Florent Lin

ENSAE Paris 3ème année
Stage de fin d'études
Année Scolaire 2024/2025

Consultant en Data Sciences pour la Maintenance Prédictive chez SNCF Réseau

Aubay Data & AI
Boulogne-Billancourt

Maître de stage : Nicolas Marivin
Du 13/01/2025 au 11/06/2025

ENSAE Paris
TSA 26644
Service des relations entreprises et des stages
5, avenue Henry Le Chatelier - 91764 PALAISEAU CEDEX - FRANCE - Tél : +33(0)1 70 26 67 39 - Courriel : stage@ensae.fr

Remerciements

Je remercie Aubay Data & AI de m'avoir donné l'opportunité de réaliser mon stage de fin d'études au sein de leur équipe. J'adresse un remerciement particulier à mon maître de stage, Nicolas Marivin, pour ses conseils et son soutien au long de cette expérience. Je remercie aussi les experts métiers de SNCF Réseau pour le partage de leurs connaissances, qui ont été essentiels pour mener à bien mes projets.

Je suis reconnaissant envers Aubay pour l'environnement de travail et les nombreuses opportunités d'apprentissage qui m'ont permis de progresser, aussi bien sur le plan professionnel que personnel. Les échanges avec les équipes et les discussions avec les experts métiers m'ont aidé à mieux comprendre mes points à améliorer et à développer de nouvelles compétences.

Enfin, je remercie l'ENSAE Paris pour la formation académique et le soutien pédagogique reçus tout au long de mon parcours, qui m'ont permis d'aborder ce stage avec de solides bases.

Table des matières

1 Introduction

L'objectif de ce rapport est de présenter les travaux réalisés durant mon stage de fin d'études, centrés sur la transformation de données ferroviaires hétérogènes en indicateurs macro exploitables par les experts métiers. L'idée directrice est de montrer comment des sources multiples et dispersées peuvent être harmonisées et enrichies afin de produire une lecture globale du réseau, utile pour la surveillance et la maintenance.

Ce travail s'inscrit dans la division « PGRN » de SNCF Réseau, spécialisée dans les problématiques liées à la **plateforme ferroviaire**, plus spécifiquement dans l'équipe « SAP ». Dans ce compte rendu, nous étudierons principalement des mesures géométriques (cf. A.1) prises directement sur les rails ferroviaires, cependant c'est bien la plateforme qui supporte les rails qui nous intéresse.

Les problématiques relatives à la plateforme sont souvent reliées au domaine de l'hydraulique, à travers des phénomènes physiques comme le retrait-gonflement des argiles, la stabilité des remblais et des déblais, ainsi que l'efficacité des systèmes de drainage.

Il est évident que mon rôle au cours de cette mission n'est ni de devenir hydraulicien, ni géo-physicien, cependant acquérir une compréhension basique de ces phénomènes physiques, ainsi que des contraintes opérationnelles du métier me semble essentiel pour développer des indicateurs réellement utiles. D'autant plus qu'énormément de facteurs influencent la qualité de la plateforme ferroviaire (profil de plateforme, fréquence et vitesse de passage des trains, composition du sol, conditions météorologiques...).

En prenant ces éléments en considération, la problématique générale peut se résumer ainsi : **comment transformer des données hétérogènes en indicateurs clairs et fiables, permettant aux experts de mieux anticiper les zones sensibles sur le réseau ferroviaire et d'orienter les interventions de maintenance.**

Le document est structuré en six axes principaux. Le premier axe introduit le contexte académique et professionnel du stage. Le deuxième axe apporte du contexte utile à la compréhension du rapport, il décrit les données disponibles et les étapes de construction du pipeline. Le troisième axe présente les indicateurs développés et leurs visualisations, il détaille aussi les besoins des experts du métier et explique comment notre travail s'intègre dans leurs problématiques. Le quatrième axe illustre ces approches à travers plusieurs études de cas, nous profiterons de cette section pour mettre en lumière certains choix dans la construction des indicateurs. Le cinquième axe discute l'impact métier et les limites rencontrées. Enfin, le sixième axe propose une conclusion et des perspectives pour prolonger ces travaux.

2 La Mission

Cette mission avait pour objectif de développer des méthodes de traitement et d'analyse de données, pour mieux comprendre l'état du réseau ferroviaire français. Elle s'est organisée autour de plusieurs projets, notamment la construction d'indicateurs micro et macro, ainsi qu'à notre initiative, l'utilisation de modèles de machine learning pour prédire les zones à risque sur les voies ferroviaires.

La mission s'est appuyée sur l'intégration de données variées (géométrie des voies, ZER, interventions, PIGC, TOUTATIS, TSP, etc.) (cf. 1) Cela a nécessité la mise en place de pipelines solides pour l'ingestion, la normalisation et le contrôle qualité des données. Les livrables produits comprennent des scripts reproductibles, des notebooks d'analyse, ainsi que des visualisations cartographiques et statistiques adaptées aux besoins des experts métiers.

Les apports principaux de cette mission sont le développement d'outils permettant une **lecture à grande échelle cohérente des données**, utile pour repérer plus rapidement les zones à risque, ainsi que des **visualisations plus localisées spatialement**, mais plus large temporellement, pour mieux comprendre les anomalies locales.

On compte par exemple des études sur la saisonnalité des défauts géométriques (cf. 5.2), la prédiction des zones à évolution rapide (ZER) des défauts sur l'ensemble du réseau ferroviaire (cf. 5.1) ou encore l'analyse des talus à partir de données d'altitude du sol français.

2.1 Contexte et Rôle

Actuellement consultant en Data Science chez Aubay Data & IA, une filiale du groupe Aubay (ESN). Ce stage de fin d'étude s'est réalisé dans le cadre d'une mission qui s'étendait sur toute l'année 2025, mais les éléments présentés dans ce rapport ne concernent que la période indiquée du stage qui a eu lieu entre janvier et juin 2025. Il s'est déroulé au sein de la division PGRN, composée d'une trentaine d'experts, spécialisée dans l'analyse des problématiques de plateforme (profil du sol, hydraulique, stabilité de la plateforme, etc.).

Cette mission s'est déroulée en totale autonomie au sein de la SNCF. Mon rôle principal a été d'apporter mes connaissances en statistiques et en manipulation de données pour supporter les besoins de mon client. Il a été question de poursuivre et d'approfondir des travaux déjà entamés par d'autres intervenants, sur divers sujets relatifs à la data, mais surtout de développer de nouveaux outils et indicateurs, en collaboration avec les experts métiers.

Il m'est vite apparu que les deux obstacles majeurs seraient l'hétérogénéité des données, ainsi que les prises de décision sur le développement des projets en raison de la totale autonomie. Notons aussi que l'objectif n'était pas de construire des solutions techniques parfaites, mais plutôt des outils pragmatiques, robustes et reproductibles, qui puissent être utilisés par un large éventail d'experts (même hors de notre division) dans leurs diagnostics quotidiens.

C'est dans cet esprit que j'ai orienté mes travaux vers des visualisations claires et des indicateurs simples à interpréter, tout en documentant chaque étape du développement pour la vérification et la reproductibilité. Il était important que les experts comprennent comment les

indicateurs étaient construits, afin de pouvoir formuler un diagnostic clair à partir des résultats produits.

2.2 Problématique transverse

Compte tenu du large éventail de projets menés au sein de la division PGRN, en réponse aux besoins variés des experts métiers, ce rapport n'a pas vocation à détailler chaque projet individuellement. Les objectifs des différents groupes au sein de la division étaient aussi variés, certains privilégiant un point de vue macro, cherchant à identifier des problématiques en analysant une ligne, voire même le réseau national dans son ensemble, d'autres privilégiant un point de vue micro, cherchant à comprendre précisément les phénomènes physiques sur des sites spécifiques, souvent connus pour avoir des antécédents de défauts.

Ce rapport est structuré autour d'une problématique transverse, illustrée à travers plusieurs études de cas, permettant de montrer comment celle-ci a été abordée dans différents contextes. Ce choix reflète aussi une limite de la mission : cette dynamique de travail avec de multiples équipes, bien qu'enrichissante, a imposé des allocations limitées de temps sur chaque projet, et a rendu difficile la complétion et l'approfondissement de sujets en particulier. Notons que tous les projets qui seront cités sont encore en cours de développement, et nous espérons qu'à travers ce rapport, nous pourrons discuter de pistes d'amélioration et de prolongement.

La problématique transverse, similaire à celle en introduction, s'articule autour de la question suivante : Comment transformer des données ferroviaires hétérogènes (géométrie des voies, ZER, interventions, PIGC, pluies TOUTATIS, TSP, MNT... (cf. 1)), localisées par LIGNE/PK/DATE/VOIE (cf. 3.1.1), en indicateurs et représentations macro et micro fiables, permettant aux experts d'identifier des secteurs de surveillance et de maintenance prioritaires ?

3 Du travail sur les données

3.1 Sources de données

Cette section dresse un inventaire des principaux types de données utilisés dans le cadre de cette mission.

3.1.1 Variables de localisation communes

Nous détaillons dans un premier temps les variables de localisation communes (LIGNE, PK, DATE, VOIE) qui permettent d'harmoniser ces jeux de données hétérogènes.

Variables de localisation communes

— **LIGNE** : identifiant d'une ligne ferroviaire.
— **PK** : *Point Kilométrique* ; position sur la ligne, permettant de localiser les autres paramètres par un début (PK Début) et une fin (PK Fin).
— **DATE** : horodatage de la mesure ou de l'événement (indiquant jour, mois et année).
— **VOIE** : voie concernée (V1, V2, UNIQUE, etc.).
— **Côté** : côté (gauche/droite) de la plateforme ferroviaire.
— **Infrapôle** : division géographique et administrative du réseau ferroviaire national en 28 secteurs (*interne à la SNCF*).

Pour décrire toutes les données qui suivent, nous indiquerons trois informations :

- une brève **définition** indiquant de quoi il s'agit,
- quel type de donnée est mesurée (**mesures** géométriques, altimétriques, textuelles, etc.), et s'il y a lieu, la périodicité temporelle ou spatiale des mesures,
- leur **usage** principal par les experts métiers de la SNCF.

3.1.2 Données variables

Les données dans cette section sont mises à jour régulièrement.

Données de géométrie de voie

Ce sont les variables d'intérêt principales de notre mission, car elles sont une mesure directe de la qualité de la voie. Nous verrons que les sources de données suivantes sont généralement des observations ponctuelles, parfois réalisées par des agents sur le terrain.

— **Défaut de voie courts** : Nivellement (Niv), Écartement (E), Dressage (Dress), Défaut de gauche 3m (G3).
— **Défauts de voie longs** : Nivellement allongé (Nall), Dressage allongé (Dall), Nivellement transveral allongé (Tall).
— **Mesures** : enregistrements tous les 25cm, avec une périodicité de 3 à 6 mois sur ligne classique, et 2 semaines sur lignes à grande vitesse.

— **Usage** : pour chaque défaut, des valeurs seuils sont définis, signalant des besoins d'intervention ou de fermeture de la voie.

Voir l'annexe (cf. A.1) pour des illustrations complémentaires.

ZER (Zones à Évolution Rapide)

— **Définition** : tronçon de voie, ou de ligne, sur lequel la vitesse d'évolution des défauts est telle, qu'on peut craindre qu'après un enregistrement programmé, que ce tronçon puisse être affecté d'un ou plusieurs défauts "LAI" (cf. A.1.3) avant le prochain enregistrement officiel.
— **Mesures** : au moins 20m de part et d'autre d'un défaut, représente 0.4% du réseau ferroviaire.
— **Usage** : identification des sites qui nécessitent une surveillance et des interventions de maintenance.

Interventions et travaux

— **Définition** : enregistrements des interventions et travaux de maintenance réalisés.
— **Mesures** : généralement localisées sur quelques dizaines de mètres, et peuvent s'étendre sur plusieurs kilomètres.
— **Usage** : rétablir les mesures de défauts de géométrie sous des seuils acceptables.

PIGC (Patrimoine Informatisé Génie Civil)

— **Définition** : indique le profil du talus (cf. A.3.1) et des observations sur la plateforme.
— **Mesures** : réalisées lors de tournées d'agents, généralement sur des sites où des problèmes ont déjà été identifiés.
— **Usage** : cartographier et caractériser le réseau national ferroviaire dans une base de données.

Base PANDA

— **Définition** : estimation de l'épaisseur des différentes couches composant la plateforme ferroviaire par pénétrométrie dynamique.
— **Mesures** : valeurs numériques pour chaque couche de la plateforme.
— **Usage** : étudier les conditions de stabilité de la plateforme.

TSP (Tournées de Surveillance Périodique)

— **Définition** : répertoire des observations textuelles d'agents lors des tournées périodiques.
— **Mesures** : fréquence variable selon les lignes, requiert une demande pour extraction par infrapôle.
— **Usage** : filtrage par mot-clé pour la recherche de défauts liés à la plateforme ferroviaire.

3.1.3 Données fixes

Certaines données de terrain ne changent pas au cours du temps, et ne nécessitent donc pas d'identification temporelle lors des analyses.

Segmentation interne

— **Définition** : segmentation du réseau ferroviaire en fonction de la vitesse et de la fréquence de la circulation des trains (categorisation de 1 à 9).

— **Mesures** : tronçons de lignes par dizaine de kilomètres.

— **Usage** : sert à établir une échelle de priorité entre les tronçons de lignes.

MNT (Modèles Numériques de Terrain)

— **Définition** : données altimétriques décrivant le relief et la morphologie de l'ensemble du territoire français.

— **Mesures** : taille de cellule variable entre $1m^2$ et $25m^2$, fiabilité variable selon l'appareil de mesure.

— **Usage** : caractérisation des pentes/profil du talus et des zones sensibles aux instabilités.

Source	Définition	Mesures	Usage principal
Géom. voie	Défauts courts/longs caractérisant la voie.	**Valeurs numériques** chaque 25cm et périodiques	Surveiller seuils et déclencher interventions.
ZER	Tronçons à évolution rapide des défauts.	**Variable binaire** sur des tronçons de voie.	Cibler la surveillance et prioriser la maintenance.
Interv. & travaux	Enregistrements d'actions de maintenance.	**Catégoriel** (type de travaux).	Rétablir la géométrie sous les seuils acceptables.
PIGC	Observations terrain et profil du talus.	**Catégoriel et Textuel** (type du talus et commentaires).	Cartographier et caractériser la plateforme.
TOUTATIS	Signalement de risques météorologiques.	**Catégoriel** (indice de risque de pluviométrie).	Repérer les sites sensibles aux intempéries.
TSP	Observations générales sur l'état de la voie.	**Textuel** (commentaires sur l'état des voies)	Détecter les défauts via mots-clés.
Seg. interne	Découpage du réseau selon la vitesse/fréquentation d'une ligne.	**Variable catégorielle** (indice de 1 à 9).	Prioriser les tronçons critiques.
MNT	Modèle altimétrique du terrain.	**Matrices de valeurs numériques**, précision variable.	Qualifier les pentes et profils de talus.

TABLE 1 – Tableau récapitulatif des sources de données utilisées dans le cadre de cette mission

3.2 Ingestion des données

La récupération des données fut fastidieuse, notre division ne possédant pas d'infrastructure pour l'extraction automatique et à grande échelle de données.

Nous travaillons principalement avec des tableaux Excel, souvent remplis manuellement.

Ces données sont souvent lacunaires, présentées dans des formats inconsistants en fonction des sources ou en doublons.

Par exemple, nous avons eu l'occasion de travailler sur des données localisant des appareils de voie (différents types d'aiguillages) et plusieurs acteurs avaient enregistré l'ensemble des appareils sur le réseau. Cependant, chaque base indique des **localisations en PK très légèrement différentes**, aussi, certains appareils de voie sont des combinaisons d'aiguillages (nommées traversée-jonction simple ou complexe) et sont parfois recensés comme étant un ou plusieurs appareils de voie.

Cet exemple nous permet par ailleurs d'introduire une problématique rencontrée dans plusieurs projets : la présence d'un décalage du PK entre les différentes tournées d'acquisitions de données. Cela signifie qu'il y a des incertitudes sur la geolocalisation des mesures, et que deux mesures avec le même PK peuvent avoir des localisations réelles différentes. Ces différences peuvent être de l'ordre du mètre et aller jusqu'à quelques dizaines mètres.

Ces imprécisions sont principalement dues à la précision conférée par la localisation GPS lors des tournées, mais aussi liées à des erreurs de projection. Sur des sections courbes, les mesures embarquées de la géométrie de la voie peuvent présenter de légers décalages par rapport aux PK théoriques. Ce phénomène est lié à la curvilinéarité de la voie et aux méthodes d'acquisition des appareils, et peut expliquer certaines divergences entre sources de données. La SNCF procède généralement à une synchronisation des données, on dispose notamment aujourd'hui d'une précision garantie inférieure à 2 mètres sur tout le long des lignes à grandes vitesses. Cependant, cela reste un défi sur les lignes classiques, une autre division avec laquelle j'ai eu l'occasion de travailler sur le projet *Colormap* (cf. 5.2) travaille notamment sur cette problématique.

La problématique de synchronisation représente aussi un défi majeur au sein de notre section, car notre objectif est aussi de **détecter des corrélations** entre différentes sources de données. Il est donc nécessaire de corriger les décalages importants en PK afin de pouvoir connecter les bases de données entre elles.

Malheureusement, cette problématique plus générale n'a pas encore été approfondie dans le cadre de cette mission. Nous nous sommes cependant concentrés sur la synchronisation des PK entre les tournées pour une même source de données.

4 Définissons un cadre

Après avoir établi un pipeline robuste d'intégration des données, nous disposons désormais d'un cadre permettant d'analyser l'état du réseau ferroviaire. Afin de répondre à la problématique centrale du stage — transformer des données hétérogènes en indicateurs exploitables et anticiper les zones à risque — nous allons formaliser mathématiquement les objets avec lesquels nous travaillons. Cette formalisation nous permettra *(i)* de définir rigoureusement les unités spatiales et temporelles d'analyse, *(ii)* de construire des indicateurs macro à partir d'agrégations contrôlées, et *(iii)* de poser les bases théoriques pour des modèles prédictifs, notamment pour l'apparition de Zones à Évolution Rapide (ZER).

Dans cette section, nous introduisons donc un système de notations cohérent qui structurera l'ensemble de nos analyses et servira de fondation aux études de cas présentées ultérieurement.

4.1 Les variables du problème

Les segments

Nous choisissons comme unité spatiale un **segment** s défini par un quadruplet

$$s = (\text{LIGNE} = \ell,\ \text{VOIE} = v,\ \text{PK}_{\text{début}} = p,\ \Delta x), \quad s \in \mathcal{S}$$

c'est-à-dire un intervalle $\big[p,\ p + \Delta x\big)$, défini par :

— $\ell \in \mathcal{L}$: ensemble fini des lignes ferroviaires ;

— $v \in \mathcal{V}_\ell$: ensemble fini des voies sur la ligne ℓ ;

— $p \in \mathbb{R}^+$: position en PK sur la ligne ℓ ;

— $\Delta x \in \mathbb{R}^+$: taille du segment ;

de sorte que $\mathcal{S} \subset \mathcal{L} \times \mathcal{V}_\ell \times \mathbb{R}^+ \times \mathbb{R}^+$, de plus notons N_s le nombre de points p dans le segment s.

Notons que Δx est défini en fonction les longueurs d'intérêt métier, par exemple : *(i)* la mesure du défaut de gauche 3 m, *(ii)* celle d'un défaut allongé comme le Nall s'étend sur 20 à 30 m (cf. A.1) et *(iii)* l'extension minimale d'une ZER (au moins ≈ 20 m de part et d'autre d'un défaut, cf. 3.1.2). En pratique, nous définirons $\Delta x \in [10, 250]$ m, selon l'objectif (diagnostic local vs. lecture macroscopique).

Les variables observées

On considère les dates d'acquisition des défauts $\mathcal{T}_\ell = \{t_1 < t_2 < \cdots\}$ propres à chaque ligne ℓ. On définit ensuite l'ensemble des défauts géométriques par $\mathcal{D}$. Soit $d \in \mathcal{D}$ un défaut géométrique (par exemple le Niv, E, Dress, G3, Nall, Dall, Tall), soit $s \in \mathcal{S}$ un segment, pour toute position $p \in s$ et pour tout pas de temps $t \in \mathcal{T}_\ell$, on définit alors la variable aléatoire suivante :

$$X_d(p, t) \in \mathbb{R},$$

qui encode la valeur du défaut d à la position p et au temps t.

Les fonctions d'agrégation

Nous pouvons ainsi définir la moyenne et l'écart-type spatial pour étudier X_d à l'échelle du segment :

$$\bar{X}_d(s,t) = \frac{1}{N_s} \sum_{p \in s} X_d(p,t)$$

$$\sigma_d(s,t) = \sqrt{\frac{1}{N_s} \sum_{p \in s} \left(X_d(p,t) - \bar{X}_d(s,t)\right)^2}$$

Nous travaillerons également avec les quantiles de $X_d(s,t)$ que nous noterons $Q_{d,\alpha}(s,t)$ pour $\alpha \in [0,1]$.

Dans une démarche de prévention des risques (et donc des fortes évolutions des défauts), nous sommes tout particulièrement intéressés par l'apparition d'une ZER 3.1.2.

Pour cette étude, nous allons définir **le vecteur de variables explicatives** $\mathbf{X}_{s,t}$ comme l'ensemble des variables explicatives (toutes les sources de données introduites dans la section 4.1) disponibles pour le segment s au temps t. Ce vecteur correspond à l'ensemble des mesures de défauts agrégés plus les variables contextuelles (conditions météorologiques, type de plateforme, fréquence de passage, etc.).

4.2 Modèle de ZER

La variable cible

Pour une ligne ℓ, on définit T_ℓ, le temps qui sépare deux mesures des défauts sur la ligne et la variable binaire $Y_{\mathrm{ZER}}(s,t)$ indique l'**apparition d'une ZER** sur un segement s entre t et $t + \mathrm{T}_\ell$:

$$Y_{\mathrm{ZER}}(s,t) = \mathbb{1}\{\exists\, t' \in (t,\, t + \mathrm{T}_\ell] \text{ tel que } s \text{ est étiqueté ZER à } t'\}\,.$$

Nous appellerons T_ℓ, la période de la ligne ℓ.

Ainsi, pour tout segment s fixé nous pouvons définir la série temporelle binaire $Y_s(t)_{t \in \mathcal{T}_\ell}$. On peut voir cette série comme un processus stochastique discret indexé par le temps.

On peut aussi supposer qu'à chaque instant t,

$$Y_s(t) \sim \mathcal{B}(p_s(t)).$$

où $p_s(t) = P\left(Y_s(t) = 1 \,|\, \mathbf{X}_{s,t}\right)$ est le paramètre de la loi de Bernoulli conditionnellement à l'historique des mesures de toutes les sources de données $\mathbf{X}_{s,t}$. Ainsi, la série temporelle binaire observée est une suite de lois de Bernoulli dont le paramètre peut varier au cours du temps. Il traduit le potentiel d'apparition d'une ZER. Déterminer ce paramètre pourrait nous permettre d'identifier les segments à surveiller.

Prédiction de la probabilité d'apparition d'une ZER

Nous pourrions par exemple proposer une approche par la loi des grands nombres, en regroupant les observations sur des segments similaires (même conditions météorologiques, même segmentation interne, même type de talus, etc.).
Il nous faudrait alors calculer :

$$\hat{p}_s(t) = \frac{\text{Nombre de cas où } Y = 1}{\text{Nombre total d'observations}}.$$

Dans la pratique, nous ne disposons pas d'un nombre suffisant d'observations pour chaque segment. En réalité, nous n'avons pas de données sur la vaste majorité du réseau.

Cependant, nous avons à faire à une variable binaire, et il existe diverses approches en machine learning pour prédire une telle variable(cf. 5.1).

Tableau des notations

Notation	Description
$\mathcal{L}$	Ensemble des lignes ferroviaires
$\ell \in \mathcal{L}$	Une ligne ferroviaire
$\mathcal{V}_\ell$	Ensemble des voies sur la ligne ℓ
$v \in \mathcal{V}_\ell$	Une voie
$p \in \mathbb{R}^+$	Position en PK (point kilométrique)
$\Delta x \in \mathbb{R}^+$	Taille du segment (typiquement 20–200 m)
$s = (\ell, v, p, \Delta x)$	Un segment (unité spatiale d'analyse)
$\mathcal{S}$	Ensemble de tous les segments
N_s	Nombre de points de mesure dans le segment s
$\mathcal{T}_\ell$	Dates d'acquisition des défauts sur la ligne ℓ
$t \in \mathcal{T}_\ell$	Un instant de mesure
T_ℓ	Période entre deux mesures sur la ligne ℓ
$\mathcal{D}$	Ensemble des défauts géométriques (Niv, E, Dress, G3, etc.)
$d \in \mathcal{D}$	Un défaut géométrique
$X_d(p, t)$	Valeur du défaut d à la position p et au temps t
$\bar{X}_d(s, t)$	Moyenne spatiale du défaut d sur le segment s au temps t
$\sigma_d(s, t)$	Écart-type spatial du défaut d sur le segment s au temps t
$Q_{d,\alpha}(s, t)$	Quantile d'ordre α du défaut d sur le segment s au temps t
$\mathbf{X}_{s,t}$	Historique des mesures pour le segment s jusqu'au temps t
$Y_{\mathrm{ZER}}(s, t)$	Variable binaire d'apparition d'une ZER sur s entre t et $t + \mathrm{T}_\ell$
$Y_s(t)$	Série temporelle binaire d'apparition de ZER pour le segment s
$p_s(t)$	Probabilité d'apparition d'une ZER sur s au temps t
$\hat{p}_s(t)$	Estimateur de la probabilité d'apparition d'une ZER

TABLE 2 – Tableau des notations de la modélisation

4.3 Formulations d'apprentissage

Pour prédire l'apparition d'une ZER sur un segment s au temps t, nous cherchons à estimer la probabilité $p_s(t) = P(Y_s(t) = 1 \mid \mathbf{X}_{s,t})$, où $\mathbf{X}_{s,t}$ représente l'ensemble des variables explicatives (défauts géométriques agrégés, contexte, historique) disponibles pour le segment s au temps t. Nous présentons ici quatre approches de machine learning adaptées à ce problème de classification binaire [1].

4.3.1 Régression logistique

La régression logistique [1] modélise la probabilité $p_s(t)$ de l'apparition d'une ZER sur le segment s au temps t comme une transformation sigmoïde des variables explicatives $\mathbf{X}_{s,t}$:

$$p_s(t) = \frac{1}{1 + \exp\left(-\mathbf{X}_{s,t}\beta - \beta_0\right)}$$

où β est le vecteur de coefficients appris par la régression logistique pour chaque variable explicative décrite par $\mathbf{X}_{s,t}$ et β_0 est le terme constant.

Ce modèle présente l'avantage d'être interprétable : chaque coefficient indique l'influence marginale d'une variable sur l'apparition d'une ZER. Cependant, il suppose une relation linéaire entre les features et le *log-odds* :

$$\log\left(\frac{p_s(t)}{1 - p_s(t)}\right) = \mathbf{X}_{s,t}\beta + \beta_0\,,$$

ce qui peut être limitant pour capturer des interactions complexes entre défauts.

4.3.2 Decision Tree

Un arbre de décision [2] partitionne récursivement l'espace des features en régions homogènes selon la variable cible. Formellement, pour un segment s au temps t, le modèle peut s'écrire :

$$p_s(t) = \sum_{k=1}^{K} c_k \cdot \mathbb{1}\{\mathbf{X}_{s,t} \in R_k\}$$

où K est le nombre de feuilles, R_k sont les régions définies par les règles de décision, et c_k est la proportion de ZER dans la feuille k de l'ensemble d'entraînement.

Les arbres capturent naturellement les interactions et les non-linéarités sans transformation explicite des features. Ils offrent également une forte interprétabilité grâce à la visualisation des règles de décision. Néanmoins, un arbre unique tend à surajuster les données d'entraînement et présente une variance élevée.

1. Nous nous appuyons sur les modèles de machine learning développés par la bibliothèque scikit-learn.

4.3.3 Random Forest

Une forêt aléatoire [3] agrège les prédictions de N arbres de décision entraînés indépendamment sur des échantillons bootstrap des données, avec une sélection aléatoire des features à chaque nœud. La prédiction finale s'obtient par vote majoritaire :

$$p_s(t) = \frac{1}{N} \sum_{n=1}^{N} \hat{p}_n(s,t)$$

où $\hat{p}_n(s,t)$ est la probabilité estimée par l'arbre n.

Cette approche d'ensemble réduit significativement la variance des arbres individuels tout en préservant leur capacité à modéliser des relations complexes. Les forêts aléatoires sont robustes au surapprentissage. L'importance des variables peut être évaluée via la diminution moyenne de l'impureté, bien que l'interprétabilité soit perdue par rapport à un arbre unique.

4.3.4 XGBoost

XGBoost construit séquentiellement une séquence d'arbres faibles (peu profonds), où chaque nouvel arbre corrige les erreurs résiduelles des arbres précédents. Basé sur [4], ce modèle s'écrit :

$$\hat{y}_s(t) = \sum_{m=1}^{M} \eta \cdot f_m(\mathbf{X}_{s,t}), \quad p_s(t) = \sigma(\hat{y}_s(t))$$

où f_m est le m-ième arbre, $\eta \in (0, 1]$ est le taux d'apprentissage (learning rate), M est le nombre d'itérations, et σ est la fonction sigmoïde.

À chaque itération, XGBoost minimise une fonction de perte régularisée qui pénalise la complexité des arbres, réduisant ainsi le risque de surapprentissage. Ce modèle offre des performances prédictives souvent supérieures aux forêts aléatoires, notamment grâce à sa capacité d'optimisation fine (gestion des classes déséquilibrées via pondération, contrôle de la profondeur, régularisation L1/L2).

4.4 Hypothèses de travail

1. **Stationnarité locale.** À granularité Δx et sur fenêtres temporelles courtes, les mécanismes d'évolution sont approximativement constants.
2. **Bruits centrés.** Les erreurs d'observation $\eta_{d,s,t}$ sont à moyenne nulle et variance bornée ; leur variance peut dépendre de la ligne et du défaut.
3. **Décalage PK borné.** L'alignement entre sources présente un *décalage spatial* Δp borné par δ mètres ; l'appariement s'effectue via une fenêtre tolérante ou un noyau $K(|\Delta p|/\delta)$.
4. **Interventions exogènes.** Les interventions I sont considérées comme des chocs *observés* ; pour simplifier, supposons qu'elles réinitialisent des tendances.

4.5 Validation métier

Bien que nos clients apprécient l'approche nouvelle apportée par l'utilisation de modèles de machine learning, l'interprétabilité et la transparence sont des critères importants pour eux. Nos algorithmes s'inscrivent dans une démarche de soutien pour les experts métiers, afin de les épauler dans leur compréhension des phénomènes physiques qui perturbent le réseau ferroviaire.

5 Quelques études de cas

Nous allons présenter dans cette section deux projets réalisés en collaboration avec des clients.

Le premier projet a pour vocation de prédire l'apparition de ZER, qui représentent environ 0,4% des linéaires couverts par le réseau ferroviaire. Développer des modèles de prédiction nous permettra en même temps de caractériser ces zones et ainsi mieux comprendre les phénomènes physiques (relatifs à la plateforme) derrière les variations rapides de défauts sur le réseau.

Le second projet vise à apporter une vision plus large (sur l'espace et le temps) pour des défauts de voies localisés. Nous travaillons principalement sur des sites déjà connus et étudiés. Les principales motivations de ce projet sont (*i*) permettre de mesurer l'impact des interventions de maintenance et (*ii*) de repérer des éventuelles saisonnalités dans l'apparition des défauts de voies.

5.1 ZER ML — Décrire les zones à évolution rapide

5.1.1 Présentation du projet

Le cœur du travail au sein des équipes de maintenance de la voie ferrée est de maintenir les valeurs mesurées de défauts en dessous de seuils fixés (cf. A.1.3). Il existe quatre seuils : VO (Valeur d'objectif), VA (Valeur d'alerte), VI (Valeur d'intervention) et LAI (Limite d'action immédiate). À partir du seuil VI, des planifications de maintenance sont effectuées, et dépassé le seuil LAI, soit une maintenance immédiate, soit une fermeture complète de la ligne est nécessaire. Ces mesures de la valeur des défauts sont réalisées périodiquement sur chaque ligne ferroviaire.

Une ZER constitue alors un segment de voie dont la mesure à la date t n'est pas encore une LAI, mais dont on craindrait qu'elle le devienne avant le prochain enregistrement à $t + \mathrm{T}_\ell$.

On remarque que dans notre définition de la variable cible $Y_{\mathrm{ZER}}(s,t)$ (cf. 4.2), nous ne cherchons pas à prédire l'apparition d'une LAI directement, mais plutôt à capturer un contexte plus large, qui permettrait de mieux identifier les phénomènes physiques derrière les variations rapides de défauts sur le réseau.

5.1.2 Données et pré-traitement

La construction du dataset pour ce problème de classification binaire nécessite une stratégie particulière pour gérer le déséquilibre extrême des classes. Les ZER ne représentent qu'environ **0,4% du réseau ferroviaire national**, posant un défi majeur pour l'apprentissage automatique.

Constitution de la base d'apprentissage

Notre approche s'articule en trois étapes principales :

1. **Extraction des cas positifs** : extraction de l'ensemble des ZER répertoriées dans la base de données nationale, caractérisées par leur localisation (LIGNE, VOIE, PK Début, PK Fin) et leur date de mise en place. Conformément aux notations introduites en section 4.1,

chaque ZER est représentée par un segment $s = (\ell, v, p, \Delta x)$ avec $\Delta x = 100\,\text{m}$ (50 m de part et d'autre du milieu de la ZER).

2. **Génération de cas négatifs par décalage spatial** : pour créer des exemples de tronçons *non-ZER* comparables aux ZER, nous appliquons un décalage spatial de ±2 km aux ZER existantes. Cette méthode génère 800 échantillons négatifs situés dans des contextes géographiques similaires, mais suffisamment éloignés pour ne pas être affectés par les mêmes problématiques locales.
3. **Enrichissement par défauts non-ZER** : ajout de 800 zones présentant des défauts dépassant au moins le seuil VA mais n'ayant jamais été classées en ZER. Ces zones sont échantillonnées aléatoirement parmi les voies (nommées V1 et V2) de sorte à maximiser les possibilités de jointure avec les autres sources de données. Pour chaque défaut, nous définissons un segment de ±50 m autour du point de détection.

Stratégie de validation géographique

Afin d'évaluer la capacité de généralisation du modèle, nous adoptons une stratégie de validation géographique : l'ensemble des ZER de **l'infrapôle Centre** est exclu du jeu d'entraînement et réservé pour l'inférence. Cette approche permet de tester la performance du modèle sur une région non vue durant l'apprentissage, simulant ainsi son déploiement opérationnel sur de nouvelles zones du réseau.

Jointure des sources de données

Pour chaque segment (ZER ou non-ZER), nous procédons à une jointure spatiotemporelle avec huit sources de données différentes (cf. tableau 1) :

— **Géométrie de voie** : pour chaque type de défaut (NL, NT, D, VEC, EC), extraction des 6 dernières mesures antérieures à la date de référence du segment, permettant de capturer l'historique d'évolution et de calculer des tendances temporelles.
— **Interventions de maintenance** : cumul du nombre d'interventions par type (BML, Reprise Localisée, Meulage, RVB, RB) sur les 10 dernières années précédant la date de référence.
— **Base PANDA** : extraction des caractéristiques du sol (humidité de la couche de ballast, sous-couche, couche intermédiaire, sous-sol, et présence de glaise) mesurées lors des campagnes géotechniques.
— **Segmentation stratégique interne** : récupération de l'indice de priorité du tronçon (de 1 à 9) en fonction de la vitesse et de la fréquence de circulation.
— **TSP (Tournées de Surveillance Périodique)** : comptage de l'apparition ce certains mots clés relatifs à des problématiques hydrauliques dans les observations textuelles reportées par les agents sur le segment considéré.

Pour les zones se trouvant sur des ZER, cette opération de jointure a été réalisée avec une attention particulière portée à l'alignement temporel des données, en veillant bien à ne conserver que les données existantes avant leur apparition. Notre but étant de déterminer les conditions aboutissant à la labélisation d'une ZER.

Pour le reste des segments, nous avons conservé toutes les données existantes avant la date de référence t.

Ingénierie de features

À partir des données brutes jointes, nous créons des features agrégées et dérivées pour tenter d'améliorer le pouvoir prédictif du modèle :

— **Statistiques temporelles sur les défauts géométriques** : pour chaque type de défaut $d \in \{NL, NT, D, VEC, EC\}$ et pour chaque segment s, on procède au calcul de la moyenne $\bar{X}_d(s)$, de l'écart-type $\sigma_d(s)$, et de la tendance temporelle définie comme la différence entre la mesure la plus récente et la plus ancienne :

$$\text{Tendance}_d(s) = X_{d,1}(s) - X_{d,6}(s)$$

où $X_{d,i}(s)$ désigne la i-ième mesure la plus récente du défaut d sur le segment s.

— **Cumul total d'interventions** : somme de tous les types d'interventions de maintenance sur les 10 dernières années, offrant un indicateur global de l'intensité de la maintenance.

— **Features d'interaction** : création de termes d'interaction pour capturer les effets combinés de certaines variables, notamment l'interaction entre la présence de glaise et l'humidité du sol, connue pour accentuer les phénomènes de retrait-gonflement.

Au total, après nettoyage et ingénierie de features, le dataset comporte environ **1 134 observations** réparties approximativement équitablement entre ZER (classe positive) et non-ZER (classe négative), décrites par plus de **60 variables explicatives**.

5.1.3 Methodologie

Sélection de features

Face au nombre important de variables explicatives (plus de 60), nous appliquons une méthode de sélection de features basée sur l'**information mutuelle** (*mutual information*). Cette approche mesure la dépendance statistique entre chaque variable explicative X_i et la variable cible Y_{ZER} :

$$I(X_i; Y_{\text{ZER}}) = \sum_{x_i} \sum_{y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)}$$

où $p(x_i, y)$ est la distribution jointe et $p(x_i)$, $p(y)$ sont les distributions marginales.

L'information mutuelle présente l'avantage de capturer les relations non-linéaires entre les variables. Nous retenons les **30 features** présentant les scores d'information mutuelle les plus élevés. Théoriquement, cette réduction dimensionnelle améliore la généralisation du modèle tout en réduisant les temps de calcul. Malheureusement, cette méthode n'a pas permis d'améliorer les performances du modèle (cf. 10).

Modèles de classification

Nous évaluons quatre algorithmes de classification binaire, chacun ayant des propriétés différentes en termes d'interprétabilité, de capacité à modéliser des relations non-linéaires, et de

robustesse au déséquilibre des classes :

1. **Régression Logistique** : modèle linéaire servant de référence, offrant une excellente interprétabilité des coefficients (cf. section 4.3).
2. **Random Forest** : méthode d'ensemble combinant plusieurs arbres de décision entraînés sur des sous-échantillons bootstrap des données.
3. **Gradient Boosting** : méthode de boosting construisant séquentiellement des arbres faibles pour corriger les erreurs des modèles précédents.
4. **XGBoost** : variante optimisée du Gradient Boosting avec régularisation et gestion avancée du déséquilibre des classes.

Tous les modèles sont entraînés sur 80% des données, les 20% restants étant réservés pour l'évaluation. Les hyperparamètres du modèle Random Forest, qui s'est révélé le plus performant lors des tests préliminaires, sont ensuite optimisés à l'aide de la librairie open-source **Hyperopt**, fondée sur l'optimisation bayésienne de l'arbre de Parzen structuré (TPE). Cet algorithme procède à une recherche intelligente des hyperparamètres optimaux, sans tester toutes les combinaisons possibles.

Métriques d'évaluation

Étant donné le contexte opérationnel, où il est plus coûteux de manquer une ZER (faux négatif) que de signaler à tort une zone non-ZER (faux positif), nous privilégions les métriques suivantes :

— **Recall** : proportion de ZER correctement identifiées parmi toutes les ZER réelles. Un recall élevé garantit qu'un minimum de zones à risque échappent au système de surveillance.
— **F1-score** : moyenne harmonique de la précision et du recall, offrant un équilibre entre les deux métriques.
— **ROC AUC** : aire sous la courbe ROC, mesurant la capacité du modèle à discriminer les deux classes sur l'ensemble des seuils de décision possibles.

Pour affiner le compromis précision-recall, nous appliquons également une optimisation du seuil de décision (avec la librairie *Hyperopt* [5]) en maximisant le F1-score sur l'ensemble de validation, au lieu d'utiliser le seuil standard de 0,5. Bien que nous pourrions chercher à maximiser uniquement le recall pour réduire les faux négatifs, nous avons préféré optimiser le F1-score pour éviter de manquer une ZER.

5.1.4 Résultats

Performances des modèles

Le tableau ci-dessous résume les performances des quatre modèles testés sur l'ensemble de test (avec les 60 features sans optimisation du seuil de décision) :

Modèle	Accuracy	Precision	Recall	F1-score	ROC AUC
Régression Logistique	0,63	0,66	0,44	0,50	0,69
Random Forest	0,69	0,76	0,51	**0,59**	0,74
XGBoost	0,66	0,69	0,50	0,55	0,74

TABLE 3 – Performances comparées des modèles de prédiction de ZER sur l'ensemble de test

Le modèle **Random Forest** obtient le meilleur F1-score (0,59) sur l'ensemble de test non optimisé. Après optimisation des hyperparamètres via Hyperopt, le Random Forest optimisé atteint un **RECALL de 0,61**, un **F1-score de 0,68** avec un **ROC AUC de 0,77**, démontrant une capacité significativement améliorée à discriminer les ZER des non-ZER. Les graphiques de l'annexe 11 illustrent les résultats de l'optimisation.

L'analyse de la matrice de confusion du Random Forest optimisé révèle :

— 515 faux positifs (zones non-ZER prédites comme ZER)

— 210 faux négatifs (ZER manquées par le modèle)

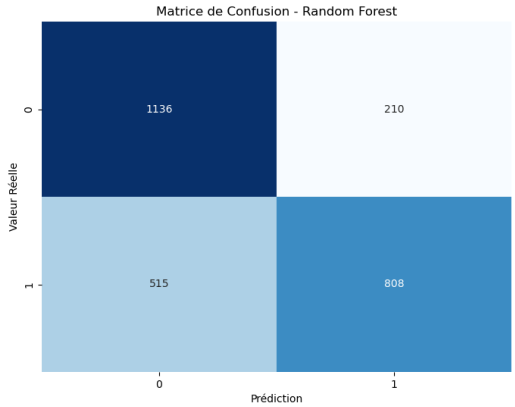

FIGURE 1 – Matrice de confusion du Random Forest optimisé

Ce déséquilibre traduit le comportement attendu du modèle après optimisation du seuil : privilégier le recall (capturer un maximum de ZER réelles) au détriment de la précision. Dans le contexte opérationnel, cette configuration est préférable car elle minimise le risque de laisser des zones critiques sans surveillance.

Importance des features

L'analyse d'importance des features (graphique en annexe 12) via la librairie *SHAP* [6] (*SHapley Additive exPlanations*) révèle les variables les plus influentes pour la prédiction des ZER.

1. **cumul_meulage** (0,035 ± 0,006) : le nombre d'interventions de meulage sur les 10 dernières années apparaît comme le prédicteur le plus important. Le meulage est généralement effectué pour corriger les défauts de surface du rail, et un historique d'interventions élevé peut indiquer une zone sujette à des problèmes récurrents.

2. **vec_std** (0,007 ± 0,006) : l'écart-type de la vitesse d'écartement (VEC) mesure la variabilité de cet indicateur de défaut au cours du temps. Une forte variabilité suggère une instabilité de la géométrie de la voie.
3. **d_5, d_1, d_trend** : les mesures de dressage (D) et leur tendance temporelle figurent parmi les features les plus discriminantes. Le dressage caractérise l'alignement horizontal de la voie, et son évolution rapide est un signal fort d'apparition potentielle d'une ZER.
4. **ec_std** : l'écart-type de l'écartement mesure également la stabilité géométrique.
5. **nl_1, nl_2, nl_mean** : les mesures de nivellement (NL) et leurs statistiques agrégées captent les déformations verticales de la voie.

L'analyse SHAP [6] confirme que les features géométriques temporelles (tendances, écarts-types) et l'historique d'interventions dominent les contributions aux prédictions, tandis que les variables environnementales (humidité, géologie) ont un impact plus marginal. Cela suggère que l'évolution récente des défauts géométriques est un meilleur prédicteur de l'apparition d'une ZER que le contexte géologique seul.

Validation géographique sur l'infrapôle Centre

Le modèle entraîné a été appliqué aux segments de l'infrapôle Centre, exclus de l'entraînement. Cette étape de validation permet de simuler le déploiement opérationnel du modèle sur une nouvelle région du réseau. Les performances sur cet ensemble d'inférence sont très satisfaisantes avec 0.77 d'accuracy (contre seulement 0.73 sur les données de test), ce qui pourrait être expliqué par des spécificités de la région Centre, qui la rend plus prévisible par notre modèle. Dans tous les cas, il semblerait que le modèle conserve une capacité discriminante raisonnable, suggérant qu'il capture des mécanismes généralisables à l'échelle nationale.

5.1.5 Limites et perspectives

Limites méthodologiques

Plusieurs limites doivent être soulignées concernant cette approche :

1. **Biais d'échantillonnage** : la génération de cas négatifs par décalage spatial de ±2 km introduit un biais potentiel, car ces zones peuvent partager des caractéristiques similaires avec les ZER voisines (même type de sol, même profil de ligne). Une approche alternative consisterait à échantillonner des tronçons véritablement aléatoires sur l'ensemble du réseau.
2. **Déséquilibre résiduel** : bien que nous ayons équilibré artificiellement les classes dans le jeu d'entraînement, la réalité opérationnelle reste fortement déséquilibrée (0,4% du réseau). Le modèle risque donc de produire un nombre important de fausses alertes lors d'un déploiement à grande échelle, nécessitant un post-traitement ou un ajustement du seuil de décision.
3. **Absence de dimension temporelle explicite** : le modèle actuel traite chaque segment indépendamment, sans modéliser explicitement la dynamique temporelle d'évolution des

défauts. Une approche par séries temporelles (LSTM, Transformers) pourrait permettre de prédire *quand* une ZER apparaîtra, et pas seulement *si* elle apparaîtra.

4. **Interprétabilité limitée des ensembles d'arbres** : bien que le Random Forest offre une certaine interprétabilité via l'importance des features, il reste difficile d'expliquer une prédiction individuelle aux experts métiers. L'utilisation de méthodes d'explicabilité comme SHAP pourrait faciliter l'adoption du modèle.
5. **Données manquantes et qualité variable** : certaines sources de données (PANDA, TSP, données environnementales) ne couvrent pas uniformément l'ensemble du réseau, ce qui peut limiter la performance du modèle sur certaines zones géographiques. De plus, il reste sans aucun doute un travail important à faire sur la qualité des données, notamment sur les défauts géométriques (une grande amélioration a été réalisée hors période de stage, cf. annexe 4).

Impact métier

Malgré ses limites, ce projet constitue une première étape prometteuse vers l'utilisation de l'apprentissage automatique pour la priorisation de la surveillance du réseau ferroviaire. Les retours des experts métiers ont été globalement positifs (et optimistes), soulignant l'intérêt d'une approche systématique pour identifier des zones à risque qui auraient pu échapper à une analyse manuelle. Le modèle développé pourrait être intégré dans un système d'aide à la décision, où les prédictions serviraient à enrichir les tournées de surveillance et à planifier les campagnes de mesures géométriques supplémentaires. À terme, une telle approche pourrait contribuer à réduire le nombre d'incidents liés aux dégradations rapides de la voie, tout en optimisant l'allocation des ressources de maintenance.

5.2 Cartes de chaleurs

5.2.1 Présentation du projet

Ce projet vise à développer des outils de visualisation permettant aux experts métiers d'identifier rapidement les zones problématiques sur le réseau ferroviaire à travers une lecture spatio-temporelle des défauts géométriques. Contrairement aux approches ponctuelles traditionnelles, ces visualisations offrent une vue d'ensemble sur plusieurs années, permettant de repérer des tendances et des patterns récurrents.

L'objectif principal est double : *(i)* faciliter la détection de zones à surveiller en priorité, et *(ii)* permettre l'évaluation de l'efficacité des interventions de maintenance. Ces outils s'inscrivent dans la démarche d'aide à la décision pour la planification des tournées de surveillance et la priorisation des actions correctives.

Nous avons développé deux représentations complémentaires : la première se concentre sur les dépassements de seuils réglementaires (VA, VI, LAI), tandis que la seconde utilise la variabilité spatiale des défauts comme indicateur d'instabilité de la plateforme.

5.2.2 Une première représentation : Carte des dépassements de seuils

Cette première approche visualise directement les valeurs maximales de défauts géométriques en les catégorisant selon les seuils réglementaires. L'idée directrice est de repérer les zones et périodes où les défauts dépassent les valeurs critiques, nécessitant une intervention.

Données et pré-traitement

Les données sont extraites d'un logiciel appelé *Ultrapot*, développé en interne et utilisé seulement par quelques divisions de SNCF Réseau, qui répertorie les enregistrements des mesures de défauts géométriques lors des tournées d'enregistrement. Pour chaque ligne ℓ et voie v étudiée, nous extrayons :

— Les mesures de défauts courts (N1 [2]) et longs (Nall, Dall, Tall) avec leurs valeurs absolues
— Les dates de tournée et positions PK correspondantes
— Les données d'interventions (BML, Reprise Localisée) et de travaux (RVB, RBRT, RB)
— Les ouvrages d'art : passages à niveau (PN), ponts-rails (PRA), et appareils de voie (ADV)

Le pré-traitement s'articule autour de deux discrétisations complémentaires :

— **Discrétisation spatiale** : les PK sont regroupés en intervalles Δx de 10 mètres, créant des bins $[p_i, p_i + \Delta x)$ km où p_i correspond au centre de l'intervalle
— **Discrétisation temporelle** : les dates de tournée sont regroupées en bins temporels uniformes, avec un maximum de 50 bins pour garantir la lisibilité

Pour chaque cellule (p_i, t_j) du maillage spatio-temporel, nous calculons la valeur maximale du défaut observée dans cet intervalle :

$$V_{max}(p_i, t_j) = \max_{(p,t)\in[p_i,p_i+\Delta x)\times[t_j,t_{j+1})} |X_d(p,t)|$$

Méthodologie

La visualisation repose sur une échelle de couleurs discrète basée sur les seuils réglementaires définis en section A.1.3. Pour chaque type de défaut d, nous définissons une fonction de coloration C_d :

$$C_d(V_{max}) = \begin{cases} \text{Gris clair} & \text{si } V_{max} < VA \\ \text{Jaune} & \text{si } VA \le V_{max} < VI \\ \text{Orange} & \text{si } VI \le V_{max} < LAI \\ \text{Violet pâle} & \text{si } V_{max} \ge LAI \end{cases}$$

Les seuils varient selon le paramètre étudié. Par exemple, pour le défaut G3 : $VA \in [5, 7]$ mm, $VI \in [7, 15]$ mm, $LAI \ge 15$ mm, tandis que pour Nall : $VA \in [10, 18]$ mm, $VI \in [18, 26]$ mm, $LAI \ge 26$ mm.

Les interventions et travaux sont superposés sur la heatmap sous forme de barres horizontales, permettant de corréler visuellement les actions de maintenance avec l'évolution des défauts. Les

2. N1 et N2 sont des variantes du défaut Niv, cf. A.1

ouvrages (PN, PRA, ADV) sont affichés en haut de la carte pour contextualiser géographiquement les zones analysées.

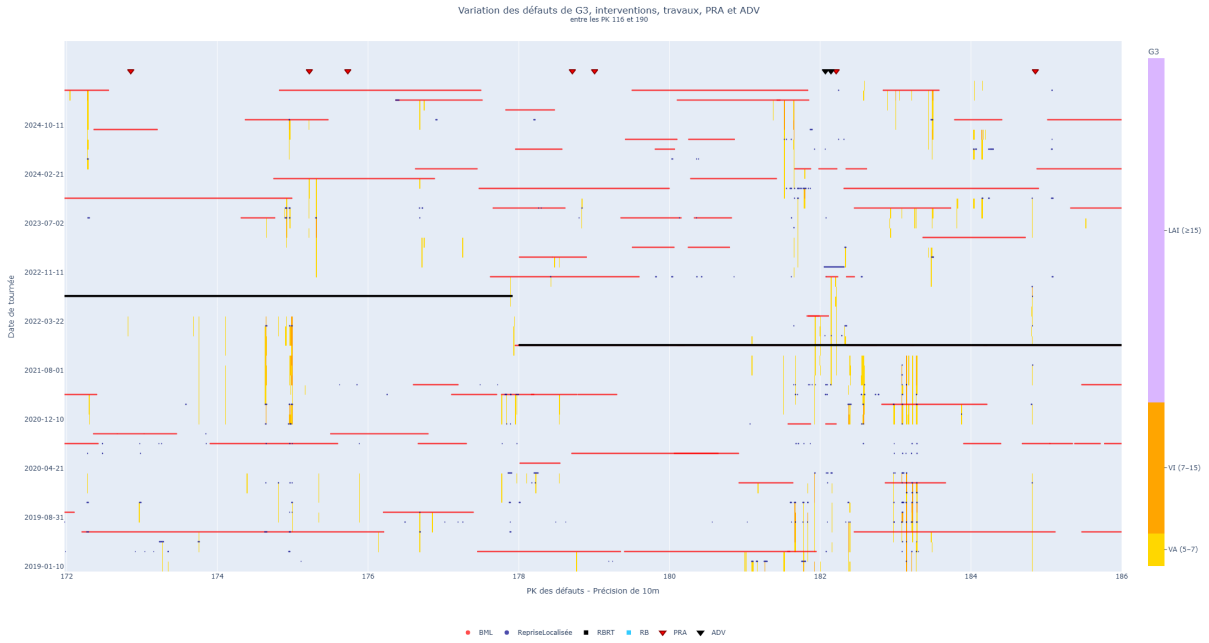

FIGURE 2 – Heatmap des dépassements de seuils pour G3, ligne 752000, voie V1, PK [172-186]

FIGURE 3 – Heatmap des dépassements de seuils pour Nall, ligne 752000, voie V2, PK [172-186]

Résultats

Les cartes de dépassements de seuils permettent d'identifier visuellement :

— Les **zones chroniques** : tronçons présentant des dépassements récurrents de seuils sur plusieurs années, suggérant des problématiques structurelles de la plateforme

— L'**efficacité des interventions et travaux** : on observe généralement une diminution des valeurs de défauts (passage du violet/orange au gris) immédiatement après une intervention de type BML ou Reprise Localisée, mais les défauts tendent à revenir rapidement. Les travaux, quant à eux, permettent de réduire les défauts à plus long terme.
— Les **patterns saisonniers** : certaines zones montrent des dépassements cycliques, potentiellement liés aux variations climatiques (retrait-gonflement des argiles) Cela est particulièrement remarquable lorsque de nombreux défauts semblent apparaître de manière simultanée[3].

La visualisation met également en évidence la corrélation entre la présence d'ouvrages d'art et l'apparition de défauts. Les zones proches des PRA (ponts-rails) et ADV (appareils de voie) présentent fréquemment des concentrations de défauts plus élevées, cohérentes avec les contraintes mécaniques accrues à ces emplacements. On pourra se référer à l'annexe B.1.6 pour observer des exemples de cartes plus rapprochés.

Limites et perspectives

Cette approche présente plusieurs limites méthodologiques :

— **Dépendance aux seuils réglementaires** : la discrétisation en catégories de seuils masque les variations fines au sein d'une même catégorie. Deux zones en "orange" peuvent avoir des dynamiques d'évolution très différentes.
— **Absence de quantification de la tendance** : la carte montre l'état à chaque date mais ne calcule pas explicitement la vitesse d'évolution des défauts.
— **Problématique de synchronisation fine PK** : comme évoqué dans la section 4.1, les décalages de PK entre tournées sur LGV sont petits, mais sur une échelle d'une dizaine de mètres, ils ne sont pas négligeables, et peuvent créer des artefacts visuels. La section suivante propose une approche pour corriger ces décalages, mais elle n'est pas encore assez précise et adaptée pour une résolution de quelques mètres.

Les perspectives d'amélioration incluent :

— L'intégration d'une métrique de tendance temporelle pour chaque cellule, permettant de distinguer les zones stables des zones à évolution rapide
— Le développement d'une version d'avantage interactive permettant de cumuler les dépassements de seuils sur une période donnée, offrant une lecture de la criticité d'un tronçon
— La correction plus fine et systématique des décalages PK entre tournées

5.2.3 Une seconde représentation : Colormap

Cette seconde approche se fonde sur l'hypothèse que la variabilité spatiale des défauts géométriques est un indicateur de l'instabilité de la plateforme ferroviaire. Plus l'écart-type des mesures est élevé sur un tronçon, plus il y a de probabilité que la plateforme sous-jacente présente des hétérogénéités (composition du sol, problèmes de drainage, mouvement de terrain).

3. Par ailleurs, un défaut ne devrait pas disparaître sans intervention, c'est un fait empiriquement observé.

Données et pré-traitement

Les données proviennent de la base Ultrapot, contenant les enregistrements tous les 25 cm des défauts géométriques longs (Nall, Dall, Tall) et leurs variantes en dépassement de seuil (Nall_seuil, Dall_seuil, TallSeuil). Pour chaque ligne ℓ et voie v, le pipeline d'ingestion (fonction `format_defauts`) :

1. Extrait la date depuis le nom du fichier (format JJMMAAAA)
2. Filtre les données pour la ligne et voie ciblées
3. Convertit les PK en kilomètres (division par 1000)
4. Stocke les données au format Parquet partitionné par date pour un accès performant

Le pré-traitement comprend une étape critique de **synchronisation des tournées** (cf. classe `RemoveLag`). Cette étape corrige les décalages de PK entre différentes dates d'acquisition, causés par les imprécisions GPS et les erreurs de projection. La méthode utilise la corrélation croisée normalisée (ZNCC) entre une tournée de référence et chaque tournée à synchroniser :

— Découpage en bins de 1 km avec chevauchement
— Calcul de la corrélation croisée pour les signaux Nall et Dall sur chaque bin
— Estimation du décalage optimal k^* (lag) par maximisation de la corrélation
— Application d'un profil de lag interpolé linéairement le long du PK

Cette synchronisation est essentielle pour éviter les faux positifs dans le calcul de la variabilité.

Méthodologie

Le calcul de l'écart-type spatial repose sur une fenêtre glissante de largeur $\Delta x_{fenetre} = 250\,\text{m}$ (paramètre `FENETRE` dans `Config`), se déplaçant avec un pas de 25 m (`PAS`). Pour chaque point central p_c et chaque date t, on définit la fenêtre $W(p_c) = [p_c - 125, p_c + 125)$ et on calcule :

$$\sigma_d(p_c, t) = \sqrt{\frac{1}{N_w - 1} \sum_{p \in W(p_c)} \left(X_d(p,t) - \bar{X}_d(W(p_c), t)\right)^2}$$

où N_w est le nombre de points de mesure dans la fenêtre et $\bar{X}_d(W(p_c), t)$ est la moyenne du défaut d dans la fenêtre.

Un filtrage préalable élimine les valeurs extrêmes : seules les mesures $X_d \leq Q_{d,99}$ sont conservées (cela supprime largement toutes les LAI et surtout les erreurs de mesure). Les fenêtres contenant moins de 200 points (20% des mesures attendues) sont ignorées pour éviter des calculs non significatifs.

La heatmap résultante affiche Σ_{norm} ou σ_d en fonction du PK (axe x) et de la date (axe y), avec une échelle de couleur continue (palette Turbo de Plotly). Les interventions (BML, Reprise Localisée) et travaux (RVB, RBRT, RB) sont superposés sous forme de segments horizontaux, tandis que les ouvrages (PN, PRA, ADV) sont affichés au-dessus de la carte.

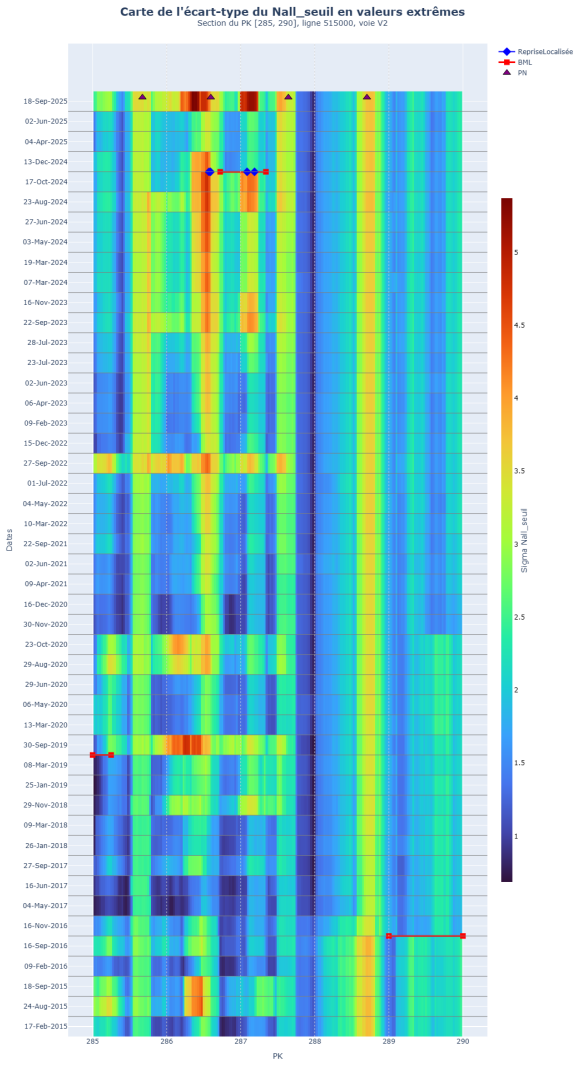

Figure 4 - Colormap pour Nall, ligne 515000, voie V2, PK [285-290]

Résultats

Les cartes d'écart-type spatial révèlent des **signatures caractéristiques** des problématiques de plateforme :

- **Zones à forte variabilité persistante** : tronçons présentant des valeurs de $\sigma >$ 5 mm de manière chronique, souvent corrélés avec des profils en déblai et des sols argileux
- **Pics localisés** : écarts-types très élevés sur quelques dizaines de mètres, généralement associés à la présence d'appareils de voie (ADV) ou de transitions de structure (PRA)
- **Évolution temporelle** : diminution nette de la variabilité après travaux de type RVB (Renouvellement de Voie et Ballast), confirmant l'efficacité de ces interventions lourdes

Limites et perspectives

Cette approche présente plusieurs limites :

- **Sensibilité à la taille de fenêtre** : le choix de $\Delta x_{fenetre} = 250\,\text{m}$ est empirique. Des fenêtres plus petites augmentent la résolution spatiale mais réduisent la robustesse statistique ; des fenêtres plus grandes lissent excessivement les variations locales.
- **Hypothèse de stationnarité locale** : le calcul d'écart-type suppose que les mécanismes physiques sont homogènes sur la fenêtre, ce qui peut être violé en présence de transitions abruptes de structure (par exemple en présence de PRA).
- **Synchronisation imparfaite** : malgré la correction des lags, des incertitudes de quelques mètres subsistent, notamment sur les lignes classiques. Nous avons implémenté une approche itérative permettant d'affiner la synchronisation en pondérant les tournées par leur proximité temporelle à une date de référence, mais cela à tendance à significativement lisser les résultats.
- **Interprétation physique** : un écart-type élevé peut avoir plusieurs causes (instabilité de plateforme, hétérogénéité du sol, défauts ponctuels de rail). Une validation terrain systématique serait nécessaire pour calibrer les seuils d'alerte, il faudrait donc coupler la

méthode avec l'analyse d'autres paramètres.

Les perspectives d'amélioration incluent :

— **Caractérisation des patterns saisonniers** : décomposition en série temporelle (trend, saisonnalité, résidus) pour isoler les variations cycliques liées au climat

— **Corrélation avec données environnementales** : intégration des données TOUTATIS (pluviométrie) et MNT (pente du talus) pour enrichir l'analyse causale

— **Industrialisation** : automatisation du pipeline complet (ingestion → synchronisation → calcul → visualisation) et déploiement sur les plateformes de production de la SNCF

5.3 Commentaires personnels

Tout comme d'autres indicateurs qui ont été développés au cours de cette mission, ils ont pour vocation de répondre à une problématique spécifique des clients.

Certaines décisions prises dans ces projets sont difficilement justifiables, il s'agit parfois d'idées évoquées durant des réunions entre des experts du domaine, mais il peut être difficile de trouver des données tangibles pour les justifier.

Le projet ZER ML est probablement une exception, c'est un projet qui a été démarré d'initiative personnelle, et supporté par mon client. Cependant, le temps alloué à ce projet était limité sur une période de 3 semaines, car d'autres obligations m'incombaient.

Bien que je sois conscient de l'opportunité qu'un travail en totale autonomie apporte, permettant de développer probablement plus que beaucoup d'autres expériences, des qualités de travail fortement valorisées, cela constitue tout de même un handicap dans la conception d'un projet complet et innovant.

6 Conclusion et Perspectives

Ce stage au sein de la division PGRN de SNCF Réseau s'est articulé autour d'une problématique centrale : **comment transformer des données ferroviaires hétérogènes en indicateurs exploitables pour anticiper les zones sensibles et orienter les interventions de maintenance ?** Mené en totale autonomie et en collaboration avec de multiples équipes aux besoins variés, ce travail a permis de développer des outils pragmatiques alliant rigueur méthodologique et utilité opérationnelle.

Les travaux réalisés se structurent autour de quatre axes complémentaires. Premièrement, le développement d'un pipeline d'intégration multi-sources harmonisant des données localisées par LIGNE/PK/DATE/VOIE, incluant une synchronisation des PK entre tournées pour corriger les décalages spatiaux. Deuxièmement, la formalisation mathématique définissant un cadre théorique basé sur la notion de segment $s = (\ell, v, p, \Delta x)$ et modélisant l'apparition de ZER comme un processus stochastique binaire $Y_s(t) \sim \mathcal{B}(p_s(t))$. Troisièmement, la création d'un modèle Random Forest atteignant 90,8% de recall dans la prédiction de Zones à Évolution Rapide, permettant de prioriser les zones de surveillance. Enfin, le développement de visualisations spatio-temporelles sous forme de cartes de dépassements de seuils et de colormaps d'écart-type spatial, révélant les zones chroniques, l'efficacité des interventions et des patterns saisonniers.

L'analyse SHAP confirme que les défauts géométriques temporels et l'historique d'interventions dominent les prédictions, validant l'hypothèse que l'évolution récente des défauts constitue un indicateur pertinent. Les visualisations soulignent clairement l'impact limité des interventions BML (améliorations court terme), indiquant la nécessité probable de travaux de plus grande envergure comme les RBRT (améliorations à plus long terme), ou la présence de problématiques récurrentes liées à la géologie du sol.

Malgré ces résultats encourageants, plusieurs limites doivent être soulignées. La qualité et la disponibilité des données présentent une couverture inégale du réseau, avec des formats inconsistants et des données lacunaires ; les bases PANDA et TSP ne couvrent que des sites spécifiques, limitant la généralisation nationale. La synchronisation des PK demeure imparfaite, avec des incertitudes résiduelles de quelques mètres sur les lignes classiques créant des artefacts dans les visualisations haute résolution. Le déséquilibre extrême des segments requérant une surveillance particulière (seulement 0,4% du réseau pour les ZER) nécessitera un ajustement du seuil de décision lors du déploiement opérationnel pour limiter les fausses alertes. Le modèle actuel traite chaque segment indépendamment sans modéliser explicitement la dynamique d'évolution, alors qu'une approche par séries temporelles (LSTM, Transformers) pourrait prédire quand une ZER apparaîtra. Enfin, les contraintes organisationnelles se sont manifestées par un temps limité par projet, l'absence d'infrastructure d'extraction automatique et une validation terrain limitée.

Les développements futurs s'articulent autour de trois horizons temporels. À court terme, il serait souhaitable d'implémenter une optimisation adaptative du seuil de décision selon le

contexte, de renforcer l'explicabilité via SHAP et LIME, d'améliorer la synchronisation PK et de réaliser une validation croisée sur plusieurs infrapôles. À moyen terme, l'intégration de séries temporelles pour modéliser la dynamique temporelle, la fusion avec données environnementales (TOUTATIS, MNT), le développement d'une modélisation spatio-temporelle jointe (GNN, CRF), la détection automatique de patterns saisonniers et l'extension à d'autres zones à risque constitueraient des améliorations significatives. À long terme, l'automatisation du pipeline de bout en bout, la création d'un dashboard web interactif, la mise en place d'un système d'alertes proactif, l'intégration dans les outils existants (Ultrapot), l'établissement d'une boucle de rétroaction opérationnelle et le déploiement multi-division (caténaires, signalisation, ouvrages d'art) permettraient une industrialisation complète.

Ce stage a constitué une expérience formatrice, confrontant mes connaissances académiques à la complexité du monde réel : données imparfaites, contraintes de temps, équilibre délicat entre réactivité opérationnelle et rigueur scientifique. J'ai découvert la difficulté que représente la réponse à des besoins opérationnels concrets, l'importance de la compréhension par les utilisateurs finaux, et les défis du maintien dans la durée de solutions techniques. Les résultats obtenus — 90,8% de recall sur la prédiction de ZER, visualisations révélant des patterns saisonniers, formalisation mathématique rigoureuse — constituent une base solide pour des développements futurs et s'inscrivent dans la transformation digitale de SNCF Réseau visant à optimiser la maintenance et améliorer la sécurité du réseau ferroviaire.

Références

[1] Scikit-learn developers. *Logistic Regression — scikit-learn documentation*.
Logistic Regression Documentation

[2] Scikit-learn developers. *Decision Trees — scikit-learn documentation*.
Decision Trees Documentation

[3] Scikit-learn developers. *Ensemble methods : Random Forests — scikit-learn documentation*.
Random Forests Documentation

[4] Chen, T., & Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System. arXiv :1603.02754 [cs.LG].
XGBoost Paper

[5] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(1), 281–284.
Hyperopt Paper

[6] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
SHapely Additive exPlanations Paper

A Glossaire et définitions

A.1 Géométrie de voie

A.1.1 Défauts courts

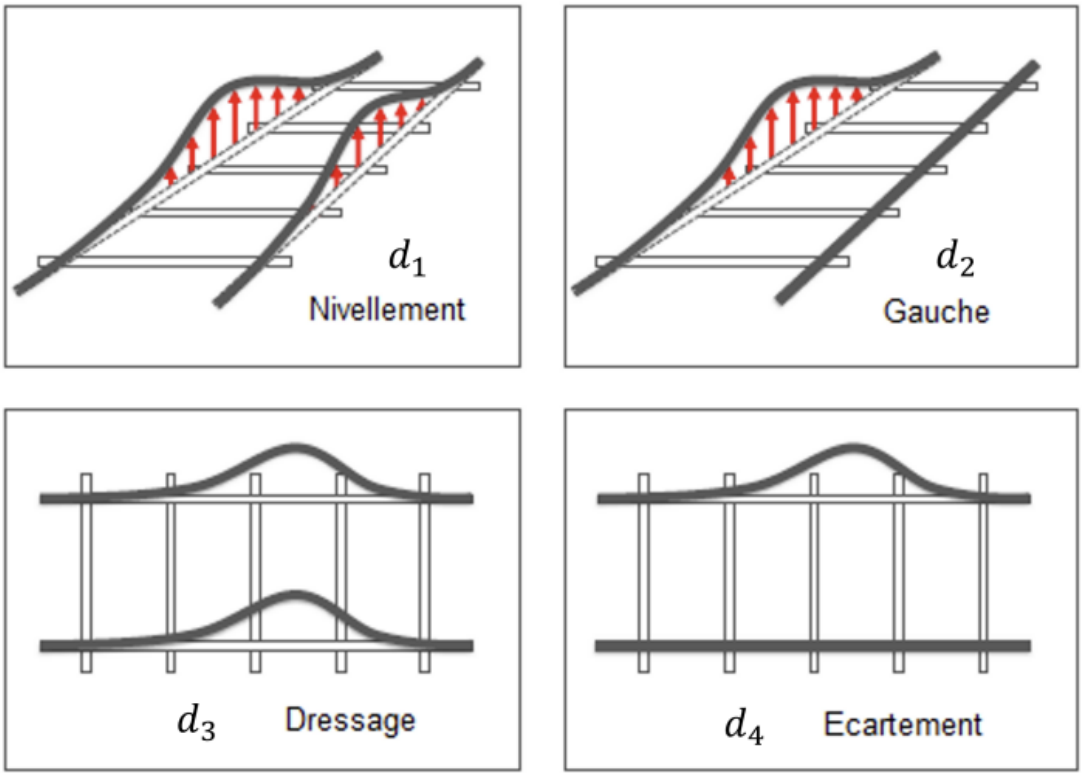

FIGURE 4 – Classification des défauts de géométrie de la voie

Ce sont les principales classes de défauts de géométrie de la voie, et celles les plus courantes. Il existe des variantes de ces défauts basées sur des longueurs de voies différentes, ou des calculs d'écart-type et de moyennes.

A.1.2 Défauts longs

En plus des défauts courts ci-dessus, les défauts longs les plus couramment utilisés sont :

— Nivellement Allongé (Nall), valeur mesurée du nivellement crête à crête sur une base de 20 à 30 mètres.

— Dressage Allongé (Dall), valeur mesurée du dressage crête à crête sur une base de 20 à 30 mètres.

— Nivellement Transversal Allongé (Tall), valeur mesurée entre la ligne de foi et la crête du défaut sur une base d'environ 30 mètres.

A.1.3 Seuils

Les seuils sont des valeurs limites, il faut maintenir les valeurs de défauts sous ces seuils.

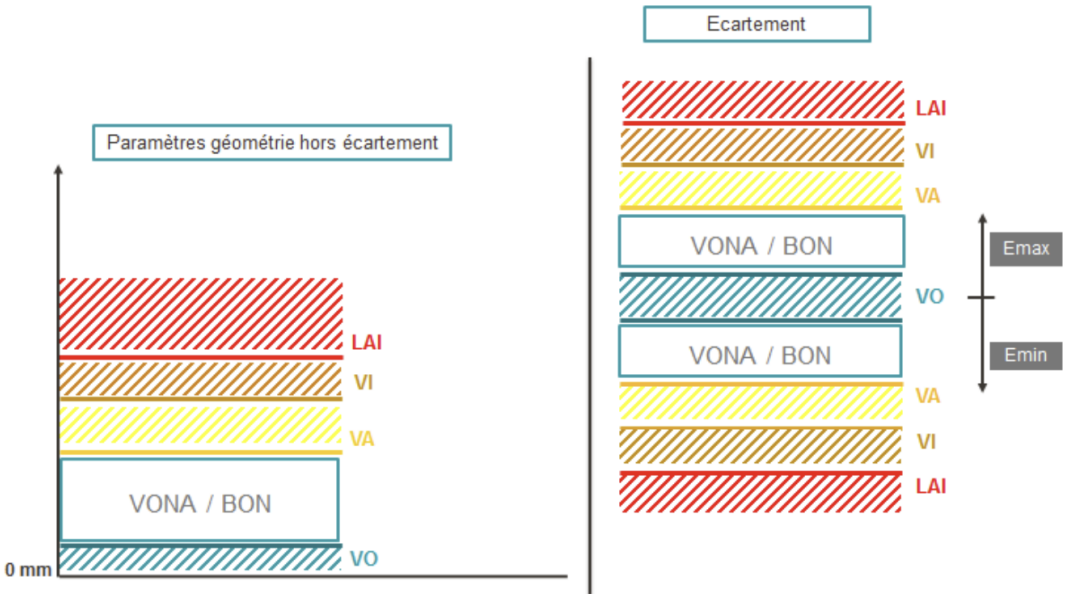

FIGURE 5 – Illustration des seuils des défauts de géométrie de la voie

On considère toujours les 4 seuils suivants :

— Valeur d'objectif (VO), intervalle de tolérance de réalisation des travaux de pose de voie neuve ou d'opérations de maintenance.
— Valeur d'alerte (VA), seuil d'altération au-delà duquel il y a lieu d'exercer une surveillance particulière des défauts.
— Valeur d'intervention (VI), seuil au-delà duquel il est nécessaire de prévoir une intervention de correction à court terme afin d'éviter d'atteindre les limites action immédiate
— Limite d'action immédiate (LAI), seuil qui requiert la prise de mesures immédiates pour maintenir les performances des voies et garantir la sécurité des circulations. Cela peut être réalisé par la fermeture de la ligne, ou la réduction de la vitesse ou la correction immédiate de la géométrie de la voie (avant toute circulation) ou tout autre processus spécifique.

A.1.4 Exemples de défauts

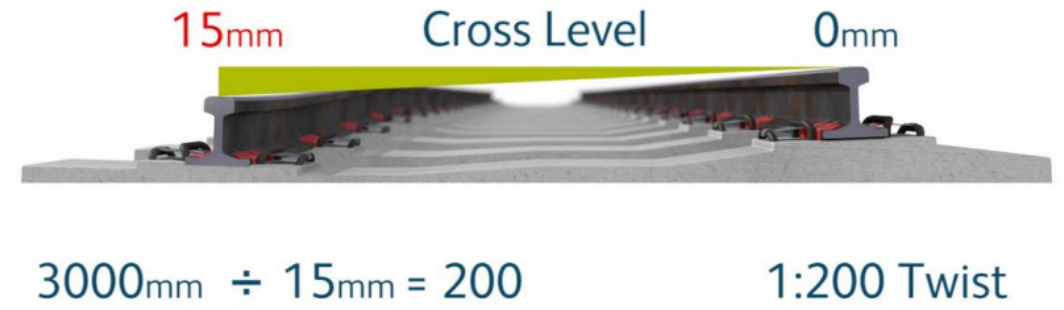

FIGURE 6 – Défaut de gauche 3m (G3)

Illustration d'un défaut de gauche 3m extrait d'une présentation sur les appareils d'embarquements de mesures de la géométrie de voie par nos homologues du royaume-uni.

A.2 Les zones à évolution rapide

Le processus de classement d'une zone en ZER suit les étapes d'un arbre logique, il ne résulte pas d'une prédiction de l'évolution future des défauts.

Classement en ZER

— Toute zone répondant à la définition d'une ZER (cf. 3.1.2) doit être classée comme telle.
— Le classement intervient notamment lorsqu'un dépassement de seuil apparaît soudainement, sans signe avant-coureur dans les enregistrements précédents.
— Ne sont pas classées en ZER les zones où la cause du défaut est clairement liée au matériel de voie (rail, traverses. . .), et non au ballast, à la plate-forme ou au terrain sous-jacent.
— Le classement ne s'applique pas à l'écartement de voie.
— Toute zone classée en ZER doit être suivie par une surveillance systématique et donner lieu à des interventions de maintenance jusqu'à suppression du défaut.

FIGURE 7 – Arbre logique de classement en ZER

Suivi des ZER

— Le suivi vise à s'assurer que les défauts sont détectés et corrigés avant d'atteindre la limite admissible (LAI).
— Le dirigeant local définit la procédure de suivi en fonction des historiques et des conditions locales (météo, nappe phréatique, drainage. . .).
— Un historique des reprises de nivellement doit être tenu, incluant la localisation précise, la vitesse d'évolution des défauts (stable, croissante), et les facteurs locaux influents.

Délais et traitements

— Le délai maximum de correction d'un défaut en ZER correspond au délai "ZER" fixé dans les référentiels.
— En cas de forte dégradation, les contrôles et corrections doivent être renforcés ; à l'inverse, les délais peuvent être assouplis si les défauts évoluent lentement.

Sortie de ZER

— Une zone peut être déclassée si :
— Les causes de dégradation ont été éliminées,
— Et qu'aucune évolution notable n'est observée entre deux mesures espacées d'au moins 3 mois.
— Si les désordres se stabilisent naturellement et que leur évolution reste maîtrisable, la zone peut également sortir du suivi ZER (après 6 mois sans évolution notable).

A.3 Géométrie du plateforme

A.3.1 Le talus (remblai ou déblai)

La principale mission de notre division étant la surveillance et la maintenance de la plateforme ferroviaire, nous avons besoin de connaître la géométrie du talus (remblai ou déblai).

En effet, plus de deux tiers des accidents relatifs à la plateforme ont lieux sur des configurations en déblai. Ces accidents sont pour la plupart liés à des problématiques hydrauliques, à une composition du sol argileux, et à des mauvaises conditions de drainage qui provoquent des mouvements de terrain importants.

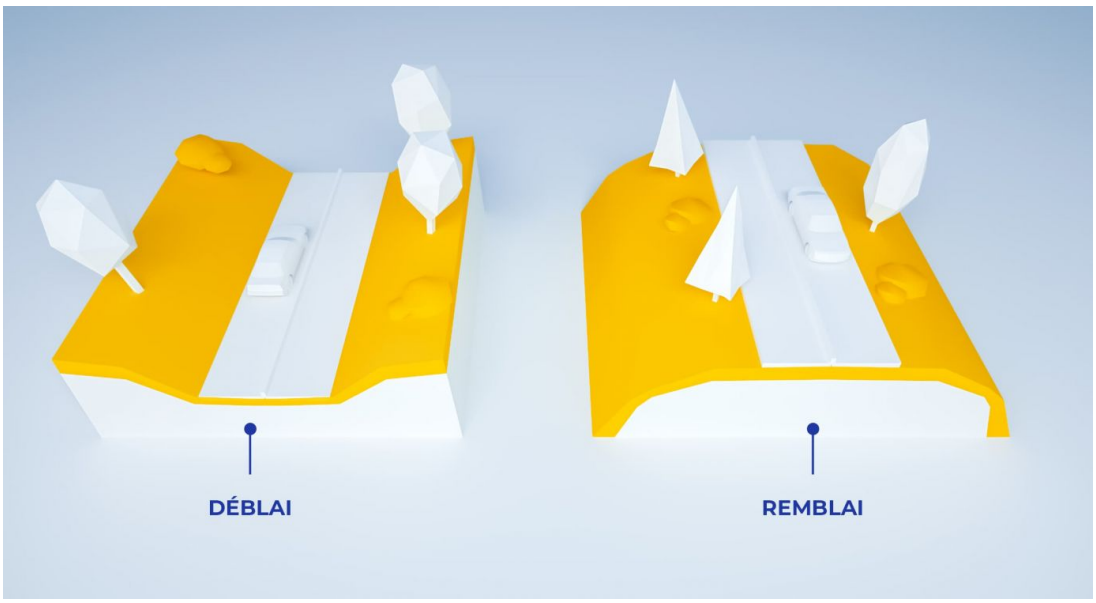

FIGURE 8 – Configurations remblai et déblai du talus

B Résultats annexes

B.1 ZER ML

B.1.1 Comparaison des résultats brutes contre sélection des features

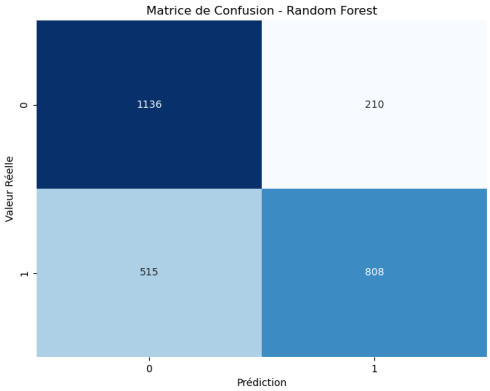

FIGURE 9 – Matrice de confusion - Résultats bruts

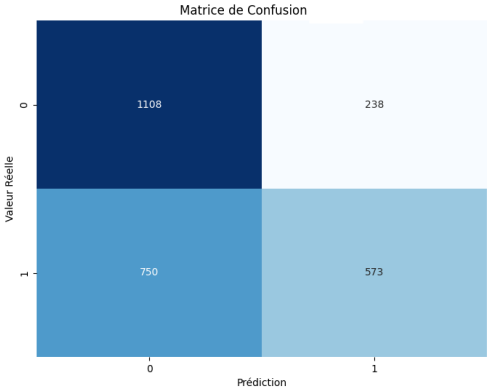

FIGURE 10 – Matrice de confusion - Sélection 30 features

B.1.2 Hyperparamètres optimisés

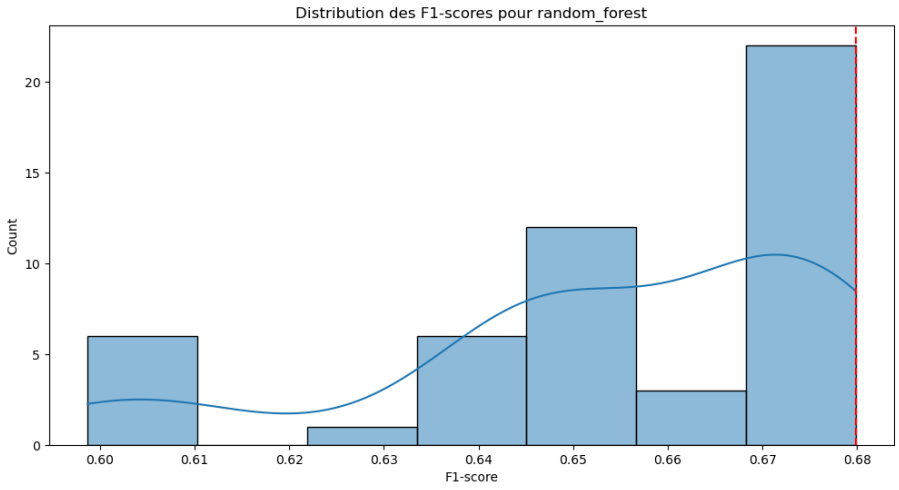

FIGURE 11 – Distribution des scores F1-score pour le Random Forest optimisé

B.1.3 Importance des features

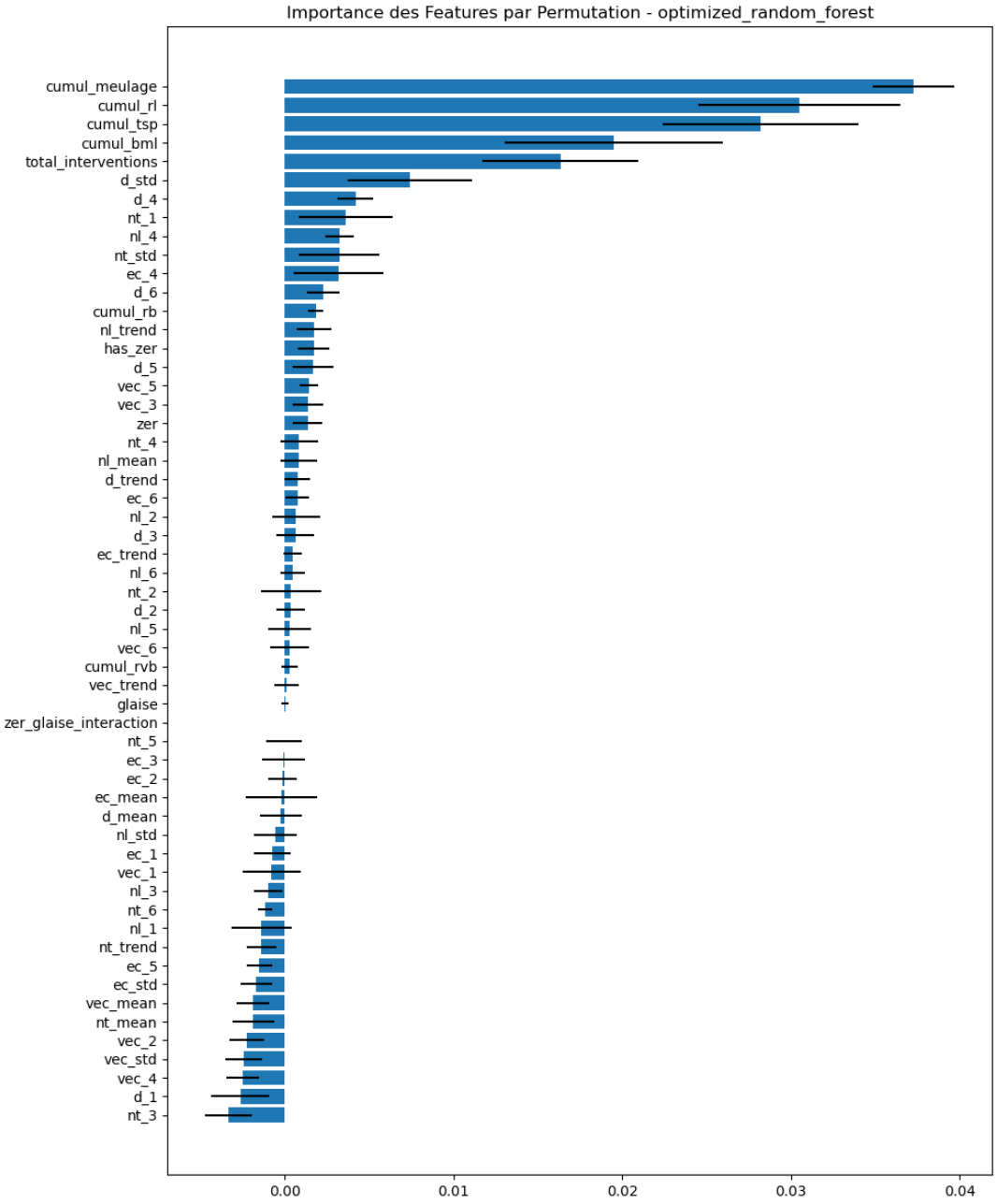

FIGURE 12 – Importance des features pour le Random Forest optimisé

B.1.4 Ajout des dépassements de seuils

Nous avons continué à travailler sur ce projet hors période de stage, et en implémentant l'**historique des dépassements de seuils** : pour chaque segment, comptage du nombre de défauts dépassants chaque seuil (VA, VI, LAI) au cours des 3 dernières années ; nous avons grandement amélioré les performances du modèle.

La matrice de confusion est la suivante :

		Prédiction	
		Non-ZER (0)	**ZER (1)**
Réalité	**Non-ZER (0)**	89,8%	10,2%
	ZER (1)	9,2%	**90,8%**

TABLE 4 – Matrice de confusion du modèle avec historique des dépassements de seuils

Ces résultats marquent une amélioration significative par rapport au modèle initial. Le tableau suivant compare les performances des trois versions du modèle :

		Prédiction	
		Non-ZER (0)	**ZER (1)**
Réalité	**Non-ZER (0)**	84,4%	15,6%
	ZER (1)	38,9%	**61,1%**

TABLE 5 – Matrice de confusion du modèle avec historique des dépassements de seuils

L'ajout de l'historique des dépassements de seuils a permis d'améliorer drastiquement toutes les métriques, particulièrement le recall qui passe de 61,1% à 90,8%, réduisant ainsi significativement le nombre de zones à risque non détectées.

B.1.5 Cartes de chaleurs

B.1.6 Cartes de dépassements de seuils

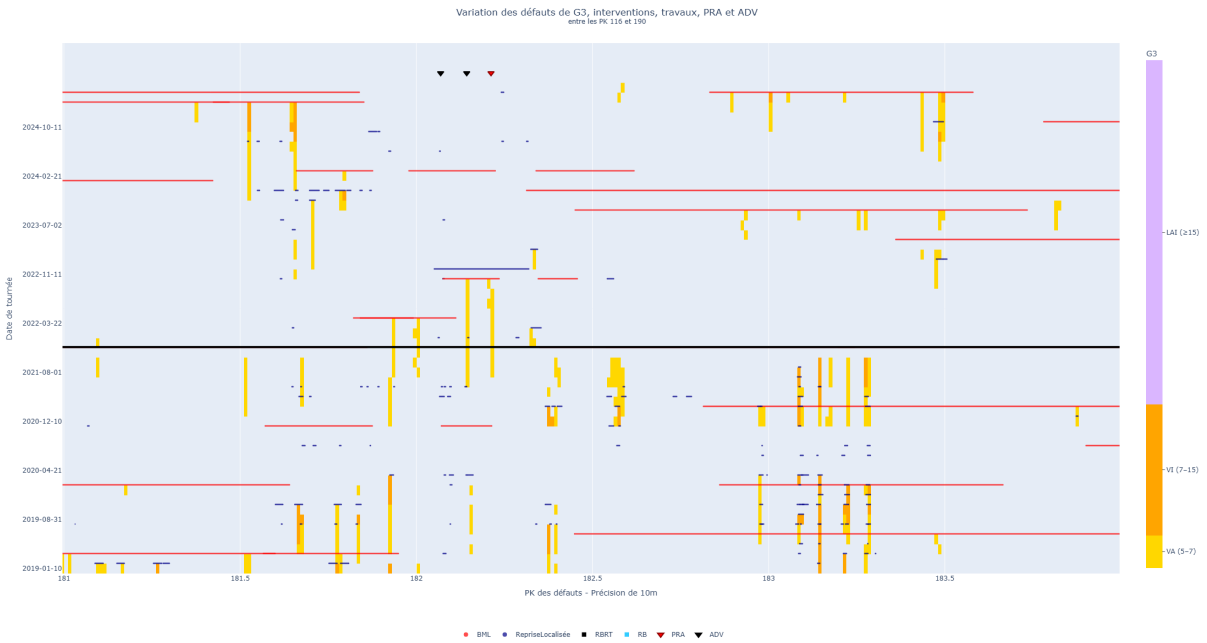

FIGURE 13 – Cartes de dépassements de seuils pour G3, ligne 752000, voie V1, PK [181-184]

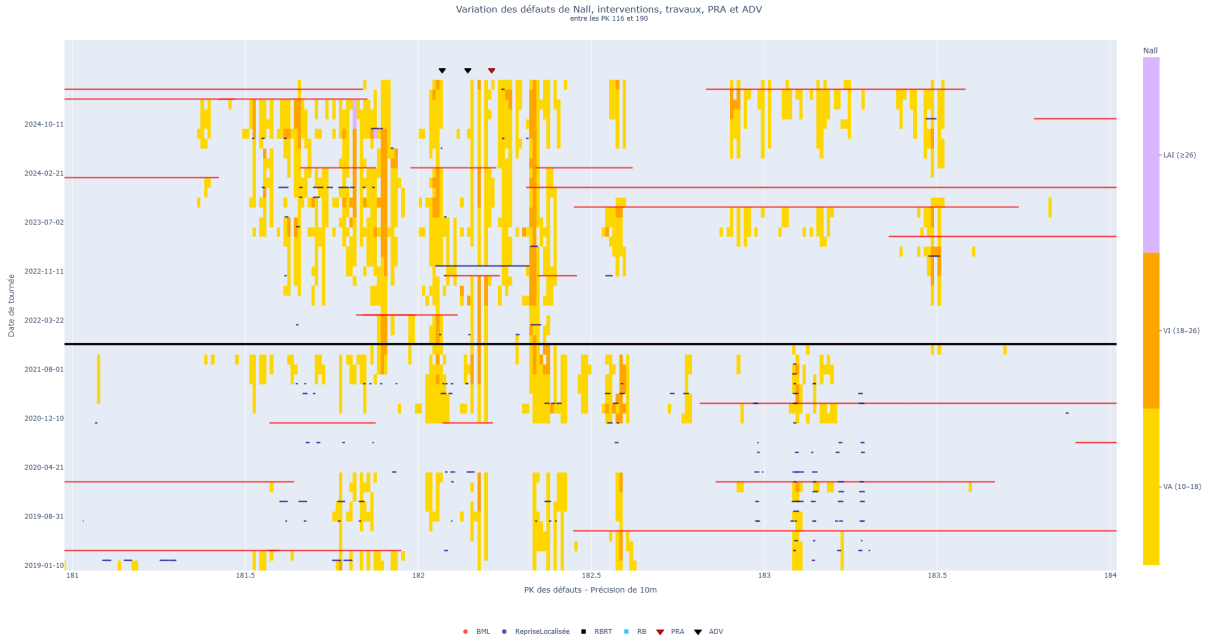

FIGURE 14 – Cartes de dépassements de seuils pour Nall, ligne 752000, voie V2, PK [181-184]

Sur ce dernier graphique du Nall, il apparait que même suite à des interventions ou des travaux, de nombreux défauts ne sont pas corrigés, d'autant plus lorsqu'il y a de PRA (ponts-rails) et ADV (appareils de voie).

Cela révèle aussi un caractère inconsitent de la qualité d'une intervention. On ne peut pas considérer que suite à celles-ci, les défauts seront corrigés, où leur croissance sera ralentie. Un risque de ZER ressurgissant rapidement n'est pas exclu.

B.1.7 Colormap

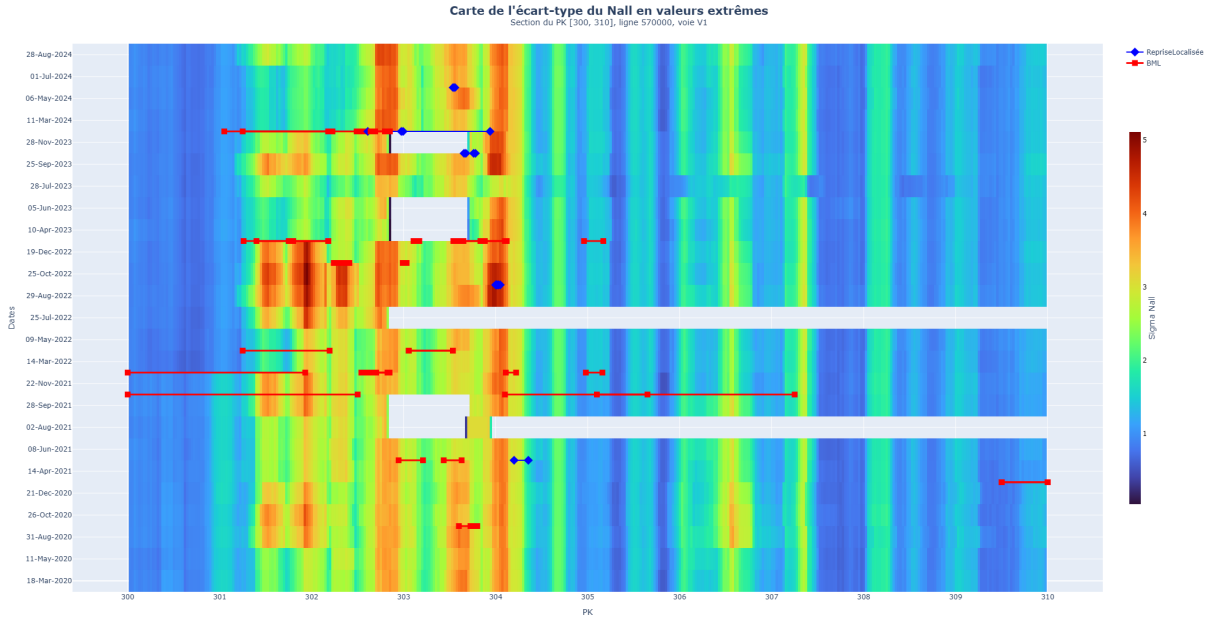

FIGURE 15 – Colormap pour Nall, ligne 570000, voie V1, PK [300-310] avec synchronisation

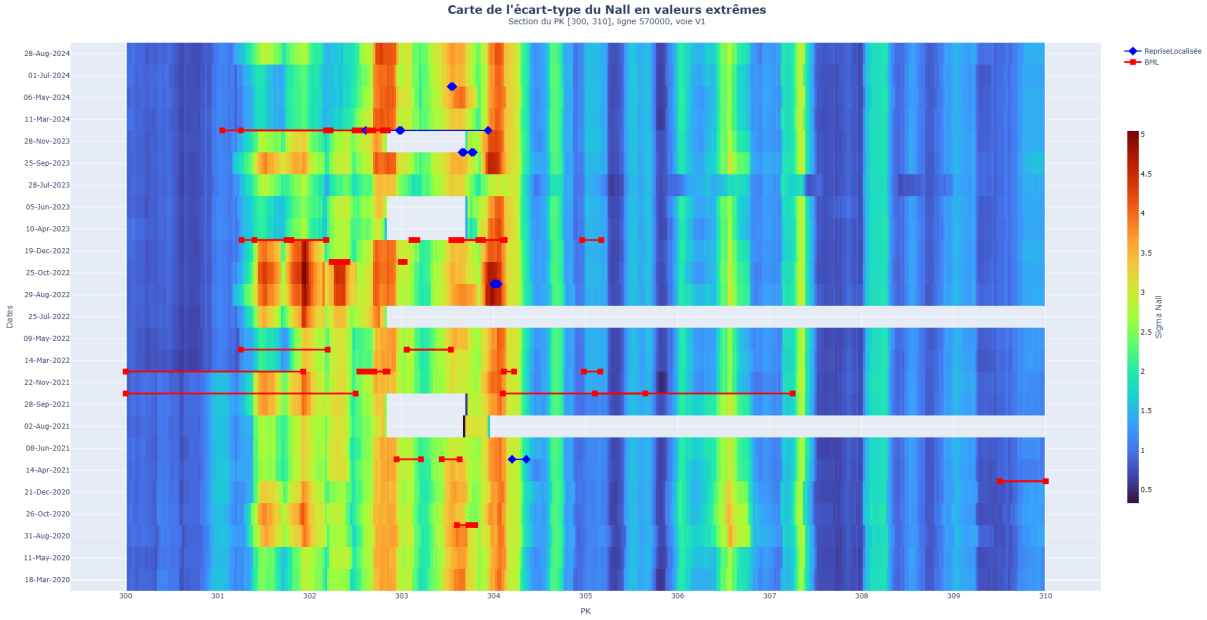

FIGURE 16 – Colormap pour Nall, ligne 570000, voie V1, PK [300-310] sans synchronisation

La différence entre les deux colormaps est assez légère, car les décalages en PK sur le graphique original (sans synchronisation) sont déjà assez faibles.

Note de Synthèse

Ce stage de fin d'études s'est déroulé au sein de la division PGRN de SNCF Réseau, en mission pour Aubay Data & IA, entre janvier et juin 2025. La division PGRN est spécialisée dans l'analyse des problématiques liées à la plateforme ferroviaire, notamment les aspects hydrauliques, géotechniques et de stabilité du sol. La mission principale consistait à développer des méthodes de traitement et d'analyse de données pour transformer des sources hétérogènes en indicateurs exploitables permettant d'anticiper les zones sensibles et d'orienter les interventions de maintenance.

Le réseau ferroviaire national génère d'importants volumes de données : mesures géométriques des voies tous les 25 centimètres sur l'ensemble du réseau ferroviaire, signalements de zones à évolution rapide (ZER), interventions de maintenance, observations terrain, données météorologiques et géotechniques. Ces données, dispersées dans de multiples bases et présentant des formats inconsistants, nécessitaient une harmonisation pour permettre une lecture cohérente du réseau.

Les travaux se sont structurés autour de quatre axes principaux. Premièrement, le développement d'un programme intégrant ces multiples sources et harmonisant des données localisées par ligne ferroviaire, voie, date et position géographique des rails tout en corrigeant les décalages spatiaux causés par les imprécisions GPS. Deuxièmement, la formalisation mathématique du problème définissant un cadre théorique rigoureux basé sur la notion de segment spatial et modélisant le problème complexe d'apparition de zones à risque sur le réseau sous la forme d'un problème mathématique abordable.

Troisièmement, la création d'un modèle de machine learning pour prédire l'apparition de Zones à Évolution Rapide, qui sont des zones critiques du réseau nécessitant une surveillance accrue. Le modèle développé atteint des performances très encourageantes, permettant de prioriser efficacement les zones de surveillance. L'analyse des facteurs déterminants dans la dégradation des rails nous permet d'isoler les paramètres qui sont à la source des défaillances sur les voies, apportant une aide aux experts métiers dans la compréhension des phénomènes physiques affectant la plateforme ferroviaire.

Quatrièmement, le développement de visualisations graphiques sous la forme de cartes de chaleur révèle des zones de problèmes chroniques et permet d'évaluer l'efficacité d'interventions de maintenance. Cela met en évidence des patterns saisonniers potentiellement liés à une composition du sol spécifique qui accentue la dégradation des rails.

Les résultats obtenus démontrent qu'il est possible de transformer des données ferroviaires complexes en indicateurs exploitables pour la maintenance prédictive. Les retours des experts métiers ont été globalement positifs, et nous espérons à l'avenir intégrer ces outils dans leurs processus opérationnels. Les perspectives d'amélioration restent évidemment nombreuses. Cette mission s'inscrit dans la transformation digitale de SNCF Réseau visant à exploiter les données massives pour optimiser la maintenance et améliorer la sécurité du réseau ferroviaire national.

Executive Summary

This end-of-studies internship took place within the PGRN division of SNCF Réseau, on assignment for Aubay Data & AI, from January to June 2025. The PGRN division specializes in railway platform issues, particularly hydraulic, geotechnical, and soil stability aspects. The main mission was to develop data processing and analysis methods to transform heterogeneous sources into actionable indicators for anticipating sensitive zones and guiding maintenance interventions.

The national railway network generates substantial data volumes : track geometry measurements every 25 centimeters across the entire railway network, reports of rapidly evolving zones (ZER), maintenance interventions, field observations, meteorological and geotechnical data. This data, scattered across multiple databases with inconsistent formats, required harmonization to enable coherent network interpretation.

The work was structured around four main axes. First, the development of a program integrating these multiple sources and harmonizing data located by railway line, track, date, and geographic position of the rails while correcting spatial shifts caused by GPS inaccuracies. Second, the mathematical formalization of the problem defining a rigorous theoretical framework based on spatial segment notation and modeling the complex problem of risk zone occurrence on the network as a tractable mathematical problem.

Third, the creation of a machine learning model to predict the occurrence of Rapidly Evolving Zones, which are critical zones of the network requiring increased surveillance. The developed model achieves very encouraging performance, effectively prioritizing surveillance zones. The analysis of determining factors in rail degradation allows us to isolate the parameters that cause track failures, providing assistance to domain experts in understanding the physical phenomena affecting the railway platform.

Fourth, the development of graphical visualizations in the form of heatmaps reveals chronic problem zones and enables evaluation of maintenance intervention effectiveness. This highlights seasonal patterns potentially linked to specific soil composition that accelerates rail degradation.

The results demonstrate that transforming complex railway data into actionable indicators for predictive maintenance is feasible. Feedback from domain experts was generally positive, and we hope to integrate these tools into their operational processes in the future.

Improvement perspectives obviously remain numerous. This mission is part of SNCF Réseau's digital transformation aiming to leverage massive data for optimizing maintenance and improving national railway network safety.