05232   4.940 7.89e-07 ***
densitePrec20      -0.33877    0.03816  -8.878  < 2e-16 ***
Matin               0.09020    0.10427   0.865 0.387005
trajets_pro         0.71876    0.09741   7.378 1.67e-13 ***
HENDAYE            -0.74727    0.12241  -6.105 1.05e-09 ***
densite20_5_origine 0.14405    0.03690   3.904 9.49e-05 ***
TOULOUSE_MATABIAU  -0.77177    0.15632  -4.937 8.00e-07 ***
LA_ROCHELLE_VILLE   0.35170    0.09714   3.621 0.000295 ***
Dep_Nuit            0.26955    0.11249   2.396 0.016580 *
dimanche            0.41597    0.10440   3.984 6.80e-05 ***
ST_MALO             0.53567    0.20554   2.606 0.009162 **
grands_departs      0.16426    0.05505   2.984 0.002852 **
trainSuivLigne      0.13674    0.04557   3.000 0.002701 **
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .

-----------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.16761    0.06983  31.043  < 2e-16 ***
trajets_pro        -0.16836    0.03722  -4.523 6.13e-06 ***
nbTraversees       -0.28754    0.02221 -12.949  < 2e-16 ***
LA_ROCHELLE_VILLE  -0.60290    0.07355  -8.197 2.63e-16 ***
Dep_Matin          -0.45146    0.04304 -10.490  < 2e-16 ***
trainPrecLigne      0.14941    0.01707   8.754  < 2e-16 ***
densite20_5_Arret  -0.03616    0.01945  -1.860    0.063 .
ST_BRIEUC          -0.96072    0.18214  -5.275 1.35e-07 ***
BREST               0.71986    0.07659   9.399  < 2e-16 ***
TARBES              0.60444    0.08254   7.323 2.52e-13 ***
TGV_D               0.36174    0.05891   6.141 8.37e-10 ***
vendredi           -0.22572    0.04217  -5.352 8.79e-08 ***
marge              -0.08084    0.01558  -5.190 2.13e-07 ***
trainSuivLigne      0.12546    0.01898   6.612 3.90e-11 ***
nbArrets           -0.14287    0.01688  -8.464  < 2e-16 ***
```

```
densite20_5_origine -0.17074    0.02532  -6.744 1.59e-11 ***
Dep_Nuit            -0.26724    0.05454  -4.900 9.68e-07 ***
ANGERS_ST_LAUD      -0.42281    0.08974  -4.712 2.47e-06 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .

-------------------------------------------------------------
No. of observations in the fit:  18276

Degrees of Freedom for the fit:  36
      Residual Deg. of Freedom:  18240
                      at cycle:  7

Global Deviance:     67871.1
            AIC:     67943.1
            SBC:     68224.38
```

TGV Départ :

```
pen =   9.8034461983559  , iterations max =   30
temps de calcul  148.565197749933   minutes
*************************************************************
Family:  c("NBItr", "right truncated Negative Binomial type I")

-------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.2755     0.1753   1.572 0.115992
densiteSuiv15   0.6632     0.1429   4.640 3.51e-06 ***
Duree           0.3096     0.1049   2.951 0.003168 **
Matin          -1.1146     0.1966  -5.670 1.45e-08 ***
ST_MALO        -1.2279     0.2602  -4.719 2.39e-06 ***
densitePrec60  -0.3022     0.1311  -2.305 0.021184 *
ST_NAZAIRE     -1.2551     0.3488  -3.598 0.000321 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .

-------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.19755    0.05650  56.596  < 2e-16 ***
densiteSuiv15 -0.25408    0.04552  -5.582 2.41e-08 ***
marge          0.18881    0.03551   5.317 1.07e-07 ***
Duree         -0.13000    0.02802  -4.639 3.52e-06 ***
grands_departs -0.17223    0.05072  -3.395 0.000687 ***
trajets_pro   -0.53681    0.09442  -5.685 1.33e-08 ***
densitePrec60  0.22605    0.05097   4.435 9.25e-06 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .

-------------------------------------------------------------
No. of observations in the fit:  18096
Degrees of Freedom for the fit:  14
      Residual Deg. of Freedom:  18082
                      at cycle:  10

Global Deviance:     22671.13
            AIC:     22699.13
            SBC:     22808.38
```

TN Arrivée :

```
pen =   10.1357890416351  , iterations max =   30
temps de calcul  180.910402135054   minutes
*************************************************************
Family:  c("NBItr", "right truncated Negative Binomial type I")

-------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.29869    0.03847  -7.764 8.51e-15 ***
densite20_5_Traversee  0.05598    0.02247   2.491 0.012749 *
temps_arret_max      -0.09002    0.02160  -4.168 3.08e-05 ***
marge                 0.02144    0.01465   1.463 0.143449
Dep_Matin             0.39611    0.03305  11.984  < 2e-16 ***
vacances             -0.22634    0.02986  -7.579 3.59e-14 ***
Duree                 0.28119    0.02856   9.846  < 2e-16 ***
temps_arret          -0.40009    0.03190 -12.540  < 2e-16 ***
RAMBOUILLET           0.35367    0.03604   9.812  < 2e-16 ***
Matin                 0.42620    0.04067  10.480  < 2e-16 ***
SEVRES_RIVE_GAUCHE   -1.09775    0.08901 -12.333  < 2e-16 ***
trainSuivLigne        0.13459    0.02042   6.592 4.41e-11 ***
densiteSuiv5          0.08907    0.01668   5.341 9.31e-08 ***
Automne               0.14929    0.02834   5.268 1.39e-07 ***
mardi                 0.12428    0.03459   3.593 0.000327 ***
WE                   -0.18694    0.04771  -3.918 8.94e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 .

-------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.17107    0.16321  19.430  < 2e-16 ***
Matin                -0.98042    0.06006 -16.323  < 2e-16 ***
SEVRES_RIVE_GAUCHE    0.41637    0.14693   2.834   0.0046 **
temps_arret           0.50226    0.05390   9.319  < 2e-16 ***
marge                -0.57094    0.03286 -17.373  < 2e-16 ***
densiteSuiv5         -0.22205    0.02936  -7.562 4.09e-14 ***
temps_arret_max       0.42161    0.03254  12.955  < 2e-16 ***
Automne               0.02802    0.06474   0.433   0.6652
densite20_5_Traversee -0.57095    0.04821 -11.843  < 2e-16 ***
nbArrets             -0.14642    0.01162 -12.601  < 2e-16 ***
RAMBOUILLET          -2.59733    0.28011  -9.273  < 2e-16 ***
temps_arret_min      -0.12040    0.02335  -5.157 2.53e-07 ***
densitePrec20         0.21054    0.03698   5.693 1.26e-08 ***
Ete                   0.47138    0.07067   6.670 2.62e-11 ***
Hiver                 0.41748    0.06662   6.266 3.75e-10 ***
nbArrets:RAMBOUILLET  0.16273    0.02017   8.066 7.58e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 .

-------------------------------------------------------------
No. of observations in the fit:  25230

Degrees of Freedom for the fit:  32
      Residual Deg. of Freedom:  25198
                      at cycle:  10

Global Deviance:     54324.07
            AIC:     54388.07
            SBC:     54648.41
```

TN Départ :

```
pen =   10.1753069757744  , iterations max =   30
temps de calcul  84.1497111320496   minutes
*************************************************************
Family:  c("NBItr", "right truncated Negative Binomial type I")

-------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.30173    0.07634 -17.051  < 2e-16 ***
densiteSuiv5    0.77785    0.05887  13.213  < 2e-16 ***
marge          -0.17595    0.03740  -4.705 2.56e-06 ***
MANTES_LA_JOLIE -0.51804    0.10415  -4.974 6.60e-07 ***
densiteSuiv15  -0.22809    0.05864  -3.889 0.000101 ***
Dep_Journee    -0.45938    0.10886  -4.220 2.45e-05 ***
grands_departs -0.25818    0.07693  -3.356 0.000792 ***
trainPrecLigne  0.17860    0.05805   3.076 0.002097 **
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .

-------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.74888    0.09204  29.865  < 2e-16 ***
```

```
densiteSuiv5  -0.39116    0.04073  -9.605  < 2e-16 ***
marge          0.41335    0.04581   9.023  < 2e-16 ***
trajets_pro   -0.76641    0.12724  -6.023 1.73e-09 ***
Dep_Nuit       0.94421    0.11940   7.908 2.71e-15 ***
densitePrec60  0.75795    0.09913   7.646 2.14e-14 ***
Dep_Journee    0.58147    0.12372   4.700 2.61e-06 ***
densitePrec20 -0.32773    0.08494  -3.858 0.000114 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .
```

```
-----------------------------------------------------------------
No. of observations in the fit:  26247
Degrees of Freedom for the fit:  16
      Residual Deg. of Freedom:  26231
                      at cycle:  7

Global Deviance:     21109.32
            AIC:     21141.32
            SBC:     21272.12
```

TER et IC Arrivée :

```
pen =   8.90191122637961  , iterations max =   30
temps de calcul  17.3123234669367   minutes
*****************************************************************
Family:  c("NBItr", "right␣truncated␣Negative␣Binomial␣type␣I")

-----------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.382724   0.159223  -2.404  0.01625 *
Matin              0.348751   0.047807   7.295 3.30e-13 ***
nbTraversees       0.041715   0.008804   4.738 2.20e-06 ***
Automne            0.304902   0.038551   7.909 2.98e-15 ***
densite20_5_Arret  0.169211   0.019810   8.542  < 2e-16 ***
temps_arret_max   -0.207189   0.021279  -9.737  < 2e-16 ***
grands_departs    -0.175795   0.037598  -4.676 2.98e-06 ***
X7200              0.411504   0.077701   5.296 1.22e-07 ***
CHARTRES          -0.194528   0.047843  -4.066 4.83e-05 ***
GRANVILLE          0.203319   0.071703   2.836  0.00459 **
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .

-----------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.98497    0.05592  17.615  < 2e-16 ***
densite20_5_Traversee -0.39670    0.04544  -8.730  < 2e-16 ***
Matin                 -0.50281    0.09497  -5.294 1.23e-07 ***
temps_arret            0.19366    0.03278   5.908 3.61e-09 ***
X7200                 -0.67208    0.13115  -5.125 3.06e-07 ***
WE                    -0.47063    0.08808  -5.343 9.40e-08 ***
trainSuivLigne         0.14713    0.03532   4.166 3.13e-05 ***
marge                 -0.07316    0.02963  -2.469 0.013577 *
Dep_Nuit              -0.29014    0.07893  -3.676 0.000239 ***
Journee               -0.28358    0.08576  -3.307 0.000949 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .

-----------------------------------------------------------------
No. of observations in the fit:  7346
Degrees of Freedom for the fit:  20
      Residual Deg. of Freedom:  7326
                      at cycle:  6

Global Deviance:     24574.26
            AIC:     24614.26
            SBC:     24752.3
```

TER et IC Départ :

```
20181031_20181101
pen =   8.91664022719884  , iterations max =   30
temps de calcul  12.3723331491152   minutes
*****************************************************************
Family:  c("NBItr", "right␣truncated␣Negative␣Binomial␣type␣I")

Call:  gamlss(formula = as.formula(paste("y", formula_mu,
sep = "␣~␣")),
    sigma.formula = as.formula(paste("", formula_sigma, sep = "␣~␣")), family = NBItr(),
    data = TERIC_Dep_tr, method = RS(30),
    mu.formula = as.formula(paste("",  formula_mu, sep = "␣~␣")), trace = F)

Fitting method: RS(30)

-----------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.63782    0.05750 -11.093  < 2e-16 ***
densitePrec60  0.38016    0.04613   8.241  < 2e-16 ***
ARGENTAN       0.43163    0.13185   3.274  0.00107 **
GRANVILLE     -0.44048    0.11102  -3.968 7.33e-05 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .

-----------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.85405    0.05346  34.681  < 2e-16 ***
densiteSuiv5 -0.46626    0.04260 -10.944  < 2e-16 ***
ARGENTAN     -1.32100    0.20476  -6.452 1.18e-10 ***
YVETOT_LE_BOTTEREAU  0.88311    0.15827   5.580 2.49e-08 ***
---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .

-----------------------------------------------------------------
No. of observations in the fit:  7455
Degrees of Freedom for the fit:  8
      Residual Deg. of Freedom:  7447
                      at cycle:  6

Global Deviance:     11363.47
            AIC:     11379.47
            SBC:     11434.8
```

Annexe D

Évaluation des prédictions

D.1 Graphes de calibration

On donne ici les graphiques de calibration des autres périodes. Les mois de juillet et février qui sont donnés dans le chapitre 4 ne sont pas rappelés. La première ligne de chaque graphe correspond dans l'ordre aux TGV, TN, TER arrivée puis TGV, TN, TER départ pour un seuil 1 minute, et la seonde ligne pour un seuil de 5 minutes.

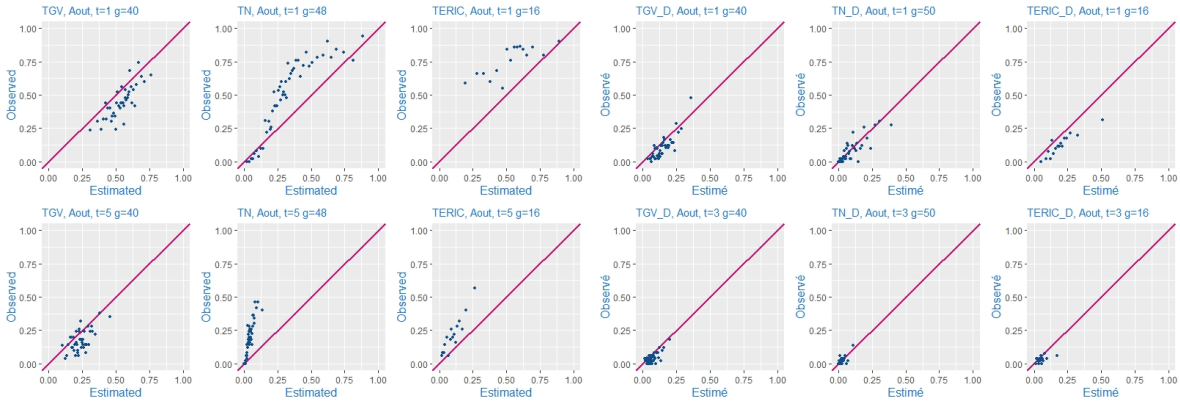

FIGURE D.1 – Graphes de calibration - août

Le graphique D.7 donne les graphes du mois d'août. Comme ce mois-ci les trains ont été bien plus perturbés qu'à la normale, les modèles n'ont pas été en mesure d'anticiper cela et on obtient des graphes très mal calibrés, avec des importantes sous-estimations du risque pour les TN et TER.

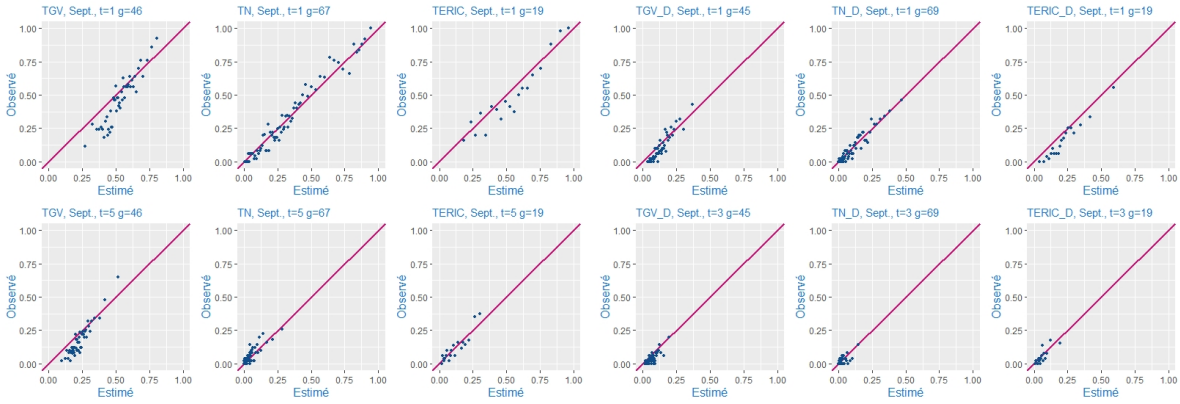

FIGURE D.2 – Graphes de calibration - septembre

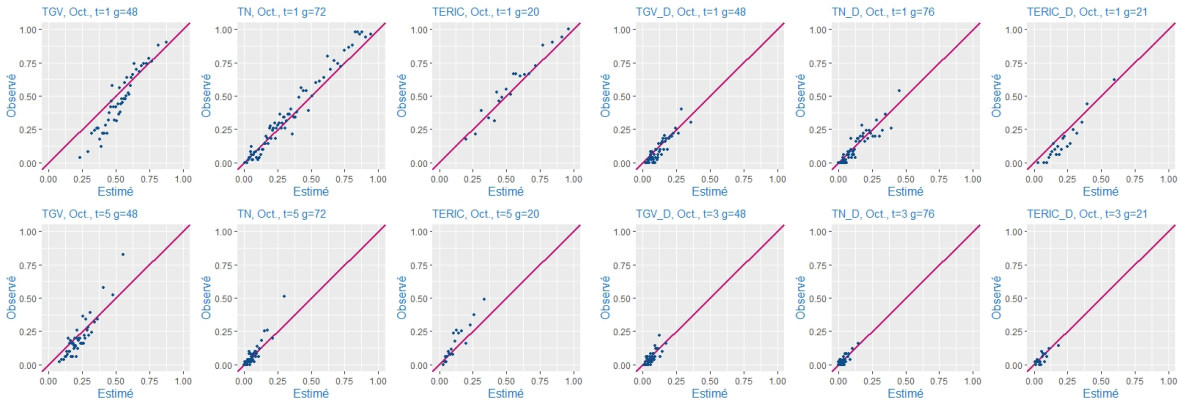

FIGURE D.3 – Graphes de calibration - octobre

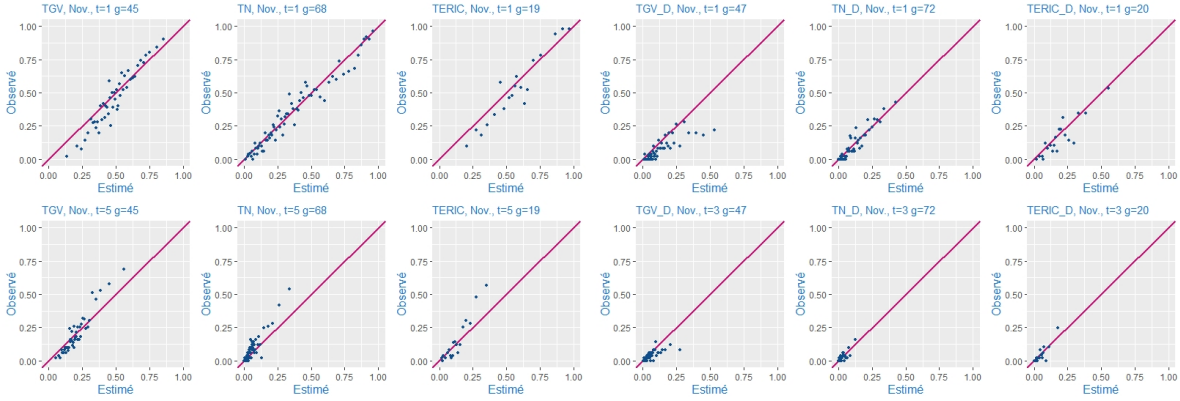

FIGURE D.4 – Graphes de calibration - novembre

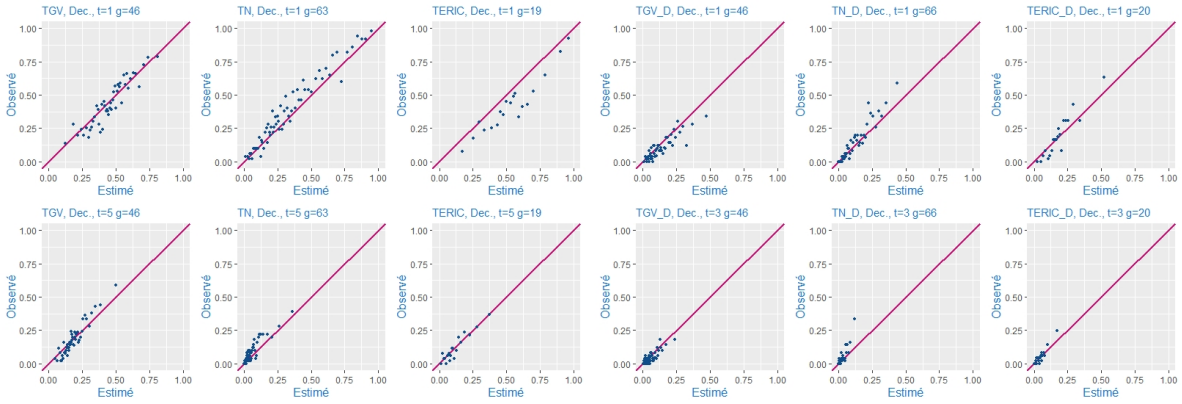

FIGURE D.5 – Graphes de calibration - décembre

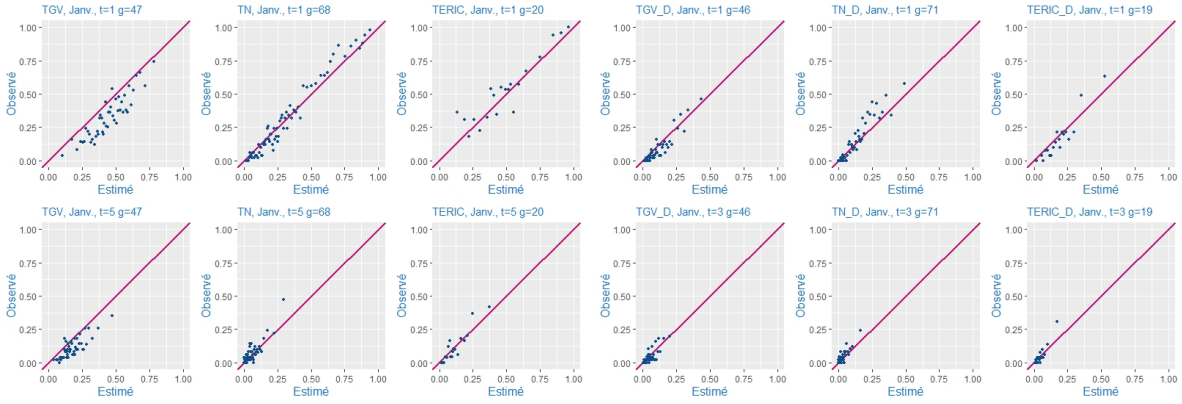

FIGURE D.6 – Graphes de calibration - janvier

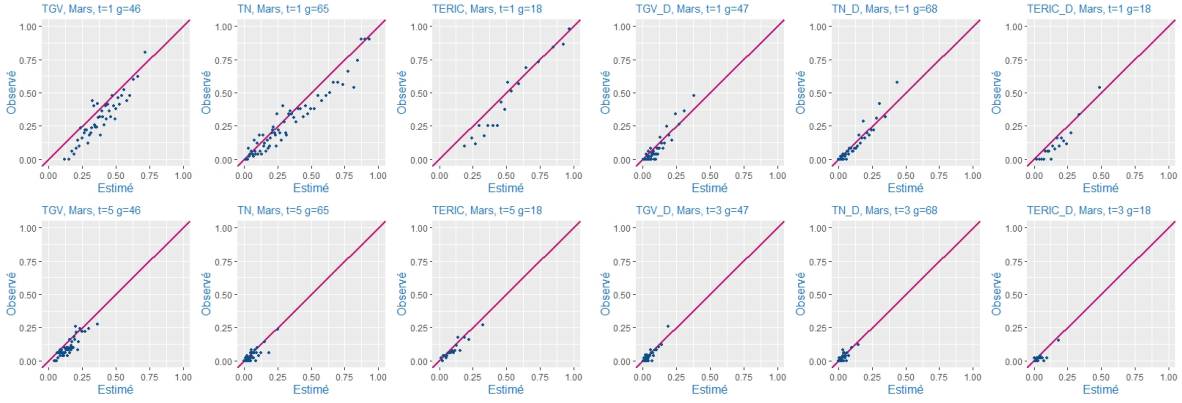

FIGURE D.7 – Graphes de calibration - mars

D.2 Résidus de quantiles randomisés

La figure D.8 représente les histogrammes des résidus obtenus par la transformation RQR (normalisation des résidus par transformation de probabilité intégrale, PIT) pour les prédictions de test du mois d'octobre. La méthodologie utilisée par plusieurs chercheurs consiste à vérifier la calibration des distributions estimées en vérifiant que cette base de résidus est normalement distribuée.

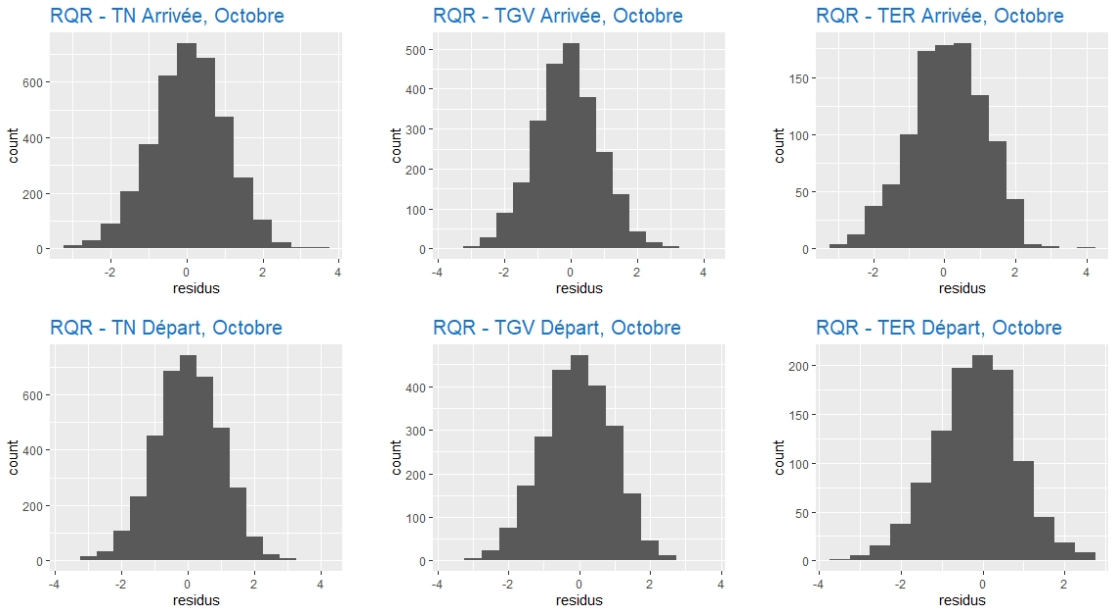

FIGURE D.8 – Visualisation des résidus RQR

Cette méthodologie a été mise de côté dans ces travaux par manque de robustesse des tests de normalité. Nous les considérons trop sensibles à la taille de l'échantillon. Le tableau D.1 montre les p valeurs obtenues en utilisant des tests de normalité classiques sur ces résidus. On constate que les différents tests ne donnent jamais des résultats concordants. Certains graphes, comme ceux des TER, montrent des déviations apparentes sur les visualisations des résidus, mais ces graphiques ne permettent pas d'interpréter l'importance de ces écarts.

Base	Taille	Shapiro	Anderson	Lillie	Kolmogorov
TN Arr	3620	0.09	0	0	0
TGV Arr	2408	0.09	0	0	0
TER Arr	1018	0.35	0	0	0.01
TN Dep	3798	0.52	1	0	0.79
TGV Dep	2397	0.56	0	0	0.25
TER Dep	1046	0.08	0	0	0

TABLEAU D.1 – Tests de normalités appliquées aux résidus de tests

Annexe E

Dimension des graphes générés au chapitre 5

Les dimensions moyennes du graphe sont données, avec en premier le nombre de sommets, le nombre d'arêtes (en milliers) et le nombre d'arêtes non nulles. Le graphe initial est complet car la solution initiale est réalisable. Les algorithmes tabou et VNS sont arrêtés après 10 minutes de calculs et la taille du graphe VNS est moyennée sur 10 itérations.

id	Initial			Glouton			Tabu			VNS		
	noeuds	arêtes	> 0	noeuds	arêtes	> 0	noeuds	arêtes	> 0	noeuds	arêtes	> 0
0307	388	75k	543	454	103k	857	1206	716k	12876	905	407k	4687
1407	264	35k	231	288	41k	293	1356	895k	14944	864	370k	4158
2507	336	56k	355	387	75k	518	1287	809k	8731	854	363k	3376
2208	342	58k	383	384	73k	521	1299	829k	10007	860	368k	3415
2608	259	33k	231	272	37k	268	1277	797k	14014	812	327k	3760
0409	387	75k	555	462	106k	885	1153	657k	12519	894	398k	4791
1409	391	76k	581	455	103k	897	1117	617k	11430	879	385k	4667
2209	241	29k	155	269	36k	200	1215	718k	12212	807	323k	3025
0410	381	72k	551	454	103k	892	1096	584k	12750	897	401k	5079
0810	401	80k	568	468	109k	879	1065	549k	11563	905	408k	4950
1810	385	74k	555	455	103k	907	1147	642k	12224	899	402k	5036
2910	393	77k	560	453	102k	853	1008	491k	15529	909	411k	4804
0911	395	78k	554	456	104k	823	1041	532k	12957	891	396k	4466
1611	398	79k	598	456	104k	918	1064	556k	14442	854	363k	4447
2511	247	30k	251	274	37k	322	1332	868k	15844	816	330k	4675
0512	374	70k	546	443	98k	878	1182	685k	15016	882	387k	4754
0912	242	29k	241	259	33k	291	1233	745k	12970	770	294k	4012
1812	380	72k	511	444	98k	786	1143	637k	13677	872	378k	4609
2712	341	58k	414	392	77k	563	1141	628k	11850	848	357k	3852
0501	245	30k	218	270	36k	273	1283	804k	12481	800	317k	3808
1501	379	72k	515	443	98k	806	1205	714k	11913	906	409k	5120
2401	381	72k	557	453	102k	886	1163	663k	15147	923	424k	5240
0602	385	74k	516	460	105k	815	1155	651k	15532	886	390k	4861
1102	385	74k	515	459	105k	795	1108	599k	17763	920	421k	5437
2402	249	31k	213	275	38k	272	1227	739k	13484	805	321k	4441
0803	389	75k	578	445	99k	843	1082	563k	14832	864	372k	4655
1803	383	73k	528	451	101k	821	1107	594k	17173	887	391k	4970

Tableau E.1 – Taille des graphes finaux

Bibliographie

[1] Categorical features and encoding in decision trees. https://medium.com/data-design/visiting-categorical-features-and-encoding-in-decision-trees-53400fa65931. Acces : 10/10/2019. 100

[2] Code 406 : Capacity. Technical report, UIC,International Union of Railways, 2013. 32

[3] Règles d'exploitation de la gestion opérationnelle des circulations. Technical report, RFF, Novembre 2013. 29, 30

[4] Directive de justification des retards dans Brehat. Technical report, SNCF Réseau, Décembre 2016. 41, 80, 112

[5] A la reconquête de la robustesse des services ferroviaires. Technical report, Comité international d'experts ferroviaires, Juillet 2017. 21, 126

[6] Antonin 3 : le modèle multimodal d'Ile-de-France Mobilités. Technical report, Ile-de-France Mobilités, Juillet 2017. 42

[7] Le guide de la production ferroviaire. Technical report, SNCF - Direction Générale de la communication et de l'image, Janvier 2018. 28

[8] Le suivi de la qualité de service et des droits des passagers dans le transport ferroviaire de voyageurs en Europe et en France. Technical report, ARAFER, Novembre 2018. 31

[9] Rapport annuel 2017. Technical report, AQST, Décembre 2018. 28, 31

[10] Normes de tracé horaire en gare pour le SA 2021. Technical report, SNCV Réseau, Juin 2019. 125, 126

[11] M Abril, F Barber, L Ingolotti, MA Salido, P Tormos, and A Lova. An assessment of railway capacity. *Transportation Research Part E : Logistics and Transportation Review*, 44(5) :774–806, 2008. 31

[12] Alan Agresti. *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015. 57

[13] Ravindra K Ahuja, Özlem Ergun, James B Orlin, and Abraham P Punnen. A survey of very large-scale neighborhood search techniques. *Discrete Applied Mathematics*, 123(1-3) :75–102, 2002. 136

[14] M Faris Muslim Al-Athari. Estimation of the mean of truncated exponential distribution. *Journal of Mathematics and Statistics*, 4(4) :284, 2008. 58

[15] Emma V Andersson, Anders Peterson, and Johanna Törnquist Krasemann. Quantifying railway timetable robustness in critical points. *Journal of Rail Transport Planning & Management*, 3(3) :95–110, 2013. 32

[16] Diogo V Andrade, Mauricio GC Resende, and Renato F Werneck. Fast local search for the maximum independent set problem. *Journal of Heuristics*, 18(4) :525–547, 2012. 135

[17] John E Angus. The probability integral transform and related results. *SIAM review*, 36(4) :652–654, 1994. 71

[18] Adam C Bartley. *Evaluating goodness-of-fit for a logistic regression model using the Hosmer-Lemeshow test on samples from a large data set*. PhD thesis, The Ohio State University, 2014. 71

[19] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2) :235–292, 2018. 133

[20] Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 2019. 133, 138, 161, 170

[21] Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations research*, 52(1) :35–53, 2004. 31, 34, 132

[22] Nikola Bešinović and Rob MP Goverde. Stable and robust train routing in station areas with balanced infrastructure capacity occupation. *Public Transport*, pages 1–26, 2019. 31, 35, 38, 39, 48

[23] J Eric Bickel. Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis*, 4(2) :49–65, 2007. 68

[24] Alain Billionnet. Using integer programming to solve the train-platforming problem. *Transportation Science*, 37(2) :213–222, 2003. 35, 37

[25] John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011. 132

[26] Immanuel M Bomze, Marco Budinich, Panos M Pardalos, and Marcello Pelillo. The maximum clique problem. In *Handbook of combinatorial optimization*, pages 1–74. Springer, 1999. 134, 135

[27] Henrik Bostrom. Estimating class probabilities in random forests. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 211–216. IEEE, 2007. 61

[28] Anne-Laure Boulesteix and Matthias Schmid. Machine learning versus statistical modeling. *Biometrical Journal*, 56(4) :588–593, 2014. 63, 64

[29] William Brazil, Arthur White, Maria Nogal, Brian Caulfield, Alan O'Connor, and Craig Morton. Weather and rail delays : Analysis of metropolitan rail in dublin. *Journal of Transport Geography*, 59 :69–76, 2017. 45

[30] Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001. 60

[31] Leo Breiman. Statistical modeling : The two cultures. *Statistical Science*, 16(3) :199–215, 2001. 43, 54, 55

[32] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, 78(1) :1–3, 1950. 66

[33] Jochen Bröcker and Leonard A Smith. Scoring probabilistic forecasts : The importance of being proper. *Weather and Forecasting*, 22(2) :382–388, 2007. 65, 68

[34] Thorsten Büker and Bernhard Seybold. Stochastic modelling of delay propagation in large networks. *Journal of Rail Transport Planning & Management*, 2(1-2) :34–50, 2012. 42

[35] Robert Burdett and Erhan Kozan. Determining operations affected by delay in predictive train timetables. *Computers & Operations Research*, 41 :150–166, 2014. 32

[36] Sofie Burggraeve, Thijs Dewilde, Peter Sels, and Pieter Vansteenwegen. Improving passenger robustness by taking passenger numbers and recurring delays explicitly into account on the tactical level. In *Proceedings of 6th International Seminar on Railway Operations Modelling and Analysis (IAROR) : RailTokyo2015*, pages 1–15, 2015. 28, 39

[37] Valentina Cacchiani, Dennis Huisman, Martin Kidd, Leo Kroon, Paolo Toth, Lucas Veelenturf, and Joris Wagenaar. An overview of recovery models and algorithms for real-time railway rescheduling. *Transportation Research Part B : Methodological*, 63 :15–37, 2014. 30

[38] Shaowei Cai and Jinkun Lin. Fast solving maximum weight clique problem in massive graphs. In *IJCAI*, pages 568–574, 2016. 135, 139

[39] Guillem Candille and Olivier Talagrand. Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society : A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 131(609) :2131–2150, 2005. 67

[40] Alberto Caprara, Matteo Fischetti, and Paolo Toth. Modeling and solving the train timetabling problem. *Operations research*, 50(5) :851–861, 2002. 27

[41] Alberto Caprara, Laura Galli, Leo Kroon, Gábor Maróti, and Paolo Toth. Robust train routing and online re-scheduling. In *OASIcs-OpenAccess Series in Informatics*, volume 14. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2010. 38, 39

[42] Alberto Caprara, Laura Galli, Sebastian Stiller, and Paolo Toth. Recoverable robustness for scheduling with precedence constraints. In *International Network Optimization Conference 2009 (INOC 2009)*, pages 1–7. Università di Pisa, 2009. 39, 162

[43] Alberto Caprara, Laura Galli, and Paolo Toth. Solution of the train platforming problem. *Transportation Science*, 45(2) :246–257, 2011. 35, 36, 39

[44] Jaime S Cardoso and Joaquim F Costa. Learning to classify ordinal data : The data replication method. *Journal of Machine Learning Research*, 8(Jul) :1393–1429, 2007. 62

[45] Jaime S Cardoso and Ricardo Sousa. Measuring the performance of ordinal classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(08) :1173–1195, 2011. 74

[46] Malachy Carey. Ex ante heuristic measures of schedule reliability. *Transportation Research Part B : Methodological*, 33(7) :473–494, 1999. 38, 141

[47] Malachy Carey and Sinead Carville. Testing schedule performance and reliability for train stations. *Journal of the Operational Research Society*, 51(6) :666–682, 2000. 38

[48] Malachy Carey and Sinead Carville. Scheduling and platforming trains at busy complex stations. *Transportation Research Part A : Policy and Practice*, 37(3) :195–224, 2003. 35, 37

[49] Fabrizio Cerreto, Bo Friis Nielsen, Otto Anker Nielsen, and Steven S Harrod. Application of data clustering to railway delay pattern recognition. *Journal of Advanced Transportation*, 2018, 2018. 46

[50] Maguelonne Chandesris and Xavier Chapuis. Non-parametric appoach for real-time prediction. In *Proceedings of TransitData, Brisbane, Australia, July, 2018*, 2018. 47

[51] X Chapuis. Arrival time prediction using neural networks. In *Proceedings of RailLille2017 : 7th International Conference on Railway Operations Modelling and Analysis, Lille, France. International Association of Railway Operations Research (IAROR)*, 2017. 47

[52] M Clarke, Chris J Hinde, Mark S Withall, Thomas W Jackson, Iain W Phillips, Steve Brown, and Robert Watson. Allocating railway platforms using a genetic algorithm. In *Research and Development in Intelligent Systems XXVI*, pages 421–434. Springer, 2010. 37

[53] A Clifford Cohen. *Truncated and censored samples : theory and applications*. CRC press, 1991. 58

[54] Francesco Corman and Pavle Kecman. Stochastic prediction of train delays in real-time using bayesian networks. *Transportation Research Part C : Emerging Technologies*, 95 :599–615, 2018. 48

[55] Francesco Corman and Lingyun Meng. A review of online dynamic models and algorithms for railway traffic management. *IEEE Transactions on Intelligent Transportation Systems*, 16(3) :1274–1284, 2014. 29, 30

[56] Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. Random forest versus logistic regression : a large-scale benchmark experiment. *BMC bioinformatics*, 19(1) :270, 2018. 63, 64

[57] Yong Cui, Ullrich Martin, and Weiting Zhao. Calibration of disturbance parameters in railway operational simulation based on reinforcement learning. *Journal of Rail Transport Planning & Management*, 6(1) :1–12, 2016. 42, 44

[58] Boris Cule, Bart Goethals, Sven Tassenoy, and Sabine Verboven. Mining train delays. In *International Symposium on Intelligent Data Analysis*, pages 113–124. Springer, 2011. 46

[59] Claudia Czado, Tilmann Gneiting, and Leonhard Held. Predictive model assessment for count data. *Biometrics*, 65(4) :1254–1261, 2009. 68, 72

[60] Antonio D'Agostino. Big Data in Railways. Technical report, European Union Agency for Railways, Octobre 2016. 39, 40

[61] RB d'Agostino and Byung-Ho Nam. Evaluation of the performance of survival analysis models : discrimination and calibration measures. *Handbook of statistics*, 23 :1–25, 2003. 71, 74

[62] Marie Milliet de Faverges, Christophe Picouleau, Giorgio Russolillo, Boubekeur Merabet, and Bertrand Houzel. Impact of calibration of perturbations in simulation : the case of robustness evaluation at station. In *RailNorrköping 2019. 8th International Conference on Railway Operations Modelling and Analysis (ICROMA), Norrköping, Sweden, June 17th–20th, 2019*, number 069, pages 320–339. Linköping University Electronic Press, 2019. 162

[63] Marie Milliet de Faverges, G. Russolillo, C. Picouleau, B. Merabet, and B. Houzel. Estimating long-term delay risk with generalized linear models. In *2018 21st IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 2911–2916, 9 2018. 88, 99, 115

[64] Marie Milliet de Faverges, G. Russolillo, C. Picouleau, B. Merabet, and B. Houzel. Modelling passenger train arrival delays with generalized linear models and its perspective for scheduling at main stations. *IET Conference Proceedings*, 1 2018. 88

[65] Xavier Delorme, Xavier Gandibleux, and Joaquin Rodriguez. Grasp for set packing problems. *European Journal of Operational Research*, 153(3) :564–580, 2004. 35, 36, 37

[66] Olga V Demler, Nina P Paynter, and Nancy R Cook. Tests of calibration and goodness-of-fit in the survival setting. *Statistics in medicine*, 34(10) :1659–1680, 2015. 71

[67] Thijs Dewilde, Peter Sels, Dirk Cattrysse, and Pieter Vansteenwegen. Robust railway station planning : An interaction between routing, timetabling and platforming. *Journal of Rail Transport Planning & Management*, 3(3) :68–77, 2013. 28

[68] Thijs Dewilde, Peter Sels, Dirk Cattrysse, and Pieter Vansteenwegen. Improving the robustness in railway station areas. *European Journal of Operational Research*, 235(1) :276–286, 2014. 35, 36, 38, 41, 139

[69] Francis X Diebold, Todd A Gunther, and Anthony S Tay. Evaluating density forecasts. *International Economic Review*, 39 :863–883, 1997. 71

[70] Peter K Dunn and Gordon K Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3) :236–244, 1996. 72

[71] Matteo Fischetti and Michele Monaci. Light robustness. In *Robust and online large-scale optimization*, pages 61–84. Springer, 2009. 34, 132

[72] Eibe Frank and Remco R Bouckaert. Conditional density estimation with class probability estimators. In *Asian Conference on Machine Learning*, pages 65–81. Springer, 2009. 63

[73] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *European Conference on Machine Learning*, pages 145–156. Springer, 2001. 62

[74] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA :, 2001. 59, 60, 62, 64, 66, 101

[75] Aurélien Froger, Olivier Guyon, and Eric Pinson. A set packing approach for scheduling passenger train drivers : the french experience. In *RailTokyo2015*, 2015. 27

[76] Virginie Gabrel, Cécile Murat, and Aurélie Thiele. Recent advances in robust optimization : An overview. *European journal of operational research*, 235(3) :471–483, 2014. 132

[77] Robin Genuer and Jean-Michel Poggi. Arbres cart et forêts aléatoires, importance et sélection de variables. 2017. 60, 64

[78] Faeze Ghofrani, Qing He, Rob MP Goverde, and Xiang Liu. Recent applications of big data analytics in railway transportation systems : A survey. *Transportation Research Part C : Emerging Technologies*, 90 :226–246, 2018. 40

[79] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477) :359–378, 2007. 65, 67

[80] Mithat Gönen and Glenn Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4) :965–970, 2005. 74

[81] Rob MP Goverde. Punctuality of railway operations and timetable stability analysis. *PhD Thesis, Delft University of Technology.*, 2005. 28, 44, 48

[82] Rob MP Goverde. A delay propagation algorithm for large-scale railway traffic networks. *Transportation Research Part C : Emerging Technologies*, 18(3) :269–287, 2010. 48

[83] Rob MP Goverde, IA Hansen, G Hooghiemstra, and HP Lopuhaa. Delay distributions in railway stations. In *9th World Conference on Transport Research, Seoul, Korea, July 22-27, 2001*. WCTRS, 2001. 45

[84] Rob MP Goverde and Ingo A Hansen. Performance indicators for railway timetables. In *Intelligent Rail Transportation (ICIRT), 2013 IEEE International Conference on*, pages 301–306. IEEE, 2013. 30, 31

[85] Jon Ketil Grønnesby and Ørnulf Borgan. A method for checking regression models in survival analysis based on the risk score. *Lifetime data analysis*, 2(4) :315–328, 1996. 71

[86] Pedro Antonio Gutierrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. Ordinal regression methods : survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1) :127–146, 2015. 62

[87] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar) :1157–1182, 2003. 96

[88] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008. 154

[89] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1) :29–36, 1982. 74

[90] Ingo A Hansen, Rob MP Goverde, and Dirk J van der Meer. Online train delay recognition and running time prediction. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 1783–1788. IEEE, 2010. 48

[91] Pierre Hansen and Nenad Mladenović. Variable neighborhood search. In *Handbook of metaheuristics*, pages 145–184. Springer, 2003. 150

[92] Pierre Hansen, Nenad Mladenović, and Dragan Urošević. Variable neighborhood search for the maximum clique. *Discrete Applied Mathematics*, 145(1) :117–125, 2004. 135, 150

[93] Frank E Harrell Jr. *Regression modeling strategies : with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015. 65, 68, 71

[94] Steven Harrod, Georgios Pournaras, and Bo Friis Nielsen. Distribution fitting for very large railway delay data sets with discrete values. In *Trafikdage 2018*, 2018. 44

[95] Dennis R Helsel. Fabricating data : how substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere*, 65(11) :2434–2439, 2006. 57

[96] Joseph M. Hilbe. *Modeling Count Data*, pages 836–839. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. 56

[97] Joseph M Hilbe. *Negative binomial regression*. Cambridge University Press, 2011. 95

[98] David W Hosmer and Stanley Lemesbow. Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9(10) :1043–1069, 1980. 70

[99] David W Hosmer, Stanley Lemeshow, and J Klar. Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small. *Biometrical Journal*, 30(8) :911–924, 1988. 71

[100] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013. 70

[101] Torsten Hothorn and Achim Zeileis. partykit : A modular toolkit for recursive partytioning in r. *The Journal of Machine Learning Research*, 16(1) :3905–3909, 2015. 89

[102] Bertrand Houzel. Opengov : la régularité à bout de cliques... *Revue Générale des Chemins de Fer*, Juillet-Août :34–46, 2016. 125

[103] Dennis Huisman, Leo G Kroon, Ramon M Lentink, and Michiel JCM Vromans. Operations research in passenger railway transportation. *Statistica Neerlandica*, 59(4) :467–497, 2005. 26, 27, 35

[104] Lars Wittrup Jensen, Alex Landex, and Otto Anker Nielsen. Evaluation of robustness indicators using railway operation simulation. *Computers in Railways XIV : Railway Engineering Design and Optimization*, 135 :329, 2014. 32

[105] Predrag Jovanović, Pavle Kecman, Nebojša Bojović, and Dragomir Mandić. Optimal allocation of buffer times to increase train schedule robustness. *European Journal of Operational Research*, 256(1) :44–54, 2017. 33

[106] Vasileios Kassiano, Anastasios Gounaris, Apostolos N Papadopoulos, and Kostas Tsichlas. Mining uncertain graphs : An overview. In *International Workshop of Algorithmic Aspects of Cloud Computing*, pages 87–116. Springer, 2016. 134

[107] Kengo Katayama, Akihiro Hamamoto, and Hiroyuki Narihisa. An effective local search for the maximum clique problem. *Information Processing Letters*, 95(5) :503–511, 2005. 135, 136

[108] Pavle Kecman, Francesco Corman, Andrea D'Ariano, and Rob MP Goverde. Rescheduling models for railway traffic management in large-scale networks. *Public Transport*, 5(1-2) :95–123, 2013. 30

[109] Pavle Kecman and Rob MP Goverde. Predictive modelling of running and dwell times in railway traffic. *Public Transport*, 7(3) :295–319, 2015. 47

[110] Harshad Khadilkar. Data-enabled stochastic modeling for evaluating schedule robustness of railway networks. *Transportation Science*, 51(4) :1161–1176, 2016. 32, 160

[111] Haris Koutsopoulos and Zhigao Wang. Simulation of urban rail operations : application framework. *Transportation Research Record : Journal of the Transportation Research Board*, (2006) :84–91, 2007. 42

[112] Leo Kroon, Gábor Maróti, Mathijn Retel Helmrich, Michiel Vromans, and Rommert Dekker. Stochastic improvement of cyclic railway timetables. *Transportation Research Part B : Methodological*, 42(6) :553–570, 2008. 34

[113] Leo G Kroon, H Edwin Romeijn, and Peter J Zwaneveld. Routing trains through railway stations : complexity issues. *European Journal of Operational Research*, 98(3) :485–498, 1997. 34

[114] Jochen Kruppa, Yufeng Liu, Gérard Biau, Michael Kohler, Inke R König, James D Malley, and Andreas Ziegler. Probability estimation with machine learning methods for dichotomous and multicategory outcome : Theory. *Biometrical Journal*, 56(4) :534–563, 2014. 61, 63

[115] Jochen Kruppa, Yufeng Liu, Hans-Christian Diener, Theresa Holste, Christian Weimar, Inke R König, and Andreas Ziegler. Probability estimation with machine learning methods for dichotomous and multicategory outcome : Applications. *Biometrical Journal*, 56(4) :564–583, 2014. 61

[116] Alex Landex and Lars Wittrup Jensen. Measures for track complexity and robustness of operation at stations. *Journal of Rail Transport Planning & Management*, 3(1-2) :22–35, 2013. 38

[117] Alex Landex, Anders H Kaas, and Otto Anker Nielsen. *Methods to estimate railway capacity and passenger delays*. Technical University of Denmark (DTU), 2008. 31

[118] Rune Larsen, Marco Pranzo, Andrea D'Ariano, Francesco Corman, and Dario Pacciarelli. Susceptibility of optimal train schedules to stochastic disturbances of process times. *Flexible Services and Manufacturing Journal*, 26(4) :466–489, 2014. 33, 42

[119] Katerina Lepenioti, Alexandros Bousdekis, Dimitris Apostolou, and Gregoris Mentzas. Prescriptive analytics : Literature review and research challenges. *International Journal of Information Management*, 50 :57–70, 2020. 133

[120] Christian Liebchen, Marco Lübbecke, Rolf Möhring, and Sebastian Stiller. The concept of recoverable robustness, linear programming recovery, and railway applications. In *Robust and online large-scale optimization*, pages 1–27. Springer, 2009. 33, 39, 133, 162

[121] Christian Liebchen, Michael Schachtebeck, Anita Schöbel, Sebastian Stiller, and André Prigge. Computing delay resistant railway timetables. *Computers & Operations Research*, 37(5) :857–868, 2010. 33

[122] James K Lindsey. *Applying generalized linear models*. Springer Science & Business Media, 2000. 58

[123] Richard Lusby, Jesper Larsen, David Ryan, and Matthias Ehrgott. Routing trains through railway junctions : a new set-packing approach. *Transportation Science*, 45(2) :228–245, 2011. 35, 36, 37

[124] Richard M Lusby, Jesper Larsen, and Simon Bull. A survey on robustness in railway planning. *European Journal of Operational Research*, 266(1) :1–15, 2018. 32

[125] Elder Magalhães Macambira and Cid Carvalho De Souza. The edge-weighted clique problem : valid inequalities, facets and polyhedral computations. *European Journal of Operational Research*, 123(2) :346–371, 2000. 134

[126] Reason L Machete. Contrasting probabilistic scoring rules. *Journal of Statistical Planning and Inference*, 143(10) :1781–1790, 2013. 68

[127] JD Malley, J Kruppa, A Dasgupta, KG Malley, and A Ziegler. Probability machines : Consistent probability estimation using nonparametric learning machines. *Methods Inf Med*, 51(1) :74–81, 2012. 54, 61

[128] Alexandre Mani. Prédiction du risque de retards pour l'exploitation des grandes gares. 2019. 116

[129] Dragos D Margineantu. Class probability estimation and cost-sensitive classification decisions. In *European Conference on Machine Learning*, pages 270–281. Springer, 2002. 60, 61

[130] Nikola Marković, Sanjin Milinković, Konstantin S Tikhonov, and Paul Schonfeld. Analyzing passenger train arrival delays with support vector regression. *Transportation Research Part C : Emerging Technologies*, 56 :251–262, 2015. 46

[131] Lars-Göran Mattsson. Railway capacity and train delay relationships. In *Critical Infrastructure*, pages 129–150. Springer, 2007. 31

[132] Susanne May and David W Hosmer. A simplified method of calculating an overall goodness-of-fit test for the cox proportional hazards model. *Lifetime data analysis*, 4(2) :109–120, 1998. 71

[133] Edgar C Merkle and Mark Steyvers. Choosing a strictly proper scoring rule. *Decision Analysis*, 10(4) :292–304, 2013. 65, 68

[134] Zhuqi Miao, Balabhaskar Balasundaram, and Eduardo L Pasiliao. An exact algorithm for the maximum probabilistic clique problem. *Journal of Combinatorial Optimization*, 28(1) :105–120, 2014. 134

[135] Yongyi Min and Alan Agresti. Modeling nonnegative data with clumping at zero : a survey. *Journal of the Iranian Statistical Society*, 1(1) :7–33, 2002. 57

[136] Allan H Murphy. Scalar and vector partitions of the probability score : Part i. two-state situation. *Journal of Applied Meteorology*, 11(2) :273–282, 1972. 67

[137] Allan H Murphy. Scalar and vector partitions of the ranked probability score. *Mon Weather Rev*, 100(10) :701–708, 1972. 67

[138] Rahul Nair, Thanh Lam Hoang, Marco Laumanns, Bei Chen, Randall Cogill, Jácint Szabó, and Thomas Walter. An ensemble prediction model for train delays. *Transportation Research Part C : Emerging Technologies*, 104 :196–209, 2019. 47

[139] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society : Series A (General)*, 135(3) :370–384, 1972. 55

[140] Alexandru Niculescu-Mizil and Rich Caruana. Obtaining calibrated probabilities from boosting. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 413–420, 2005. 62

[141] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632. ACM, 2005. 73, 74

[142] Luca Oneto, Emanuele Fumeo, Giorgio Clerico, Renzo Canepa, Federico Papa, Carlo Dambra, Nadia Mazzino, and Davide Anguita. Advanced analytics for train delay prediction systems by including exogenous weather data. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pages 458–467. IEEE, 2016. 47

[143] Luca Oneto, Emanuele Fumeo, Giorgio Clerico, Renzo Canepa, Federico Papa, Carlo Dambra, Nadia Mazzino, and Davide Anguita. Train delay prediction systems : A big data analytics perspective. *Big Data Research*, 2017. 47

[144] Carl-William Palmqvist, Nils Olsson, and Lena Hiselius. Some influencing factors for passenger train punctuality in sweden. In *Annual Conference of the Prognostics and Health Management Society 2017*, volume 8. PHM Society, 2017. 45

[145] Carl-William Palmqvist, NOE Olsson, and Lena Hiselius. Delays for passenger trains on a regional railway line in southern sweden. *International Journal of Transport Development and Integration*, 1(3) :421–431, 2017. 45

[146] Prabasaj Paul, Michael L Pennell, and Stanley Lemeshow. Standardizing the power of the hosmer–lemeshow goodness of fit test in large data sets. *Statistics in medicine*, 32(1) :67–80, 2013. 70, 71

[147] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011. 61, 62, 66, 74, 99, 100

[148] Jan Peters, Bastian Emig, Marten Jung, and Stefan Schmidt. Prediction of delays in public transportation using neural networks. In *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, volume 2, pages 92–97. IEEE, 2005. 47

[149] David Pisinger and Stefan Ropke. Large neighborhood search. In *Handbook of metaheuristics*, pages 399–419. Springer, 2010. 136

[150] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3) :61–74, 1999. 62

[151] Suporn Pongnumkul, Thanakij Pechprasarn, Narin Kunaseth, and Kornchawal Chaipah. Improving arrival time prediction of thailand's passenger trains using historical travel times. In *Computer Science and Software Engineering (JCSSE), 2014 11th International Joint Conference on*, pages 307–312. IEEE, 2014. 47

[152] Foster Provost and Pedro Domingos. Tree induction for probability-based ranking. *Machine learning*, 52(3) :199–215, 2003. 60, 61, 64

[153] Yuanqi Qin, Xiaoliang Ma, and Sida Jiang. A stochastic process approach for modeling arrival delay in train operations. In *Transportation Research Board 86th Annual Meeting*, 2017. 45

[154] B Rigby, M Stasinopoulos, and C Akantziliotou. Instructions on how to use the gamlss package in r. *Computational statistics and Data analysis*, 2 :194–195, 2008. 56

[155] RA Rigby, DM Stasinopoulos, GZ Heller, and Fernanda De Bastiani. Distributions for modelling location, scale, and shape : Using gamlss in r. *URL www. gamlss. org.(last accessed 5 March 2018)*, 2017. 56, 87, 95

[156] Robert A Rigby and D Mikis Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 54(3) :507–554, 2005. 56

[157] Miguel A Salido, Federico Barber, and Laura Ingolotti. Robustness in railway transportation scheduling. In *Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on*, pages 2880–2885. IEEE, 2008. 32

[158] Anita Schöbel. Line planning in public transportation : models and methods. *OR spectrum*, 34(3) :491–510, 2012. 27

[159] Peter Sels, Dirk Cattrysse, and Pieter Vansteenwegen. Automated platforming & routing of trains in all belgian railway stations. *Expert Systems with Applications*, 62 :302–316, 2016. 35

[160] Peter Sels, Thijs Dewilde, Dirk Cattrysse, and Pieter Vansteenwegen. Expected passenger travel time for train schedule evaluation and optimization. In *Proceedings of the 5th international seminar on railway operations modelling and analysis, Copenhagen, Denmark*, 2013. 31, 41

[161] Peter Sels, Pieter Vansteenwegen, Thijs Dewilde, Dirk Cattrysse, Bertrand Waquet, and Antoine Joubert. The train platforming problem : The infrastructure management company perspective. *Transportation Research Part B : Methodological*, 61 :55–72, 2014. 35

[162] Galit Shmueli. To explain or to predict ? *Statistical Science*, 25(3) :289–310, 2010. 43

[163] Mikis D Stasinopoulos, Robert A Rigby, Gillian Z Heller, Vlasios Voudouris, and Fernanda De Bastiani. *Flexible regression and smoothing : using GAMLSS in R*. Chapman and Hall/CRC, 2017. 56, 66, 87

[164] Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models : a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1) :128, 2010. 54, 65

[165] Yoko Takeuchi, Norio Tomii, and Chikara Hirai. Evaluation method of robustness for train schedules. *Quarterly Report of RTRI*, 48(4) :197–201, 2007. 31, 33

[166] Theja Tulabandhula and Cynthia Rudin. Robust optimization using machine learning for uncertainty sets. *ECML/PKDD 2014*, page 121, 2014. 133

[167] Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, and Ewout W Steyerberg. A calibration hierarchy for risk models was defined : from utopia to empirical data. *Journal of clinical epidemiology*, 74 :167–176, 2016. 68

[168] Niels van Oort, Daniel Sparing, Ties Brands, and Rob MP Goverde. Data driven improvements in public transport : the dutch example. *Public transport*, 7(3) :369–389, 2015. 40, 43

[169] Pieter Vansteenwegen and Dirk Van Oudheusden. Developing railway timetables which guarantee a better service. *European Journal of Operational Research*, 173(1) :337–350, 2006. 41

[170] Rafael Velasquez, Matthias Ehrgott, David Ryan, and Anita Schöbel. A set-packing approach to routing trains through railway stations. In *40th annual conference of the operations research society of New Zealand*, pages 305–314, 2005. 36, 37

[171] Michiel Vromans. *Reliability of Railway Systems*. PhD thesis, Erasmus Research Institute of Management, 2005. 28, 31, 33, 41

[172] Pu Wang and Qing-peng Zhang. Train delay analysis and prediction based on big data fusion. *Transportation Safety and Environment*, 2019. 47

[173] Ren Wang and Daniel B Work. Data driven approaches for passenger train delay estimation. In *IEEE 18th International Conference on Intelligent Transportation Systems (ITSC), 2015*, pages 535–540. IEEE, 2015. 47

[174] Yiyuan Wang, Shaowei Cai, and Minghao Yin. Two efficient local search algorithms for maximum weight clique problem. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. 135

[175] Chao Wen, Zhongcan Li, Javad Lessan, Liping Fu, Ping Huang, and Chaozhe Jiang. Statistical investigation on train primary delay based on real records : evidence from wuhan–guangzhou hsr. *International Journal of Rail Transportation*, 5(3) :170–189, 2017. 28, 44

[176] Nigel HM Wilson, Jinhua Zhao, and Adam Rahbee. The potential impact of automated data collection systems on urban public transport planning. In *Schedule-based modeling of transportation networks*, pages 75–97. Springer, 2009. 40, 41

[177] Qinghua Wu and Jin-Kao Hao. A review on algorithms for maximum clique problems. *European Journal of Operational Research*, 242(3) :693–709, 2015. 134, 135

[178] Hideyuki Yabuki, Taku Ageishi, and Norio Tomii. Mining the cause of delays in urban railways based on association rules. In *Proc. 13th Conf. Advanced Systems in Public Transport*, 2015. 46

[179] Masoud Yaghini, Mohammad M Khoshraftar, and Masoud Seyedabadi. Railway passenger train delay prediction via neural network model. *Journal of advanced transportation*, 47(3) :355–368, 2013. 46

[180] Xin Yang, Xiang Li, Bin Ning, and Tao Tang. A survey on energy-efficient train operation for urban rail transit. *IEEE Transactions on Intelligent Transportation Systems*, 17(1) :2–13, 2015. 31

[181] Yuxiang Yang, Ping Huang, Qiyuan Peng, LI Jie, and Chao Wen. Statistical delay distribution analysis on high-speed railway trains. *Journal of Modern Transportation*, pages 1–10, 2019. 44

[182] Oleksandra Yezerska, Sergiy Butenko, and Vladimir L Boginski. Detecting robust cliques in graphs subject to uncertain edge failures. *Annals of Operations Research*, 262(1) :109–132, 2018. 134, 135

[183] Jianxin Yuan. Stochastic modelling of train delays and delay propagation in stations. *PhD Thesis, Delft University of Technology.*, 2006. 44

[184] Jianxin Yuan and Ingo A Hansen. Closed form expressions of optimal buffer times between scheduled trains at railway bottlenecks. In *2008 11th International IEEE Conference on Intelligent Transportation Systems*, pages 675–680. IEEE, 2008. 33

[185] Ghazal Zakeri and Nils OE Olsson. Investigating the effect of weather on punctuality of norwegian railways : a case study of the nordland line. *Journal of Modern Transportation*, 26(4) :255–267, 2018. 45

[186] Beiyao Zheng and Alan Agresti. Summarizing the predictive power of a generalized linear model. *Statistics in medicine*, 19(13) :1771–1781, 2000. 74

[187] Peter J Zwaneveld, Leo G Kroon, H Edwin Romeijn, Marc Salomon, Stephane Dauzere-Peres, Stan PM Van Hoesel, and Harrie W Ambergen. Routing trains through railway stations : Model formulation and algorithms. *Transportation science*, 30(3) :181–194, 1996. 34, 36, 37

le cnam

Marie Milliet de Faverges

Développement et implémentation de modèles apprenants pour l'exploitation des grandes gares

SNCF RÉSEAU

Résumé : Cette thèse traite de l'incertitude et de la robustesse dans les problèmes de décision, avec le cas d'application des affectations de quais en gare en cas de retards. Une méthodologie en deux parties est proposée pour aborder ce problème. Dans un premier temps, les archives de données de retards sont utilisées pour construire des modèles de prédiction de distribution de probabilités conditionnellement aux valeurs d'un ensemble de variables explicatives. Une méthodologie de validation et d'évaluation de ces prédictions est mise en place afin d'assurer leur fiabilité pour de la prise de décision. Le problème d'affectations de quais pouvant être vu comme une recherche de clique de taille maximum, ces distributions prédites sont utilisées dans une seconde partie pour ajouter des pondérations pénalisant les risques de rupture des arêtes en cas de retard. Des algorithmes de recherche locale ont été utilisés et les expériences ont montré une importante baisse de conflits.

Mots clés : Apprentissage statistique, Analyse de donnée, Recherche opérationnelle, Problème d'allocations de quais, Exploitation ferroviaire

Abstract : This thesis deals with uncertainty and robustness in decision problems, with the case of the train platforming problem subject to delays. A two-part methodology is proposed to address this problem. First, delay records are used to build models predicting probability distributions conditionnaly to a set of explanatory variables. A methodology to validate and evaluate these predictions is proposed to ensure their reliability for decision-making. As the train platforming problem can be seen as a weighted clique problem, these predicted distributions are used in a second part to add weights on edges to penalize risk of conflict. Local search algorithms are used and experiments show a significant decrease in conflicts.

Keywords : Machine Learning, Data Analysis, Operations Research, Train Platforming Problem, Railway operations