Titre: Title:	Modélisation et analyse de la non-qualité de planches en bois par apprentissage automatique
Auteur: Author:	Ons Masmoudi
Date:	2021
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Masmoudi, O. (2021). Modélisation et analyse de la non-qualité de planches en bois par apprentissage automatique [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/9912/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/9912/
Directeurs de recherche: Advisors:	Soumaya Yacout, & Mohamed-Salah Ouali
Programme: Program:	Maîtrise recherche en génie industriel

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Modélisation et analyse de la non-qualité de planches en bois par apprentissage automatique

ONS MASMOUDI

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie industriel

Décembre 2021

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Modélisation et analyse de la non-qualité de planches en bois par apprentissage automatique

présenté par **Ons MASMOUDI**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Bernard CLÉMENT, président

Soumaya YACOUT, membre et directrice de recherche

Mohamed-Salah OUALI, membre et codirecteur de recherche

Amel JAOUA, membre

DÉDICACE

À mes parents qui m'ont toujours soutenue et encouragée

À mes sœurs pour leur soutien continu

À mes chers amis,

REMERCIEMENTS

Je tiens à exprimer ma plus vive reconnaissance à ma directrice de recherche, Madame Soumaya YACOUT à qui j'ai eu le privilège d'être son élève. J'apprécie son dévouement et ses commentaires perspicaces. Elle m'a appris la méthodologie pour mener à bien la recherche et présenter les travaux de recherche aussi clairement que possible.

Toute ma gratitude va aussi à mon codirecteur de recherche, Monsieur Mohamed Salah OUALI pour son importante contribution à l'élaboration de ce travail et sa qualité d'encadrement. Je suis très reconnaissante pour son attention, ses conseils judicieux et son soutien continu inestimable au cours de mes recherches et de mes études. La pertinence de ses remarques m'a absolument aidé dans le cheminement durant mon projet de maîtrise.

De même, j'aimerais remercier les membres du jury, Monsieur Bernard CLÉMENT en tant que président du jury et Madame Amel JAOUA en tant que membre de jury, d'avoir accepté d'évaluer le présent travail, malgré leurs emplois du temps chargés. Leurs compétences et leur notoriété dans leurs domaines respectifs me font honneur.

RÉSUMÉ

L'industrie du sciage occupe une place économique significative au Québec. Pour répondre aux exigences relatives à la productivité et à la qualité dans le processus de sciage, les industries forestières doivent faire face à l'évolution rapide des technologies de l'information et de la communication. Ce projet se focalise sur les problèmes relatifs à la qualité du produit final, dans ce cas, les planches de bois.

L'objectif global de cette mémoire consiste à modéliser la non-qualité des planches lors du processus du sciage pour identifier les causes racines de la mauvaise qualité du produit. Cet objectif est subdivisé en deux sous-objectifs principaux, notamment la détection de la non-qualité dans le produit fini lors de la coupe des billes de bois et l'identification des conditions sous lesquelles les critères de non-qualité apparaissent dans les planches afin de réduire éventuellement la quantité de planches de mauvaise qualité.

Deux méthodes d'analyse sont développées et appliquées. La première méthode suivie pour atteindre le premier objectif consiste à générer des règles de classification qui représentent chaque type de non-qualité du produit fini. Pour ce faire, la technique de l'analyse logique des données est utilisée en se basant sur le logiciel cbmLAD.

Pour le second objectif, des modèles analytiques basés sur la régression à multi-sorties sont formulés pour expliquer les différents types de non-qualité trouvée dans les planches produites lors du processus de sciage. Une analyse d'importance est également effectuée sur les variables entrantes afin de détecter celles qui contribuent significativement au modèle construit. Enfin, une comparaison des résultats des modèles de classification et de régression est établie.

Un cas d'étude est utilisé tout au long du mémoire pour implémenter et évaluer les modèles de classification et de régression proposés.

ABSTRACT

The sawmill industry occupies a significant economic place in Quebec. To meet the demands for productivity and quality in the sawing process, forest industries must cope with rapid developments in information and communication technologies. This project focuses on issues relating to the quality of the final product, in this case, wood planks.

The goal of this master thesis is to model the wood planks of poor quality during the sawing process in order to identify the root causes of poor product quality. This goal is subdivided into two main sub-objectives, namely the detection of poor-quality product when cutting logs and the identification of the conditions under which the non-quality criteria appear in the boards.

Two analysis methods are developed and applied. The first method consists of generating classification rules, known as patterns, that represent each type of non-quality product. For this purpose, the technique of logical data analysis (Aguilar et al.) is used based on the cbmLAD software. In order to achieve the second sub-objective, analytical models based on multi-output regression are developed to explain the different types of non-quality. A significance analysis is also performed on the input variables to detect those that significantly contribute to the constructed model. Finally, a comparison of the results of the classification and regression models is established.

A case study is used throughout the master thesis to implement and evaluate the classification and regression models.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

BPFI : Ball Pass Frequency Inner

CNN : Convolutional neural network

FC: Facteur de Crête

FFT : « Fast Fourier Transform

GB : Gradient Boosting

k-NN : k-Nearest Neighbors

KS : Kurtosis spectral

FTF : Fundamental Train Frequency

Ku : Kurtosis

LAD : Logical Analysis of Data

MPMP : Mille Pieds Mesure de Planche

MILP : Mixed Integer Linear Programming

MTR : Multi-Target Regression

PF : Produit fini

PDF : Probability Density Function

RC : régression à chaine

RF : Random Forest

RPM : Revolutions Per Minute

RMS : Root Mean Square

SS : Spectral Skewness

ST : Single-Target

STFT : Short-Time Fourier Transform

VC : Variable contrôlable

VI : variable incontrôlable

VSS : Vertical Shape Saw

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

1.1 Contexte et cadre théorique

Grâce aux progrès réalisés dans les technologies de l'information et de la communication, les données décrivant le processus de production, l'état des équipements et les produits dans les usines industrielles deviennent de plus en plus disponibles. Avec ces données et ces technologies, il est possible de découvrir des connaissances pertinentes permettant de satisfaire les exigences grandissantes de qualité et de productivité dans les industries. Parmi les exemples d'analyse les plus fréquemment abordés dans ce domaine, nous distinguons la modélisation de la non-qualité des produits finis, de la dégradation et des défauts dans les équipements telles que l'usure des engrenages et des roulements. Cette modélisation permet d'identifier les causes primaires de la mauvaise qualité des produits et/ou de la diminution de la productivité.

L'industrie forestière doit, comme toute industrie, faire face à l'évolution rapide des technologies de l'information pour améliorer sa compétitivité par la maîtrise opérationnelle de ses activités, notamment le maintien d'une disponibilité accrue des équipements de sciage en état fonctionnel, et la gestion de la qualité du produit fini (PF). Cette industrie représente un secteur manufacturier clé dans la production du bois, occupant une place économique importante au Québec. En effet, les forêts représentent plus que la moitié du territoire de cette province, soit environ 900 000 kilomètres-carrés. Il existe de nombreuses scieries au Québec. Plusieurs d'entre elles souhaitent avoir une meilleure maîtrise de leurs processus de production et une meilleure gestion de qualité. Nous intervenons dans ce projet pour soutenir cette démarche, tout en se basant sur les technologies de l'information et de la communication.

Dans ce contexte, nous nous concentrons sur la modélisation de la non-qualité des planches de bois lors du processus du sciage pour identifier les causes primaires de la mauvaise qualité du PF. Les activités de sciages sont principalement caractérisées par les intrants contrôlables et non-contrôlables. Les dimensions de planches à produire, et le nombre d'interventions de maintenance préventive et anticipée effectuée sur l'équipement de coupe de bois sont considérés des variables contrôlables (VC) étant donné qu'elles peuvent être maîtrisées par l'opérateur de la ligne de sciage.

Parmi les variables considérées comme non-contrôlables, mesurés pendant le processus du sciage, nous trouvons la dimension des billes de bois et la qualité de la matière première MP ainsi que le nombre d'interventions de maintenance corrective effectuée sur l'équipement de coupe. Les quantités d'énergie vibratoire produites par le moteur de l'équipement de sciage sont aussi considérées comme variables non-contrôlables (VNC). Ces vibrations peuvent représenter des indicateurs importants de l'état de l'équipement, plus précisément de son état de dégradation (Ghasemi et al., 2009). La Figure 1.1 résume les activités globales du processus de sciage avec ces différents intrants.

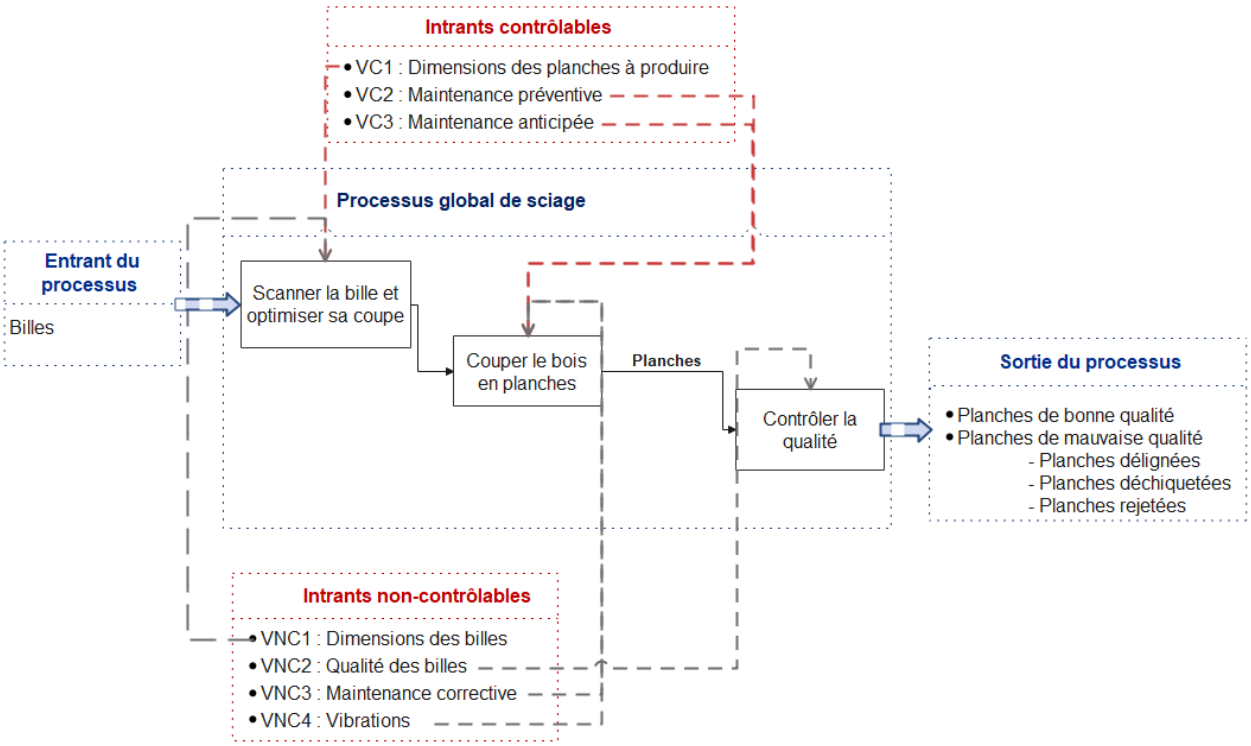


Figure 1.1 Processus global de sciage avec ses intrants

Le processus global de sciage commence par scanner la bille de bois entrante et optimiser son profil de coupe selon sa dimension et selon la dimension de planches programmées à produire. Ensuite, chaque bille est transformée en planches en passant par plusieurs équipements notamment l'équipement de débitage, appelé VSS « Vertical Shape Saw ». Parmi les variables contrôlables reliées à cette activité, le nombre d'interventions de maintenance prédictive et celui de la maintenance anticipée effectuée sur le VSS sont considérés. La maintenance anticipée est définie dans notre cas comme étant une stratégie de maintenance opportuniste basée sur le remplacement

ou la réparation de composants qui s'avèrent défectueux ou doivent être remplacés dans un avenir immédiat, en tirant parti de l'arrêt planifié ou non d'un équipement lors de sa maintenance (Ab-Samat & Kamaruddin, 2014). En maintenance préventive, les actions sont menées à intervalles périodiques en fonction d'un critère simple qui reflète l'état de l'équipement comme le temps d'utilisation ou le nombre d'opérations (Ragab, A. R. A., 2014). En revanche, le nombre d'interventions de maintenance corrective est considéré comme une variable non-contrôlable, étant donné que celle-ci dépend de l'apparition de défaillances ou de signes de défaillance potentielle qui requiert souvent une réparation imprévue.

Au niveau du contrôle de qualité, trois types de non-qualité sortante dans les planches de bois sont observés : les planches délignées, les planches déchiquetées et les planches rejetées. Ces imperfections indésirables doivent être étudiées pour trouver les causes primaires de la mauvaise qualité, afin de l'éliminer et ainsi réduire les coûts de la non-qualité et maximiser la production. Par exemple, la qualité du bois entrant peut impacter celle du produit fini. Cette variable est considérée comme non-contrôlable étant donné qu'elle dépend de facteurs environnementaux et de facteurs saisonniers tels que l'humidité, l'apparition des insectes et la saison de chasse.

Dans cette étude, une analyse profonde des problèmes liés aux caractérisations de non-qualité présentées ci-dessus sera abordée, en se basant sur les intrants. Par la suite, la production des planches de mauvaise qualité pourra être minimisée éventuellement en agissant sur les variables contrôlables. Pour ce faire, une bonne compréhension du processus complet de sciage et des types de non-qualité des planches produites est nécessaire.

1.2 Description du processus et de la non-qualité

Le processus de sciage commence par l'arrivée de bois à la scierie sous forme de billes. À l'entrée de la ligne, ces derniers sont séparés en 2 types selon leurs dimensions, des billes de diamètre proches de 17 pouces et des billes de diamètre proches de 25 pouces. La première activité du processus consiste à écorcer les billes, c'est-à-dire, enlever l'écorce des billes. Ensuite, celles-ci sont transportées sur un convoyeur et scanner par un lecteur optique pour produire un patron de la coupe optimisée des billes écorcées et ainsi minimiser les résidus de bois. Ensuite, une grande variété de produits est obtenue selon la dimension de la bille et selon le patron de coupe retenue, comme le montre la Figure 1.2.

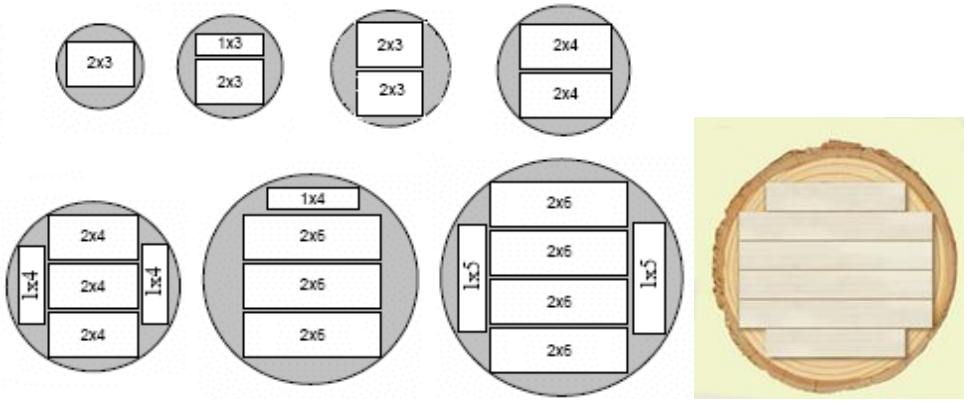


Figure 1.2 Exemples de patron de coupe de bille (Côté, 2013)

La bille sera positionnée de façon à suivre le patron pour une coupe optimale. Ensuite, les côtés de la bille sont déchiquetés pour créer deux surfaces plates parallèles sur toute sa longueur. Une fois cette activité terminée, les côtés du bois sont sciés pour obtenir jusqu'à quatre planches qui sont généralement de petites dimensions. L'activité suivante du sciage consiste à couper le bois horizontalement pour obtenir des planches de dimensions égales sur l'équipement VSS. Une fois les pièces de bois obtenues, elles sont éboutées pour éliminer les défauts, tout en tenant compte des dimensions en pouces x pouces correspondant aux normes établies (1 x 3, 1 x 4, 1 x 6, 2 x 3, 2 x 4, 2 x 6, 2 x 8, 2 x 10, 3 x 3 et 5 x 4). La Figure 1.3 cartographie le processus de sciage et de contrôle qualité après la coupe.

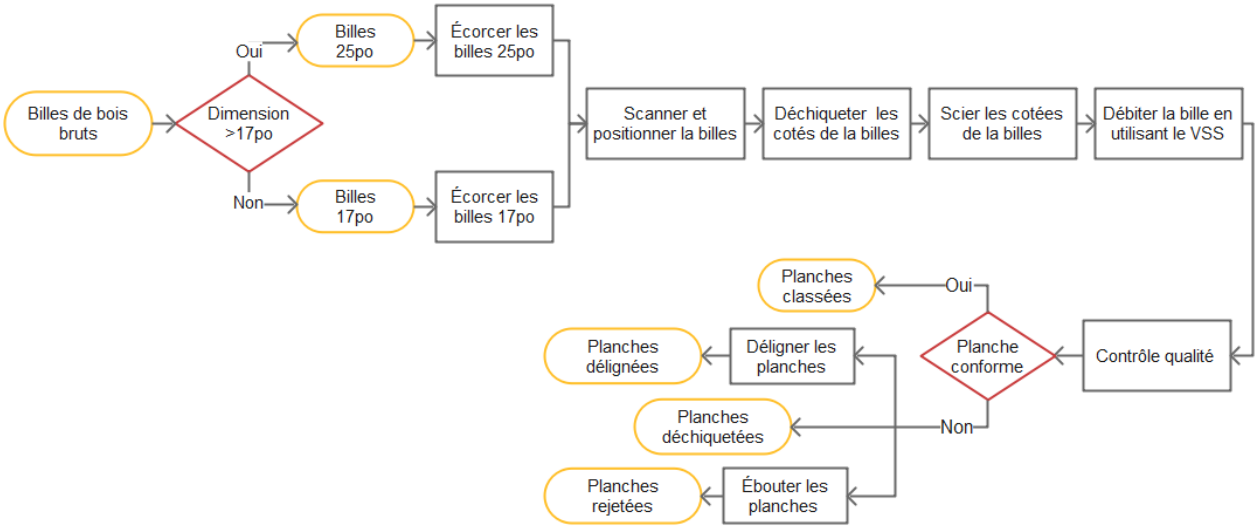


Figure 1.3 Processus de sciage et de contrôle qualité

Le contrôle qualité est effectué juste après que le bois a été coupé par l'équipement VSS. Les planches de bonne qualité sont classées suivant leurs dimensions et sont dirigées vers une empileuse automatique. Le volume des planches classées est mesuré en mille pieds mesure de planche (MPMP) (Parent, 2010) . Le PMP est une mesure équivalente à un pied carré de surface par un pouce d'épaisseur. Les planches de mauvaise qualité, quant à elles, seront soit retravaillées, soit déchiquetées, introduisant ainsi trois types de non-qualité qui sont décrits et analysés dans les 3 sous-sections suivantes.

1.2.1 Planches délignées

Parmi les caractéristiques de non-qualité pouvant être présentes sur la ligne de sciage, le délignement des planches est le plus fréquent. Ces planches se distinguent par la présence des flaches de rives qui sont considérées des défauts de type géométrique. En effet, les flaches de rive représentent des côtés longitudinaux arrondis de la planche avec ou sans écorce, restant apparentes même après le sciage (Côté, 2013). Les pièces délignées doivent être donc réusinées en largeur par la déligneuse pour respecter la qualité et les dimensions normalisées, comme le montre la Figure 1.4. Par exemple, une pièce 1po x 6po de longueur 16po, ayant des flaches de rive devrait être retournée à la déligneuse pour être retransformée en une pièce de dimension 1po x 4po 16po.

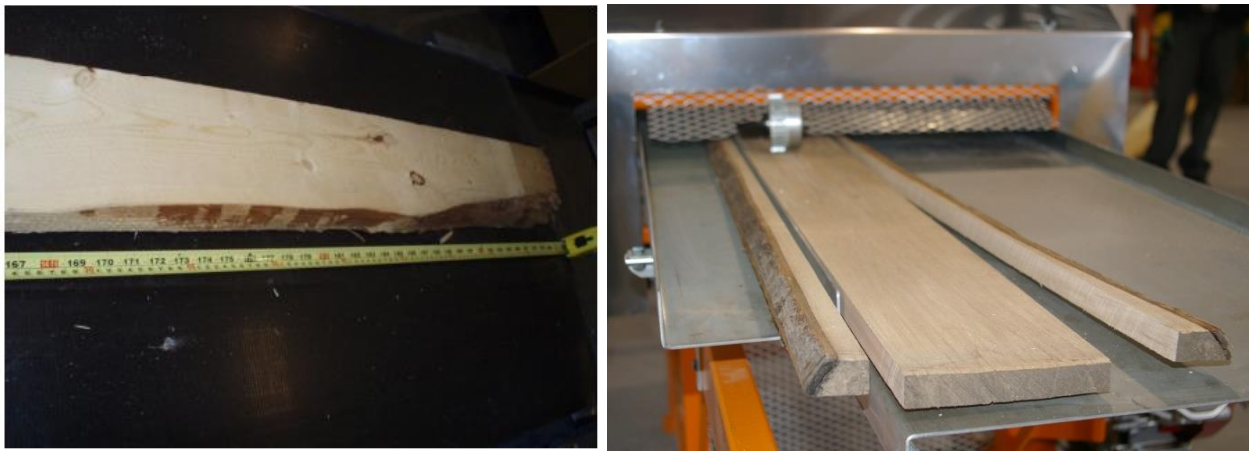


Figure 1.4 Représentation de planches délignées (Côté, 2013)

Le délignement de la planche provient principalement du mauvais positionnement (rotation) de la bille sur l'équipement VSS, effectué après son scannage par le lecteur optique durant le processus de sciage. L'origine de ce défaut peut être également expliquée par la dimension très importante

des billes. Ce type de non-qualité peut être aussi causé par le mal entretien des scies de l'équipement, notamment lorsqu'il y a un désalignement entre l'arbre de scies et les arbres des guides. Ainsi, le délignement est principalement affecté par la fréquence de la maintenance effectuée sur le VSS et par le diamètre de billes entrantes.

1.2.2 Planches déchiquetées

Une des étapes du contrôle qualité consiste également à choisir la meilleure face de chaque pièce, ayant le moins de défauts, de telle sorte que l'ébouteuse enlève le minimum possible de portions aux extrémités de la planche lors de la numérisation. Ce contrôle qualité est fait manuellement par un opérateur de ligne. Ces portions éliminées dues à la présence d'un défaut de surface sont déchirées grossièrement en petits morceaux de taille variée et sont donc déchiquetées. Ces dernières, transformées en copeaux, permettent de fabriquer des pâtes et des papiers. Les défauts de surface sont considérés comme des défauts de type visuel (Côté, 2013). Des exemples de ces défauts souvent détectés sur les planches sont les fissures et les cassures, comme le montre la Figure 1.5 (Desfor, 2003).

Figure 1.5 Exemple de fissures de planches

Il existe aussi des planches ayant des défauts de qualité sur toute la surface, de telle sorte qu'elles seront éboutées en bloc de 2' pour être retransformées en copeaux, introduisant alors un deuxième type de non-qualité, les pièces déchiquetées. Ces pièces peuvent être aussi le résultat d'un mauvais délignement, c'est-à-dire que les pièces sont devenues très étroites, ne respectant plus les dimensions normalisées, suite à leur réusinage par la déligneuse. Ainsi, une pièce déchiquetée est une pièce qui ne peut pas être retravaillée pour être ensuite classée, et elle ne peut être que déchiquetée.

L’origine de cette non-qualité est expliquée, soit par le mauvais positionnement de la planche dans la déligneuse, soit par la présence des défauts de surface. Ces défauts peuvent survenir des usures présentes dans les scies de l’équipement VSS ou encore de la qualité de la matière première. Ce type de non-qualité peut être aussi expliqué par les diamètres très faibles des billes et par les dimensions de planches à produire. En effet, lorsqu’une planche de dimension 2po x 3 po est éboutée sur l’un de ses côtés, en présence d’un défaut, cette planche devient de dimensions très petites, soit de 2po x 2po ou 2po x 1po, ce qui donne un produit non standard. La pièce sera donc déchiquetée. En revanche, si la planche est plus grande, on découpe seulement la partie contenant le défaut. Celle-ci sera déchiquetée et la partie non affectée sera classée.

1.2.3 Planches rejetées

Le dernier type de non-qualité est représenté par les pièces rejetées. Ces dernières sont des pièces considérées « perdues » dans ce processus. En effet, ces planches n’étaient pas lues correctement par les photocellules de présence lorsqu’elles passent après l’optimiseur (Figure 1.6).

Figure 1.6 Photocellules de présence

Les photocellules sont placées avant et après chaque scie de l’ébouteuse pour détecter l’emplacement de coupe de la planche et pour vérifier la longueur de la planche une fois coupée. Ces pièces devraient ainsi repasser sur la ligne d’éboutage. La non-détection de la planche par les photocellules est due à la présence des pièces tordues ou mal-transformées.

1.3 Problématique

Afin d'identifier les causes possibles du problème de non-qualité d'une manière simple et compréhensible, le diagramme de causes et effets, plus connu sous le nom du diagramme d'Ishikawa a été développé. Les aspects influençant la non-qualité ont été identifiés en se basant sur l'expertise des responsables de l'usine de sciage et sur les travaux de recherche précédents effectués dans les industries forestières (Quesada-Pineda & Arias, 2015; Smoljan & Ohran, 2015). Ainsi, un diagramme représentant les causes possibles de non-qualité des plus générales aux plus détaillées a été obtenu, comme le montre la Figure 1.7.

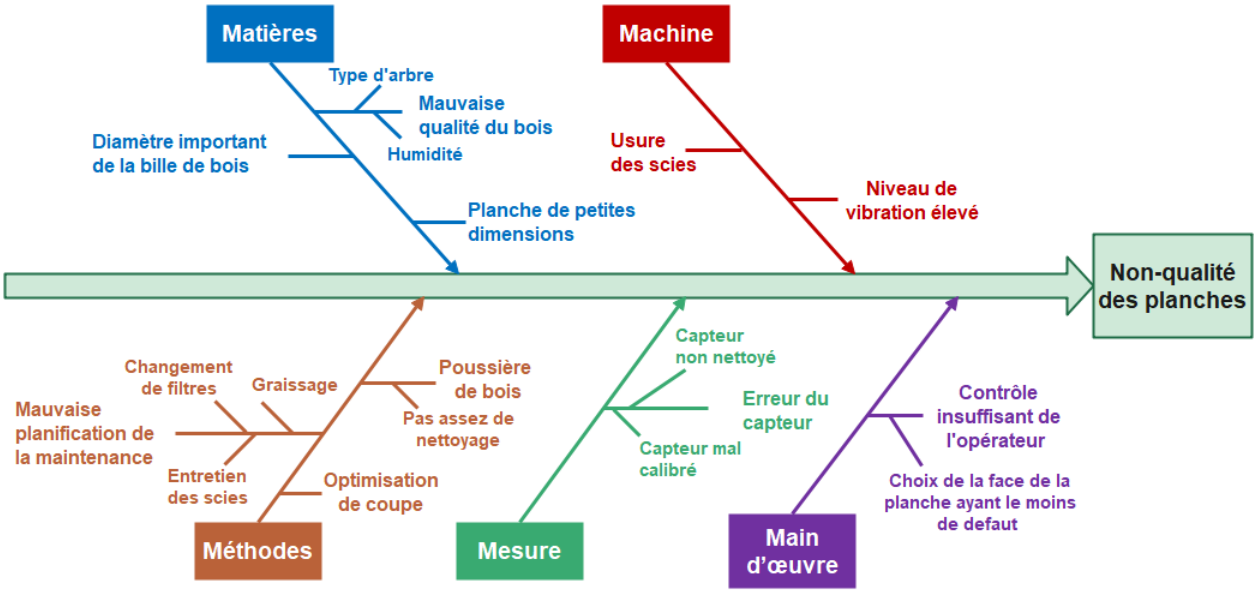


Figure 1.7 Les 5M du diagramme d'Ishikawa

Le diagramme d'Ishikawa est structuré en cinq catégories selon les 5M : matières, machines, méthodes, mesure et main d'œuvre. Pour chacune de ces catégories, des causes principales et secondaires sont mises en évidence. Ce diagramme fournit un support visuel qui met en évidence et hiérarchise les causes potentielles qui génèrent le problème de la mauvaise qualité dans les planches de bois. Par exemple, en analysant la 1ère catégorie, la matière, 3 causes principales qui engendrent la non-qualité des planches sont identifiées : le diamètre très grand de la bille, la dimension très petite des planches que la scierie souhaite produire et la mauvaise qualité du bois. Cette dernière a pour causes secondaires l'humidité et le type de bois utilisé pour obtenir la matière première. En considérant, d'un autre côté, la main d'œuvre, la représentation du diagramme nous

permettra d’identifier que parmi les causes principales dans cette catégorie qui peuvent affecter la qualité du produit fini, est le contrôle de l’opérateur. En effet, lors du contrôle de la qualité du produit fini, l’opérateur s’assure à bien choisir la face de la planche contenant le moins de défauts, de telle sorte que le minimum possible de matière est enlevé.

Une des causes de la non-qualité des planches présentée dans le digramme d’Ishikawa est liée aux problèmes d’équipements de la scierie qui peuvent être détectés par les niveaux vibratoires anormalement élevés. Les vibrations peuvent être provoquées par une ou plusieurs causes à un moment donné, comme le désalignement de l’arbre des scies, l'usure de certaines composantes (roulements à billes, des courroies d'entraînement ou d’engrenages) et le desserrage des bagues de roulement. Les vibrations peuvent aussi provenir du poids lourd des billes traversant l’équipement VSS. Si les vibrations ne sont pas contrôlées, elles peuvent causer des dommages ou une détérioration accélérée de l’équipement ainsi qu’une production de planches non conformes à la qualité souhaitée.

Chacun des intrants contrôlables ou non-contrôlables présentés dans la Figure 1.1 est relié à au moins une des causes principales extraites du diagramme d’Ishikawa illustré dans la Figure 1.7, comme le montre la Figure 1.8. Par exemple, les 3 causes relatives à la 1ère catégorie, matière, correspondent aux 2 intrants non contrôlables, les dimensions de billes et la qualité de billes, ainsi qu’à l’intrant contrôlables, les dimensions de planches.

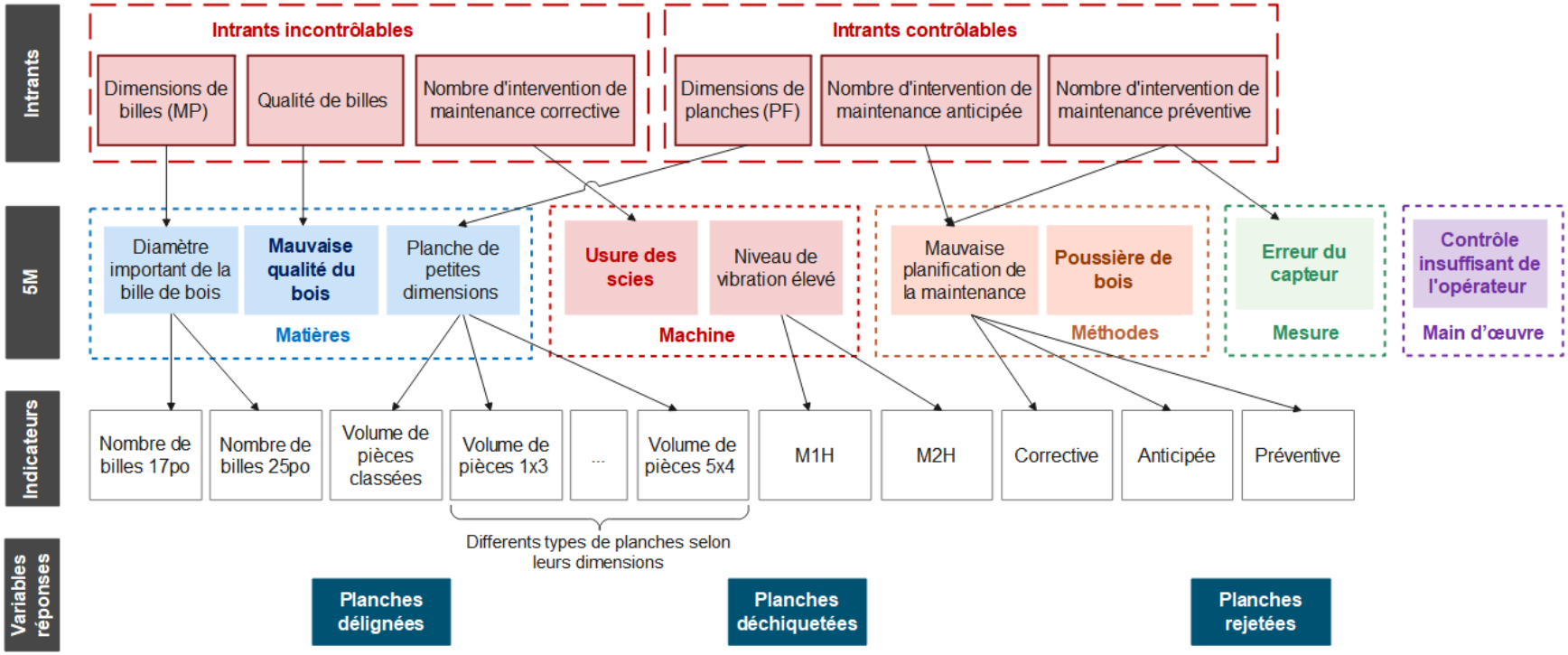


Figure 1.8 Relation des intrants avec les 5M et les indicateurs

Tout bien considéré, l'industrie de sciage est confrontée à plusieurs défis, notamment :

1. Comment détecter la non-qualité dans le produit fini?
2. Sous quelles conditions chacun des trois types de non-qualité apparaît sur les planches?
3. La production d'un certain type de non-qualité affecte-t-elle celle des autres types de non-qualité dans les planches?

Une fois ces défis sont relevés, il est possible d'intervenir par la suite sur les variables contrôlables pour réduire éventuellement la quantité de planches de mauvaise qualité. Pour ce faire, une bonne compréhension des effets et des relations entre les facteurs étudiés et les quantités de non-qualité produites lors du sciage est nécessaire.

Ainsi, le champ est largement ouvert pour la proposition et l'implémentation de méthodes d'apprentissage qui font actuellement leurs preuves dans le domaine industriel. Des modèles de classification et de régression seront donc développés pour étudier ces relations. Par la suite, une analyse de sensibilité sera appliquée en se basant sur le modèle de régression ayant la meilleure performance, afin d'identifier les variables les plus influentes sur ces différentes caractéristiques de non-qualité.

1.4 Description des données collectées

Afin de modéliser la non-qualité des planches lors du processus du sciage, une fusion des données procurées pendant 4 mois a été effectuée, regroupant des données de production et de maintenance ainsi que des données de vibration.

Ces données représentent certaines causes extraites du diagramme d'Ishikawa (Figure 1.7), à travers des indicateurs, aussi connus sous le nom de « features ». Cette représentation est mise en évidence dans la Figure 1.8. Cependant, des données qui peuvent caractériser certaines de ces causes n'ont pas pu été obtenues telles que la qualité du bois, l'usure des scies, l'état des capteurs des photocellules. En se basant sur l'analyse des causes effectuée à l'aide du diagramme d'Ishikawa, le Tableau 1.1 présente les indicateurs disponibles et manquants dans notre cas d'études.

Tableau 1.1 Indicateurs disponibles et manquants par rapport aux 5M

5M	Causes	Disponible	Manquant
Matières	Diamètre important de la bille de bois	X	
	Mauvaise qualité du bois		X
	Planche de petites dimensions	X	
Machine	Usure des scies		X
	Niveau de vibration élevé	X	
Méthode	Mauvaise planification de la maintenance	X	
	Poussière de bois		X
Mesure	Erreur du capteur		X
Main d'œuvre	Contrôle insuffisant de l'opérateur		X

Dans l'ensemble de données obtenues, chaque ligne décrit une observation enregistrée au cours d'une journée de production. Chaque observation est représentée par : la date ; des variables réponses représentant chacune le nombre d'un des 3 types de non-qualités décrites dans la section précédente ; et d'autres variables explicatives (indicateurs), considérées dans cette étude, organisées en 4 familles (matière première, produit fini, vibration et maintenance). Le tableau 1.2 présente et décrit ces variables.

Tableau 1.2 Description des indicateurs de l'ensemble des données

Famille	Nom de l'indicateur	Type	Définition de l'indicateur
Matière première	Nombre de billes 17po	Continue	Nombre de billes de bois de diamètre proche de 17 pouces utilisées par jour
	Nombre de billes 25po	Continue	Nombre de billes de bois de diamètre proche de 25 pouces entrantes par jour

Tableau 1.2 Description des indicateurs de l'ensemble des données (suite et fin)

Famille	Nom de l'indicateur	Type	Définition de l'indicateur
Produit fini	Nombre de planches classées	Continue	Nombre de planches produites de bonne qualité par jour
	Volume de planches classées [Mpmp]	Continue	Volume de planches produites de bonne qualité par jour
	Volume de planches 1x3 [Mpmp]	Continue	Quantité de planches de bonne qualité par type. Il y a 10 différents types de produit final selon la dimension : 1 x 3, 1 x 4, 1 x 6, etc.
	…	…	
	Volume de planches 5x4 [Mpmp]	Continue	
Vibration	M1H	Continue	Quantité d'énergie vibratoire obtenue du capteur M1H placé à l'arrière du moteur du VSS
	M2H	Continue	Quantité d'énergie vibratoire obtenue du capteur M2H placé à l'avant du moteur du VSS
Maintenance	Corrective	Ordinale	Nombre d'interventions de maintenance corrective
	Anticipée	Ordinale	Nombre d'interventions de maintenance anticipée
	Préventive	Ordinale	Nombre d'interventions de maintenance préventive

Notons que les variables de vibration relatives aux capteurs M1H et M2H sont extraites en se basant sur une nouvelle méthode d'analyse vibratoire utilisée par l'usine de sciage, appelée « Peakvue » (James & Robinson, 2001). Cette méthode sera expliquée davantage dans la section 2.3.1.

L'ensemble de données comprend 78 observations. La description statistique de cet ensemble notamment la moyenne, l'écart-type, la valeur minimale et maximale de chaque variable est fournie dans l'annexe A. Les valeurs uniques de chaque indicateur catégoriel sont également présentées.

Afin d'avoir une compréhension générale des relations présentes entre ces indicateurs, une matrice de corrélations a été calculée et représentée à travers le corrélogramme illustré dans l'annexe B. Le corrélogramme représente les corrélations pour toutes les paires de variables. Les corrélations positives sont affichées en bleu et les corrélations négatives en rouge. L'intensité de la couleur est proportionnelle au coefficient de corrélation. Donc, plus la corrélation est forte, c'est-à-dire la valeur est proche de -1 ou 1, plus les cases sont sombres. La légende des couleurs sur le côté droit du corrélogramme montre les coefficients de corrélation et les couleurs correspondantes. Une corrélation négative implique que les deux variables considérées varient en sens inverse, c'est-à-dire que si une variable augmente l'autre diminue et vice versa. Une corrélation positive implique que les deux variables considérées varient dans le même sens, c'est-à-dire que si une variable augmente, l'autre augmente.

Cette matrice indique que les variables de production sont fortement corrélées entre elles, étant donné que les coefficients de corrélation correspondants sont très proches de 1. Le corrélogramme montre également une corrélation élevée des deux variables réponses, le nombre de pièces délignées et celui des pièces déchiquetées, avec les variables de production, notamment la quantité de matières premières et la quantité des différents types du produit fini.

Cette forte corrélation est expliquée par la dépendance directe du nombre de planches de mauvaise qualité à celui de toutes les planches produites. Dans ce cas, il n'est pas recommandé d'utiliser des nombres bruts de planches sans tenir compte de la quantité totale produite, comme montré dans (Kleinosky et al., 2007; Salewski et al., 2003). En effet, pour un jour i, la quantité de planches délignées peut être deux fois plus grande que celle d'un autre jour j. Lorsque des nombres bruts sont utilisés dans cette étude, la quantité de planches délignées produites le jour j semble être acceptable comparé à l'autre jour. Si, toutefois, la quantité totale de planches produite du jour i est 4 fois plus grande que celle du jour j, la quantité de planches de bonne qualité produites le jour i serait nettement meilleure comparé au jour j. L'utilisation des nombres bruts peut biaiser les résultats en raison de cette dépendance des données. Par conséquent, des pourcentages sont utilisés plutôt que des nombres pour exprimer les quantités de planches, car la quantité de billes utilisées ou de planches produites varie d'un jour à un autre. Nous avons ainsi procédé à une transformation des quantités de chaque type de non-qualité en pourcentage, en divisant le nombre des pièces d'un

des types de non-qualité par la quantité totale de planches produites. Par exemple, le pourcentage de pièces délignées est déterminé comme suit :

$$\%\ de\ pièces\ délignées = \frac{Nombre\ de\ pièces\ délignées}{Nombre\ totale\ de\ pièces\ produites} * 100$$

Il serait justement plus correct d'utiliser des pourcentages de non-qualité plutôt que des quantités, pour ne pas tenir en compte de la production journalière des planches de bois. C'est-à-dire, si la quantité totale de planches produites est fixe quel que soit le jour, la comparaison des quantités de non-qualité serait juste, n'ayant pas une dépendance directe avec les variables de production.

De même, le pourcentage de pièces classées est déterminé comme suit :

$$\%\ de\ pièces\ classées\ = \frac{Nombre\ de\ pièces\ classées\ (de\ bonne\ qualité)}{Nombre\ totale\ de\ pièces\ produites} * 100$$

Afin d'étudier l'effet des différentes dimensions des planches de bois à produire sur les variables liées à la non-qualité, les volumes de chaque type de pièce ont été également transformés en pourcentages. Pour ce faire, le volume d'un type de planche est divisé par celui de toutes les planches classées. Par exemple, pour transformer le volume de pièces 2x3 en pourcentage, la formule ci-dessous est appliquée :

$$\%\ de\ pièces\ 2x3 = \frac{Volume\ \text{de}\ pièces\ \ 2\text{x}3}{Volume\ de\ pièces\ classées} * 100$$

De manière similaire, la quantité d'un type de bois est convertie en pourcentage, afin de comprendre l'effet du diamètre de bille sur les variables de non-qualité, comme suit :

$$\%\ de\ billes\ 17po = \frac{Nombre\ \text{de}\ billes\ \ 17\text{po}}{Nombre\ total\ de\ billes\ (17po\ et\ 25po)} * 100$$

Une fois la transformation des quantités en pourcentage effectuée, nous obtenons des variables qui sont de même nature, permettant par la suite une meilleure interprétation des résultats. L'annexe C présente la matrice de corrélation de l'ensemble de données avec ces nouvelles variables transformées. Le corrélogramme correspondant montre que les couleurs des cases sont plus claires par rapport à celui de l'annexe B, indiquant des corrélations moins fortes entre les variables, une fois transformées en pourcentage.

En se basant sur les causes de non-qualité déterminées dans la Section 1.2 et dans la Section 1.3, le Tableau 1.3 présente les indicateurs possibles qui peuvent affecter chacune des variables réponses. En effet, chacune des causes identifiées peut être représentée par un ou plusieurs indicateurs qui sont à leurs tours capables d'expliquer chaque variable réponse. Une analyse d'importance basée sur les modèles d'apprentissage est ensuite effectuée sur les indicateurs, afin de vérifier si la non-qualité des planches est justement provoquée ou non par les causes présentées dans le Tableau 1.3.

Certaines causes de chaque type de non-qualité identifiées dans la Section 1.2 ne peuvent pas être représentées par des indicateurs des données procurées, telles que les données relatives aux lecteurs optiques des photocellule (positionnement de la bille ou encore le patron de coupe), dans le cas du délignement des planches. D'autres causes peuvent être représentées par des indicateurs, comme les dimensions de billes. En effet, celle-ci est représentée par l'indicateur « Billes 17po » qui est défini comme étant la quantité en pourcentage de billes ayant un diamètre proche de 17po. Considérant les 2 catégories de billes selon leurs diamètres, si le pourcentage de bille 17po est inférieur au pourcentage standard, le pourcentage des billes 25po est élevé, ce qui peut expliquer la production des planches délignées. Les causes relatives au mauvais entretien des scies et au désalignement des arbres de l'équipement VSS sont représentées par les données vibratoires.

Tableau 1.3 Relation des variables réponses avec les indicateurs

Variable réponses	Indicateurs	Causes hypothétiques
Pièces délignées [%]	• Billes 17po [%]	Diamètre important de la bille de telle sorte que certaines planches obtenues aient des côtés arrondis.
	• Pièces 1x3 [%] … • Pièces 5x4 [%]	Production des planches de grandes dimensions
	• Maintenance corrective • Maintenance préventive • Maintenance anticipée	Mal entretien des scies de l'équipement VSS causant la présence des flaches de rive dans la planche lors de la coupe.
	• Données vibratoires	Désalignement de l'arbre de scies

Tableau 1.3 Relation des variables réponses avec les indicateurs (suite et fin)

Variable réponses	Indicateurs	Causes hypothétiques
Pièces déchiquetées [%]	• Maintenance corrective • Maintenance préventive • Maintenance anticipée	Usure des scies de l'équipement VSS qui peut créer des défauts surfaciques sur la planche.
	• Pièces 1x3 [%] … • Pièces 5x4 [%]	Dimensions petites de planches mal délignées qui deviennent encore plus étroites, ne respectant plus les normes.
	• Données vibratoires	Usure des scies de l'équipement VSS
Pièces rejetées [%]	• Billes 17po [%]	Les photocellules n'arrivent pas à lire la planche du lecteur optique. Les données utilisées par le lecteur correspondent aux dimensions de la matière première et ceux des planches.
	• Pièces 1x3 [%] … • Pièces 5x4 [%]	
	• Données vibratoires	Planches mal-transformées par l'équipement

1.5 Objectifs du mémoire

Compte tenu des défis inhérents à l'évolution technologique et à la compétitivité industrielle auxquelles l'industrie de sciage est constamment confrontée, les exigences en termes de productivité et de coûts réduits de la non-qualité dans le processus de sciage sont cruciales. Dans ce mémoire, l'objectif global consiste à modéliser la non-qualité des planches de bois lors du processus du sciage dans le but d'identifier les causes primaires de la mauvaise qualité des produits. Ainsi, des techniques de modélisation seront développées pour expliquer les causes relatives à la production des produits de mauvaise qualité et pour identifier les variables affectant chaque type de non-qualité dans les planches. Par conséquent, cette étude vise à trouver la meilleure combinaison de variables explicatives réduisant les défauts dans le produit fini. Deux sous-objectifs seront poursuivis dans ce projet de recherche :

1. Détecter la non-qualité dans le produit fini durant le processus de sciage en utilisant les modèles de classification. Pour ce faire, les activités suivantes seront réalisées :
    a. Transformer les variables réponses qui correspondent aux quantités de chacun des 3 types de non-qualité des planches en variables binaires, par dichotomisation.
    b. Extraire les caractéristiques statistiques à partir des signaux vibratoires et les prétraiter.
    c. Générer des règles de classification représentant chaque type de non-qualité du produit fini en utilisant la méthode de classification LAD pour Logical Analysis of Data (Boros et al., 2000). Cette méthode sera décrite dans la Section 2.4.1.
2. Identifier les conditions sous lesquelles chaque type de non-qualité apparaît sur les planches, en se basant sur des modèles de régression à multi-sorties, pour expliquer les différents types de non-qualité trouvée dans les planches produites lors de la coupe. La réalisation de ce sous-objectif implique trois activités :
    a. Vérifier les hypothèses d'application de la régression afin de voir l'adéquation du modèle aux données observées avec les variables réponses.
    b. Analyser les données en utilisant les algorithmes de régression et développer un modèle capable d'expliquer les différents types de non-qualité.
    c. Effectuer une analyse d'importance basée sur les modèles développés des indicateurs sélectionnés.
    d. Développer des modèles de régression capables de prendre en considération les dépendances entre les différents types de non-qualité.

1.6 Organisation du mémoire

Le mémoire est organisé en quatre chapitres. Le Chapitre 2 présente une revue de la littérature des études et des analyses effectuées pour identifier les méthodologies et les travaux précédents en lien avec la modélisation de la non-qualité dans les industries.

Le Chapitre 3 applique des algorithmes de classification afin de détecter la non-qualité du produit fini. Particulièrement, l'analyse logique des données LAD « Logical Analysis of Data » est proposée comme technique pour la classification.

Le chapitre 4 développe et évalue différents modèles de régression à multi-sorties afin d'expliquer le problème de la non-qualité des planches. Une analyse d'importance relative est effectuée pour étudier la relation entre les indicateurs considérés et les 3 types de non-qualité.

Enfin, le Chapitre 5 présente la conclusion générale de ce travail ainsi que les limites et les perspectives d'amélioration.

CHAPITRE 2 REVUE DE LITTÉRATURE

2.1 Méthodes de contrôle statistique de la qualité

Il existe une riche littérature sur les méthodes statistiques de qualité des produits finis dans les industries, particulièrement dans l'industrie forestière. (Young et al., 2007) ont implémenté un système de contrôle de processus statistique ou encore SPC « Statistical Process Control » en temps réel pour réduire la variation de l'épaisseur des planches et améliorer ainsi la qualité du produit fini. En effet, une variation excessive de l'épaisseur du bois peut réduire la qualité du produit. Le système mesure l'épaisseur du bois et affiche ces mesures sur des cartes de contrôle et des histogrammes distribués en temps réel à toutes les scieries. En utilisant ce système, quatre scieries situées aux États-Unis ont pu améliorer la qualité des planches produites, tout en utilisant plus judicieusement les ressources forestières. Le SPC peut réduire, dans certains cas, la quantité des produits défectueux en minimisant l'intervalle de temps entre la visualisation des données de processus et l'action sur la variabilité du produit. (Marenče et al., 2020) évaluent l'impact de la qualité de la bille sur la qualité des planches, en se basant sur l'inspection par échantillonnage. Les auteurs montrent qu'au fur et à mesure que la qualité des arbres diminue, la qualité des planches diminue et la part de copeaux et de pâtes augmente. L'un des principaux facteurs ayant un effet considérable sur la qualité des planches est la présence des nœuds dans les bois. Ces nœuds produisent des défauts surfaciques sur les planches, entraînant leur déchiquetage. Il est également révélé que le diamètre de la bille influence la qualité du produit fini. En effet, si le diamètre est important, des flaches de rives peuvent être détectées dans les planches. Une flache consiste à une portion de surface arrondie avec ou sans écorce, restant apparente même après le sciage. Cependant, étant donné que ces résultats sont basés sur un échantillon, il y a toujours une certaine probabilité que les déductions sur la qualité du lot soient fausses. (Ribarits et al., 2007) ont examiné l'effet de la géométrie des lames de scie, des angles de coupe et de cisaillement sur la qualité de surface des planches, en se basant sur des méthodes d'expérimentation, particulièrement les plans d'expérience DOE « Design Of Experiments ». L'objectif était de comprendre comment chacun de ces paramètres et les interactions entre eux influençaient la qualité surfacique des planches, afin de

minimiser la quantité des planches déchiquetées. Les résultats montrent que la qualité est fortement corrélée avec les angles de coupe et le cisaillement. Ces auteurs indiquent que des méthodes plus développées, comme les modèles d'apprentissage, sont nécessaires pour explorer plus en détail certaines des interactions complexes entre ces paramètres et leurs effets sur la qualité de surface des planches.

2.2 Méthodes de modélisation de la qualité

La modélisation de la qualité des produits finis s'intéresse à établir les liens de causes-à-effets entre la détection d'une mauvaise qualité sur les produits et les facteurs explicatifs.

De nombreuses études effectuées dans le domaine forestier sont étroitement orientées sur la qualité de la matière première, les billes de bois, ayant prouvé que celle-ci a un impact considérable sur la qualité des planches produites.

Par exemple, (Lampinen & Smolander, 1994) ont développé une approche de reconnaissance de formes, en se basant sur l'algorithme de classification, les réseaux de neurones, pour la reconnaissance des défauts surfaciques des planches causés par la mauvaise qualité des matières premières. Cette approche permet les experts ainsi de décider, selon le type des défauts surfaciques, de conserver ou de déchiqueter la planche. Dans cette étude, les indicateurs utilisés sont extraits des échantillons d'images qui représentent les surfaces des planches de bois. Les performances de l'algorithme de classification ont été évaluées avec un ensemble de plus de 400 échantillons d'image de planches, avec une précision d'environ 85 %. (Tumenjargal et al., 2019) trouvent que les nœuds et les flaches étaient les principaux facteurs de déclassement des planches, en se basant sur les modèles d'analyse de variance (Grmanová et al.). Dans ce cas, deux types de non-qualité de planches sont considérés. En effet, les planches délignées se distinguent par la présence des flaches, tandis que les planches déchiquetées sont généralement dues aux défauts de surface comme les nœuds. Ces auteurs montrent que ces facteurs dépendent principalement de la qualité du bois à son état initial qui est souvent affecté par l'environnement. (Zolotarev et al., 2019) proposent une méthode automatique basée sur le modèle des réseaux multimodaux d'auto-encodeur pour prédire la qualité des planches dans les premières étapes du processus de sciage, selon la qualité des billes. Ce modèle est capable de détecter les nœuds de bois en se basant sur des nuages de points acquis par les systèmes de balayage laser effectués sur les surfaces de billes et de planches. Les données

utilisées considèrent 50 billes de 5 catégories différentes selon leurs diamètres et le nombre de nœuds et 274 planches sciées à partir de celles-ci. L'effet de ces différentes catégories de billes sur la qualité des planches a été étudié et il a été conclu que le modèle démontre une bonne précision de 87%.

Cependant, la qualité du produit fini n'est pas seulement causée par l'état de la matière première. Il existe d'autres facteurs influents notamment les dimensions choisies des planches à produire, l'état des équipements pendant le processus de sciage, qui est généralement caractérisé par les signaux vibratoires ou par la maintenance. Par exemple, (Rudakov et al., 2018) ont introduit le problème de la détection de la qualité des planches, plus précisément des défauts surfaciques des planches de bois, comme les fissures ou les flaches. Le but de cette recherche est de développer un système de vision par ordinateur capable de détecter et de reconnaître les dommages du bois scié provoqué mécaniquement par les équipements. Pour ce faire, le réseau de neurones convolutifs (CNN) ou encore « Convolutional neural network » a été sélectionnée comme technique pour la classification d'images et la détection de ces défauts surfaciques. Cette technique a finalement produit de bons résultats avec une précision de classification importante d'environ 92 %. (Kuljich et al., 2017) évaluent les effets de l'état de la scie sur la qualité de surface de planches, traitée par les équipements de coupe, en se basant sur les modèles d'ANOVA. Les résultats montrent que les scies coudées ou désalignées induisent des vibrations dans le bord d'inclinaison, ce qui pourrait expliquer la variation de la qualité de surface de la planche et ainsi son déchiquetage. Plusieurs travaux ont démontré que les vibrations accentuées induites sont des indicateurs de la dégradation de l'état de santé des machines (Iskra & Hernández, 2012; Jackson et al., 2002; Lemaster et al., 2000). Ainsi, la qualité de surface de la planche peut diminuer au fur et à mesure que l'état de fonctionnement de l'équipement se dégrade. (Nasir & Cool, 2019) confirment aussi que lors de l'utilisation de scies, la qualité des planches est influencée par le comportement dynamique de l'équipement de coupe, comme les vibrations et la déviation de la lame de scie.

Après avoir examiné ces articles, nous remarquons que la plupart se concentrent sur les deux premiers types de non-qualité des planches, le délignement et le déchiquetage. D'après ces travaux, les facteurs qui expliquent ces non-conformités du produit fini sont principalement relatifs à la dimension et à la qualité de la matière première, la dimension des planches à produire et l'état de santé des équipements de sciage. Ce dernier facteur est généralement caractérisé par le niveau

vibratoire des composants en rotation tels que les roulements et les engrenages de l'équipement. Dans cette étude, des caractéristiques statistiques sont extraites à partir des signaux vibratoires qui sont capturés en se basant sur la méthode d'enveloppe, appelée Peakvue. Une présentation de cette méthode est ainsi effectuée dans la section suivante. Les méthodes de modélisation utilisées pour expliquer les différents types de non-qualité sont également exposées.

2.3 Préparation des données vibratoires

2.3.1 Méthode Peakvue

Les signaux de vibration peuvent être acquis facilement à l'aide des capteurs de vibration, les accéléromètres, créant une source d'informations précieuse pour la surveillance de l'état des machines (Plante et al., 2015; Ragab, A. et al., 2019) et pour le contrôle qualité (Aguilar et al., 2016; Carnero et al., 2010). La méthode classique employée dans l'analyse vibratoire consiste à :

1. Capturer une forme d'onde temporelle à partir de ce type de capteur pour une période spécifique, obtenant ainsi un signal.
2. Transmettre les signaux de hautes fréquences et bloquer les signaux de très basses fréquences qui représentent les événements non répétitifs en utilisant un filtre passe haut.
3. Supprimer tous les signaux de hautes fréquences, à travers un filtre passe-bas.
4. Transformer le signal temporel filtré en un signal dans le domaine fréquentiel en se basant sur la transformation de Fourier rapide FFT « Fast Fourier Transform » ou/et en un signal dans le domaine temps-fréquence en se basant par exemple sur la Transformée de Fourier à court terme STFT « Short-Time Fourier Transform ».

Cependant, en appliquant cette méthode d'analyse, l'élimination des signaux de hautes fréquences peut entraîner une perte importante d'informations relatives à la dégradation du roulement. En effet, la sortie analogique d'un accéléromètre inclut principalement deux composantes sur toute la bande passante. La première correspond à une activité classique induite par les vibrations normales de la machine et la deuxième correspond aux ondes de stresse provenant des impacts, des chocs, des fissures ou d'usure abrasive. Les ondes de stresse sont observées dans l'initiation et la progression des défauts dans les roulements. Généralement, la composante relative à la vibration

normale couvre une bande de fréquence inférieure à celle de la composante de l'activité des ondes de stresse. Ainsi, si la méthode d'analyse de vibration classique est appliquée, des informations pertinentes, comme l'apparition des frictions ou des fissurations dans le roulement, peuvent être absentes.

Pour remédier à ce problème, des techniques d'analyse d'enveloppe à haute fréquence ont été développées (Lebold et al., 2000). Ces techniques permettent de détecter principalement l'énergie associée aux événements d'impact ou d'impulsion qui correspondent à la présence d'un défaut dans le roulement. Pour ce faire, la composante induite par l'onde de stresse du signal est séparée de la vibration normale en acheminant le signal à travers un filtre analogique passe-haut d'ordre élevé.

Parmi ces méthodes d'analyse d'enveloppe, Emerson a proposé une technique appelée, « Peakvue », qui consiste à échantillonner le signal et à sélectionner les valeurs de pointe (les valeurs de crête) ou encore les « Peak Values » à de plus hautes fréquences que la méthode classique (James & Robinson, 2001). Le spectre Peakvue est ensuite calculé à partir des données temporelles en se basant sur l'algorithme FFT. La Figure 2.1 résume les étapes de la méthode d'analyse vibratoire classique et celles de la méthode Peakvue (James & Robinson, 2001).

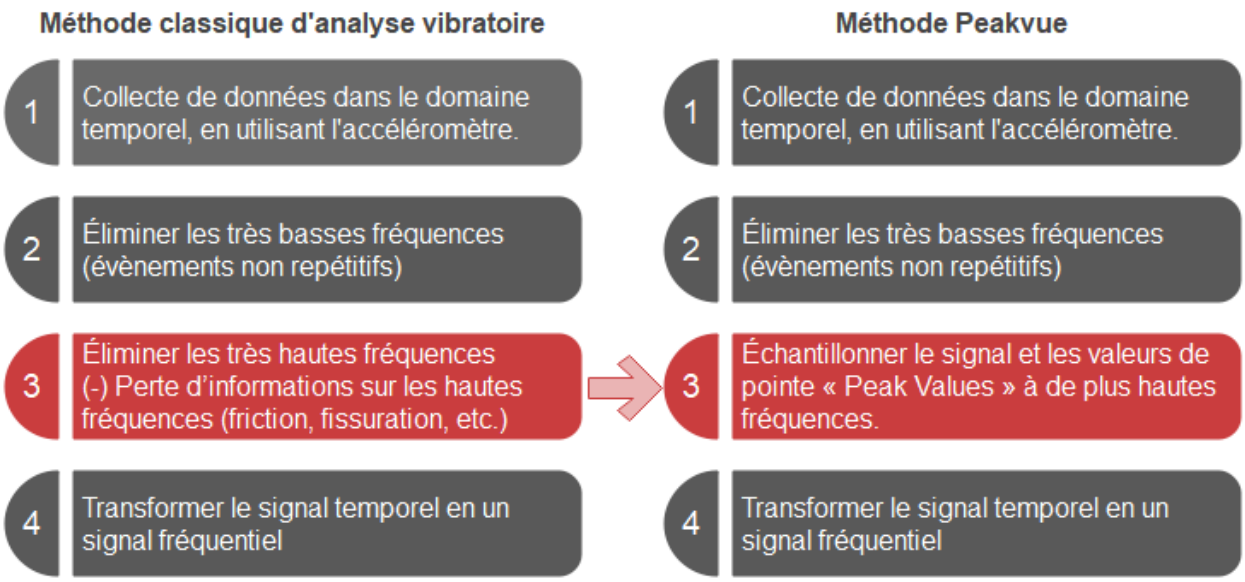


Figure 2.1 Comparaison entre la méthode d'analyse vibratoire classique avec la méthode Peakvue

Une étude de comparaison de la méthode Peakvue avec d'autres techniques d'enveloppe à haute fréquence similaires comme la méthode de démodulation d'amplitude a été également effectuée. Les résultats de (James & Robinson, 2001) montrent que les pics dans le spectre de démodulation fournissent peu ou pas d'indication de l'apparition d'un défaut dans le roulement, pour différentes raisons. Parmi ces raisons, le choix d'une bande de fréquence peut entraîner une perte considérable du signal d'amplitude. En revanche, ce choix n'aura qu'un effet léger sur la réponse en amplitude dans le signal Peakvue, comme démontré dans (James & Robinson, 2001).

Les motifs observés dans le spectre de fréquence et l'onde temporelle de la technique Peakvue sont importants pour le diagnostic de l'endommagement du roulement. Peakvue fournit particulièrement des résultats pertinents dans le domaine fréquentiel. En effet, l'analyse du signal extrait permet de révéler des fréquences de répétition même lorsque celles-ci ont une petite fluctuation aléatoire. Le spectre Peakvue montre ainsi des pics parfaitement visibles, malgré l'amplitude relativement faible de l'impact, à travers les harmoniques présentes dans le signal spectral, comme le montre la Figure 2.2.

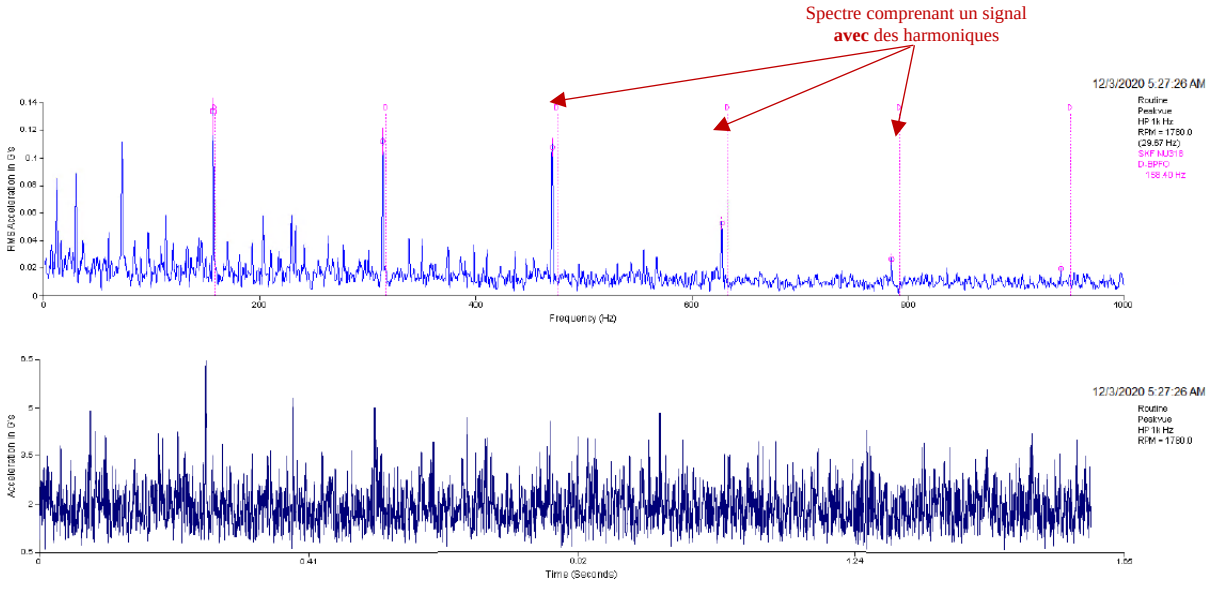


Figure 2.2 Exemple d'un spectre Peakvue

Avant de collecter les données Peakvue, des configurations de mesure doivent être effectuées. Ceci inclut le choix du filtre passe-haut de la bande passante pour éliminer les basses fréquences venant

du fonctionnement normal des composants, le choix de la fréquence d'échantillonnage et du nombre d'échantillonnages dans les données spectrales.

Le filtre passe-haut est choisi dépendamment de la bande passante d'analyse Peakvue, précisément de la fréquence maximale (F_{max}). Celle-ci est déterminée par la fréquence de défaut la plus élevée possible. En l'absence d'engrenage, la fréquence de défaut de la bague intérieure du roulement BPFI « Ball Pass Frequency Inner » est généralement la fréquence la plus élevée pour les roulements. La valeur de F_{max} doit être ainsi réglée à plus de 3 fois ou 4 fois BPFI.

Une fois la valeur de la fréquence maximale déterminée, les valeurs de crête sont collectées à un taux de $2{,}56 \times F_{max}$. Autrement dit, la fréquence d'échantillonnage est à 2,56 fois la fréquence maximale (James & Robinson, 2001).

Après avoir sélectionné le filtre passe-haut et la bande passante pour l'acquisition de données, le paramètre suivant à sélectionner est le nombre d'échantillonnages ou encore le nombre de crêtes capturées dans le domaine fréquentiel. Le critère de contrôle est de fournir une résolution adéquate sur la fréquence de défaut la plus basse possible qui est celle de la cage FTF « Fundamental Train Frequency ». Il est donc important d'avoir une résolution suffisante pour résoudre clairement la fréquence de la cage. Cela se traduit par un nombre d'échantillons d'au moins 15 révolutions qui devraient être enregistrées dans les données spectrales.

Tout compte fait, il y a plusieurs éléments clés à respecter concernant l'utilisation de la méthode PeakVue. Celle-ci permet la détection des impacts causés par les défauts de roulement dans le spectre plus clairement. Elle peut souvent détecter de nombreux défauts manqués dans l'analyse de vibration classique. Cette méthode peut être également appliquée aux machines sur une large plage de vitesse, de fractionnaire à quelques milliers de RPM « revolutions per minute ». Dans ce qui suit, une extraction des caractéristiques à partir des signaux vibratoires obtenues de la méthode Peakvue est effectuée.

2.3.2 Extraction de caractéristiques de signaux vibratoires

Dans notre cas d'étude, les données de vibrations ont été collectées en utilisant la méthode Peakvue. Les signaux vibratoires obtenus sont relatifs à l'équipement VSS qui est composé d'un moteur en position vertical, fonctionnant à une vitesse constante de 1780 RPM. Le moteur est parallèle au

mandrin qui tient les scies coupant le bois. Il entraîne le mandrin avec un système Poulies-Courroie. Deux accéléromètres, notés « M1H » et « M2H » sont placés à deux endroits différents sur ce moteur. M1H est à l'arrière du moteur (côté ventilateur), loin de l'effort d'entraînement, contrairement à M2H qui est positionné à l'avant du moteur (côté poulie).

Ces données vibratoires ont été collectées sur une période de 4 mois, soit 80 jours, si on tient en compte seulement des jours de production, mais de manière discontinue, soit durant quelques minutes par jour. Le nombre total de signales collectés par capteur pour chaque jour est 4095. Après la transformation du signal temporel au signal fréquentiel, le spectre a été échantillonné afin d'obtenir les 24 plus grandes « peak values ». C'est-à-dire, le nombre de crêtes capturées a été configuré à 24 par l'industrie. Ces données vibratoires ont été collectées par l'entremise d'un fournisseur. Nous disposons ainsi de divers échantillons sur une période de 4 mois à raison d'un spectre au maximum par jour.

Afin de distinguer les phénomènes physiques que renferment les spectres, des caractéristiques du domaine temporel et fréquentiel sont déterminées à partir du signal de vibration en se basant sur des statistiques. Ces caractéristiques permettent aux experts d'inspecter les signaux vibratoires et d'acquérir les informations relatives aux défauts concernant les roulements pour ainsi les aider dans le diagnostic des machines. Idéalement, un modèle de diagnostic se base sur un signal de vibration complet en entrée. Cependant, la quantité de données que chaque signal contient est trop importante, ce qui entraîne un coût de calcul très élevé. Pour résoudre ce problème, des caractéristiques sont extraites à partir des signaux de vibration pour obtenir des informations plus significatives. Ainsi, un signal de vibration composé de 30 000 observations de données peut être réduit à un vecteur caractéristique de moins de 30 éléments.

Par exemple, (Helmi & Forouzantabar, 2019) utilisent des mesures statistiques de signaux vibratoires dans le domaine fréquentiel pour diagnostiquer les défauts de roulement dans les moteurs électriques. Parmi les caractéristiques extraites, la moyenne, la variance, le kurtosis, l'asymétrie et la moyenne quadratique (RMS) « Root mean square » ont été utilisés pour leur cas d'études. (Abbasion et al., 2007; Subrahmanyam & Sujatha, 1997) ont utilisé un ensemble de caractéristiques du domaine temporel pour le diagnostic, tandis que (Sun et al., 2007) ont employé une combinaison de caractéristiques temporelles et fréquentielles.

Similairement, dans notre cas d'études, deux ensembles de caractéristiques représentant le signal temporel ainsi que le signal spectral sont définis pour permettre de distinguer la non-qualité présente dans les planches. La plupart de ces variables déterminées à partir du signal vibratoire sont des statistiques descriptives ou des statistiques d'ordre élevé. Le Tableau 2.1 inclut une description de ces variables.

Six caractéristiques dans le domaine temporel sont extraites, étant les plus utilisées (Mortada, M. A. et al., 2011; Sun et al., 2007; Xu et al., 2016). Les formules de ces caractéristiques sont présentées dans le Tableau 2.1, où $z(t)$ représente l'amplitude du signal brut pour un instant t et $\bar{z}$ est défini comme étant la moyenne de l'amplitude dans une période T du signal temporel, comme suit $\bar{z} = \frac{1}{T}\int_0^T z(t)dt$. La valeur pic, plus connue sous le nom de « PeakValue » représente la valeur maximum du signal dans le domaine temporel et permet la détection de défauts graves dans le roulement. Le RMS qui correspond à la moyenne quadratique du signal augmente progressivement au fur et à mesure que le défaut se développe. L'écart-type mesure la dispersion d'un signal autour de leur valeur moyenne comme référence. Le facteur de crête (FC) est un bon indicateur des premiers stades de la défaillance d'un roulement. Une valeur de FC inférieure à 3 indique un état normal, 3 à 8 indique l'initiation d'une faille et une valeur de 8 à 10 signale une croissance de faille (Mortada, M. A. et al., 2011). Cependant, dans le cas où le signal vibratoire est déformé ou lorsque plusieurs défauts sont présents, ces seuils sont moins efficaces dans la détection des défauts (Safizadeh, 2001). Kurtosis Ku est une statistique d'ordre élevé qui détecte les pics dans le signal temporel de vibration. Si le roulement fonctionne normalement, la fonction de densité de probabilité PDF « Probability Density Function » est une distribution normale et la valeur de Kurtosis est égale à 3 (Dyer & Stewart, 1978). Le coefficient d'asymétrie est une statistique d'ordre supérieur similaire au Kurtosis, impliquant le moment de troisième ordre. Il quantifie le comportement d'asymétrie du signal de vibration grâce au PDF.

Toutes ces caractéristiques statistiques extraites dans le domaine temporel sont considérées pour surveiller le comportement vibratoire qui permet de détecter la présence de défauts et un ainsi un bon diagnostic de l'équipement. Ces caractéristiques sont choisies, car elles se sont avérées utiles pour l'identification des défauts des roulements et des engrenages (Li et al., 2011; Rafiee et al., 2009).

Le domaine fréquentiel peut divulguer des informations qui ne peuvent pas être découvertes dans le domaine temporel. Ainsi, cinq caractéristiques du domaine fréquentiel sont également utilisées comme indiqué dans le Tableau 2.1, où $s(f_k)$ est l'amplitude spectrale qui correspond à la fréquence f_k, définie comme étant la fréquence de la k-ème valeur de crête échantillonnée du spectre et K est le nombre total de raies spectrales (Sun et al., 2007; Xu et al., 2016).

La première caractéristique fréquentielle extraite, la moyenne, est une mesure simple du spectre vibratoire. Si elle est basse, les valeurs de crête élevées sont moins fréquentes ou inexistantes dans le spectre et vice-versa. Cela dit, il est aussi possible que des spectres très différents, dont certains avec des valeurs de crêtes éloignées les unes des autres, et d'autres avec des valeurs de crêtes proches les unes des autres, aient des moyennes égales. Ainsi, la moyenne ne tient pas compte de la variabilité dans le signal. Une autre mesure statistique, la variance, est ainsi considérée pour représenter la variabilité de l'énergie vibratoire par rapport au comportement moyen des amplitudes échantillonnées dans le spectre. Son évolution d'un spectre à un autre peut aussi potentiellement indiquer de nouvelles perturbations ou des modifications. Le RMS spectral est aussi un indicateur important de la sévérité des vibrations. Le coefficient d'asymétrie spectrale « Spectral Skewness SS » et le Kurtosis spectral (KS) sont les mesures statistiques avancées. Le SS mesure l'asymétrie de la distribution du spectre autour de sa moyenne, tandis que le KS mesure la distribution des valeurs spectrale et se compare à une distribution gaussienne (Lerch, 2012). (Antoni & Randall, 2006) utilisent le KS pour surveiller efficacement l'état en se basant sur les vibrations des machines tournantes. Les auteurs démontrent le potentiel élevé de cette mesure à détecter et à caractériser des signaux non-stationnaires. Contrairement à l'analyse du Kurtosis classique, le KS fournit un moyen robuste de détecter les défauts naissants même en présence de bruits. De plus, le KS offre un moyen presque unique de concevoir des filtres optimaux pour filtrer la signature mécanique des défauts. La première propriété est importante à des fins de surveillance, tandis que la seconde s'avère très utile pour le diagnostic. Notons que le KS et le SS sont déterminés en utilisant l'écart-type σ et le centre de fréquence $\bar{f}$ définit comme étant la moyenne pondérée de la fréquence affectée à l'amplitude spectrale (Helmi & Forouzantabar, 2019).

Tableau 2.1 Extraction de caractéristiques des signaux vibratoires

Caractéristiques temporelles	**Caractéristiques fréquentielles**
$PeakValue = \max(z(t))$	$Moyenne = \frac{\sum_{k=1}^{K} s(f_k)}{K}$
$RMS = \sqrt{\frac{1}{T}\int_0^T z(t)dt}$	$RMS\ spectral = \sqrt{\frac{\sum_{k=1}^{K} f_k^2 s(f_k)}{\sum_{k=1}^{K} s(f_k)}}$
$Variance = \sqrt{\frac{1}{T}\int_0^T (z(t)-\bar{z})^2 dt}$	$Variance\ spectral = \frac{\sum_{k=1}^{K}\left(s(k) - \frac{\sum_{k=1}^{K} s(f_k)}{K}\right)^2}{K-1}$
$FC = \frac{PeakValue}{RMS}$	$KS = \frac{\sum_{k=1}^{K}(f_k-\bar{f})^3 s(f_k)}{\sigma^3 K}$
$Ku = \frac{\frac{1}{T}\int_0^T (z(t)-\bar{z})^4 dt}{\left(\frac{1}{T}\int_0^T (z(t)-\bar{z})^2 dt\right)^2}$	$SS = \frac{\sum_{k=1}^{K}(f_k-\bar{f})^4 s(f_k)}{\sigma^4 K}$
$Skewness = \frac{\frac{1}{T}\int_0^T (z(t)-\bar{z})^3 dt}{\left(\frac{1}{T}\int_0^T (z(t)-\bar{z})^2 dt\right)^{3/2}}$	Tels que $\bar{f} = \frac{\sum_{k=1}^{K} f_k\, s(f_k)}{\sum_{k=1}^{K} s(f_k)}$ et $\sigma = \sqrt{\frac{\sum_{k=1}^{K}(f_k-\bar{f})^2\, s(f_k)}{K}}$

Ainsi, 11 indicateurs représentant les signaux vibratoires collectés des deux accéléromètres M1H et M2H sont obtenus. Ces indicateurs préservent les informations relatives aux défauts qui couvrent le domaine temporel et le domaine fréquentiel permettant par la suite d'identifier si les défauts présents dans l'équipement de coupe VSS causent la production de la non-qualité dans les planches ou non.

2.4 Méthodes de modélisation

2.4.1 Problème de classification basé sur l'analyse logique des données

LAD est une technique de classification qui est basée sur l'extraction de connaissances à partir d'un ensemble de données sous forme de modèles ou encore des « patterns ». Ces patterns représentent la caractérisation des phénomènes physiques, tels que la présence de la non-qualité dans le produit fini. Comme tous les algorithmes d'apprentissage supervisé, LAD est appliquée sur deux ensembles de données, d'entraînement et de test. L'ensemble d'entraînement, composé d'observation dont l'état est classé comme positif (Ω^+) ou négatif (Ω^-), est utilisé pour extraire les patterns les plus pertinents afin d'expliquer les phénomènes. Le reste des données est ensuite utilisé pour tester la précision des « patterns » générée (Boros et al., 2000).

Comme indiqué dans (Dupuis et al., 2012; Hammer et al., 2012; Mortada, M.-A. et al., 2014), la technique LAD présente deux attraits principaux :

- LAD n'est basée sur aucune analyse statistique, permettant de traiter des covariables fortement corrélées et variant dans le temps, sans avoir besoin de satisfaire à des hypothèses statistiques.
- LAD repose sur la recherche de patterns interprétables qui caractérisent, dans ce cas, la qualité du produit. Ces patterns extraits sont utilisés pour formuler un modèle de décision qui affecte les observations dans l'une des classes. Cette technique permet ainsi de contrôler le pouvoir discriminant entre les classes en sélectionnant le nombre minimum de patterns qui doivent séparer chaque observation d'une classe à une autre.

Les principales étapes de la technique LAD sont la binarisation des données, la génération de patterns et la formation des théories (Boros et al., 2000).

i. Binarisation des données

Dans l'étape de binarisation, chaque variable explicative de l'ensemble d'entraînement, noté X_k ($k = 1, \dots, K$), est remplacée par au moins une variable binaire, K étant le nombre de variables explicatives. Si la variable explicative X_k est catégorielle, sa binarisation est effectuée en se basant sur la méthode d'encodage one-hot, présentée dans (Haq et al., 2018). Et, si la variable explicative

X_k est continue, la procédure de binarisation commence par ranger, par ordre décroissant, toutes les valeurs distinctes ${u_{X_k}}^{(1)}, {u_{X_k}}^{(2)}, \dots, {u_{X_k}}^{(i)}, \dots, {u_{X_k}}^{(l)}$ de X_k. A chaque changement de classe, un « cut-points » est introduit. En d'autres termes, un cut-point est établi entre chaque paire de valeurs ${u_{X_k}}^{(i)}$ et ${u_{X_k}}^{(i+1)}$, lorsqu'il existe des observations égales à la valeur ${u_{X_k}}^{(i)}$ appartenant à Ω^+ et des observations égales à la valeur ${u_{X_k}}^{(i+1)}$ appartenant à Ω^- et vice-versa. La valeur du cut-point $\alpha_{X_k,j}$ correspond à la moyenne de ${u_{X_k}}^{(i)}$ et ${u_{X_k}}^{(i+1)}$. Ensuite, un attribut binaire est défini par rapport au cut-point déterminé en divisant l'espace en deux parties représentées par 0 et 1, comme le montre l'équation 1.

$$b_j = \begin{cases} 1 & si \quad u \geq \alpha_j \\ 0 & si \quad u < \alpha_j \end{cases} \qquad j = 1, \dots, q \qquad (1)$$

Le nombre des attributs binaires obtenus est ainsi égal au nombre de cut-points. On note q le nombre de tous les cut-points résultants de toutes les variables explicatives.

ii. Génération des patterns

La deuxième étape de LAD vise à déterminer les patterns cachés qui séparent les classes. Dans notre cas d'étude, la classe positive correspond à la non-qualité et la classe négative correspond à la bonne qualité du produit. Un pattern positif est défini comme une conjonction de littéraux qui est vrai pour au moins une observation positive et faux pour toutes les observations négatives dans l'ensemble de données d'apprentissage. Un modèle négatif est défini de la même manière. Un pattern couvre une certaine observation dans l'ensemble binarisé si tous ses littéraux sont vrais pour l'attribut binaire correspondant. Par exemple, considérons un ensemble de 3 attributs binaires ($b_{X_j,1}, b_{X_j,2}, b_{X_j,3}$) obtenu de la transformation de la variable X_j. Une conjonction de littéraux $\overline{b_{X_j,1}}\ b_{X_j,2}$ est considérée comme un pattern positif si au moins une observation positive possède des valeurs respectives (0, 1) pour les attributs ($b_{X_j,1}, b_{X_j,2}$), tandis qu'aucune observation négative ne possède ces valeurs.

Une des techniques la plus utilisée pour la génération de patterns est basée sur la programmation linéaire. En effet, le MILP « Mixed Integer Linear Programming » donne des solutions optimales, générant ainsi des patterns forts qui permettent LAD de mieux se généraliser sur de nouvelles observations, comme démontré dans (Ryoo & Jang, 2009).

La procédure pour générer un pattern positif p^+ est formulée en un problème de minimisation. Les variables de décision sont le vecteur booléen du pattern w, le degré du pattern d et un vecteur y dont la valeur y_i est définie comme suit.

$$y_i = \begin{cases} 0 & si \quad l'observation\ i\ est\ couverte\ par\ le\ pattern\ p \\ 1 & sinon, \quad i = 1, \dots, nombre\ d'observations \end{cases} \tag{2}$$

L'objectif est de minimiser le nombre d'observations positives qui ne sont pas couvertes par le pattern positif p, comme indiqué dans l'équation 3.

$$\min_{w,y,d} \sum_{i \in \Omega^+} y_i \tag{3}$$

La taille du vecteur du pattern w est le double du nombre d'attributs binaires obtenus q. Les éléments $w_1, w_2, \dots, w_j, \dots, w_q$ de w sont relatifs aux attributs binaires obtenus de la 1ère étape de LAD, tandis que les éléments $w_{q+1}, w_{q+2}, \dots, w_{2q}$ de w sont définis tel que w_{q+j} est la négation de w_j, $j = 1, \dots, q$. C'est-à-dire, si $w_j = 1$ alors $w_{q+j} = 0$ et vice-versa. D'où la condition suivante.

$$w_j + w_{q+j} \leq 1 \qquad j = 1,2, \dots, q \tag{4}$$

Il existe trois contraintes principales du problème MILP pour générer un pattern optimisé de la fonction objectif. La première est définie telle que le pattern résultant doit couvrir une observation positive $i \in \Omega^+$. Cette contrainte est présentée comme suit.

$$\sum_{j=1}^{2q} a_{ij}\, w_j + q y_i \geq d \qquad \forall i \in \Omega^+ \tag{5}$$

w est le vecteur du pattern et a_i est le vecteur de chaque observation $i \in \Omega^+$.

La deuxième contrainte implique que le pattern positive ne doit pas couvrir aucune observation négative, comme formulée ci-dessous.

$$\sum_{j=1}^{2q} a_{ij}\, w_j \leq d\ - 1 \qquad \forall i \in \Omega^- \tag{6}$$

La dernière contrainte consiste à ne pas générer le même pattern plus qu'une fois lors des itérations. La formulation mathématique de cette contrainte est la suivante.

$$\sum_{j=1}^{2q} v_{kj}\, w_j \leq d_k - 1 \qquad \forall i \in \mathrm{V} \tag{7}$$

V est l'ensemble contenant les vecteurs booléens de tous les patterns générés. Initialement, l'ensemble V est vide et cette contrainte n'est pas prise en compte. Cependant, à chaque itération, un pattern positif p_k^+ de degré d_k est généré et le vecteur v_k qui lui est associé est ajouté à l'ensemble V. Ainsi, une nouvelle condition est ajoutée pour chaque pattern déjà trouvé.

Le problème MILP est finalement présenté comme suit, en considérant que m est le nombre d'observations positives dans Ω^+.

$$\min_{w,y,d} \sum_{i\in\Omega^+} y_i \tag{8}$$

$$s.t. \begin{cases} w_j + w_{q+j} \leq 1 & j = 1,2,\dots,q \\ \sum_{j=1}^{2q} a_{ij}\, w_j + q y_i \geq d & \forall i \in \Omega^+ \\ \sum_{j=1}^{2q} v_{dj}\, w_j \leq d\ - 1 & \forall i \in \mathrm{V} \\ \sum_{j=1}^{2q} v_{dj}\, w_j \leq d\ - 1 & \forall i \in \mathrm{V} \\ 1 \leq d \leq q \\ \sum_{j=1}^{2q} w_j = d \\ w \in \{0,1\}^{2q} \\ y \in \{0,1\}^m \end{cases} \tag{9}$$

iii. Formation des théories

La formation de la théorie est l'étape finale de la technique LAD où les patterns générés à l'étape précédente sont utilisés pour former une fonction de décision, telle que donnée dans l'équation 10.

$$\Delta(O) = \sum_{i=1}^{N+} \gamma_i^+\, p_i^+(O) - \sum_{i=1}^{N-} \gamma_i^-\, p_i^-(O) \tag{10}$$

O est une représentation binaire d'une observation non classée. Les valeurs γ_i^+ et γ_i^- représentent respectivement des poids attribués à un patterns positifs p_i^+ ou un pattern négatif p_i^-, tel que $p_i^+(O) = 1$ si le pattern p_i^+ couvre l'observation O et 0 sinon. Le poids γ_i^+ est défini comme étant égales au nombre d'observations positives couvertes par le pattern p_i^+ divisé par la somme de la couverture de tous les modèles positifs. Le poids négatif p_i^- est calculé de la même manière.

Cette équation permet de calculer un score compris entre -1 et 1. Les nouvelles observations sont binarisées et entrées dans la fonction de décision. Un score positif indique une mauvaise qualité du produit, tandis qu'un score négatif signifie que la planche produite est de bonne qualité (Mortada, M.-A. et al., 2014)

Ces différentes étapes de la technique LAD sont effectuées à l'aide du logiciel cbmLAD développé par (Salamanca & Yacout, 2007). LAD nous permettra d'analyser le problème de la non-qualité d'une manière générale et afin d'identifier les causes racines de ce problème, la modélisation par régression serait utilisée.

2.4.2 Problème de régression

2.4.2.1 Problème de régression à multi-sorties

La régression multi-sorties, également connue sous le nom de « Multi-Target Regression » MTR, vise à modéliser plusieurs variables réponses continues en fonction des variables explicatives. Les approches utilisées pour résoudre le problème de MTR permettent une modélisation en considérant non seulement les relations sous-jacentes entre les variables explicatives et les variables réponses correspondantes, mais aussi les relations entre ces variables réponses, garantissant ainsi une meilleure représentation et interprétabilité des problèmes réels (Spyromitros-Xioufis et al., 2016).

Le MTR considère l'ensemble des données d'entraînement D de m variables explicatives $X_1, \dots, X_m$ et de d variables réponses $Y_1, \dots, Y_d$, i.e., $D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$, n étant le nombre d'observations de l'ensemble d'entraînement D. Une observation $l \in \{1, \dots, n\}$ est ainsi caractérisée par un vecteur d'entrée des m variables explicatives $x^{(l)} = \left(x_1^{(l)}, \dots, x_m^{(l)}\right)$ et par un vecteur de sortie des d variables réponses $y^{(l)} = \left(y_1^{(l)}, \dots, y_m^{(l)}\right)$.

La résolution du problème de MTR consiste à trouver une fonction f qui affecte un vecteur y à chaque observation donnée par le vecteur x, comme suit:

$$f: \Omega_{X_1} \times \dots \times \Omega_{X_m} \rightarrow \Omega_{Y_1} \times \dots \times \Omega_{Y_d}$$
$$x = (x_1, \dots, x_m) \rightarrow y = (y, \dots, y_d),$$

Ω_{X_j} et Ω_{Y_i} désignent les espaces d'échantillonnage de chaque indicateur X_j, pour tout $j \in \{1, \dots, m\}$, et de chaque variable réponse Y_i, pour tout $i \in \{1, \dots, d\}$, respectivement. Le modèle de régression multi-sorties construit sera ensuite utilisé pour prédire les valeurs $\{\hat{y}^{(n+1)}, \dots, \hat{y}^{(n')}\}$ de toutes les variables réponses de l'ensemble des données test $D' = \{(x^{(n+1)}, y^{(n+1)}), \dots, (x^{(n')}, y^{(n')})\}$.

De nombreuses approches ont été proposées pour résoudre le problème de MTR dans la littérature. L'approche de base la plus simple est basée sur la régression à sortie unique plus connue sous le nom de « Single-Target (ST) approach » qui modélise indépendamment chaque variable réponse, en décomposant le problème MTR en plusieurs problèmes de régression à sortie unique (Spyromitros-Xioufis et al., 2016). ST est une approche rapide et simple à appliquer, mais ne tient pas compte des dépendances inter-sorties. Une des approches les plus utilisées, qui considère ces dépendances, est l'approche de régression à chaînes (RC) (Spyromitros-Xioufis et al., 2016). Celle-ci décompose également le problème MTR en plusieurs modèles à sortie unique, mais elle enchaîne ces modèles. C'est-à-dire, chaque modèle à sortie unique inclut les valeurs prédites des variables réponses dans les variables explicatives. L'approche RC permet une explication plus complète des variables réponses favorisant l'extraction de nouvelles informations des données (Wu & Lian, 2020). Les sections 2.4.2.1.1 et 2.4.2.1.2 présentent ces deux approches.

2.4.2.1.1 Approche directe basée sur la régression à sortie unique

Cette approche consiste à résoudre le problème MTR en construisant d modèles indépendants chacun lié à une des d variables réponses, puis à utiliser ces modèles pour modéliser indépendamment chacune des variables réponses $Y_1, \dots, Y_d$. Ainsi, dans l'approche ST, un modèle multi-sorties f est composé de d modèles à une variable réponse unique où chaque modèle f_i est entraîné sur un ensemble de données d'apprentissage $D_i = \left\{\left(x^{(1)}, y_i^{(1)}\right), \dots, \left(x^{(n)}, y_i^{(n)}\right)\right\}$, pour modéliser une seule variable réponse Y_i, $i \in \{1, \dots, d\}$ (Spyromitros-Xioufis et al., 2016).

Une représentation graphique de cette approche est présentée dans la figure suivante, en considérant dans ce cas, $d = 3$. C'est-à-dire, seulement 3 variables réponses sont considérées dans le problème MTR.

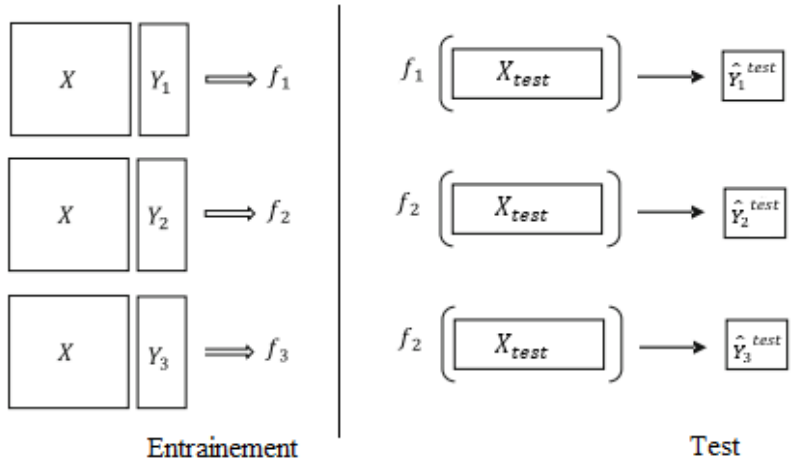


Figure 2.3 Illustration graphique de l'approche directe (Demirel et al., 2019)

Dans cette figure, il existe 3 variables réponses Y_1, Y_2, Y_3 et les indicateurs $X = (X_1, \dots, X_m)$. La méthode ST consiste ainsi à ajuster des modèles f_1, f_2, f_3 pour les variables réponses Y_1, Y_2, Y_3, respectivement, en se basant sur les indicateurs X. Durant la phase de test, la prédiction de chacune des 3 variables réponses est effectuée d'une manière indépendante, comme les modèles de régression usuels ayant un output unique.

2.4.2.1.2 Approche de régression à chaîne

L'approche RC est basée sur la construction de modèles de régression pour chaque variable réponse en entraînant séquentiellement les variables réponses dans l'ordre d'une chaîne déterminée. Cette approche consiste ainsi à sélectionner une chaîne aléatoire de l'ensemble de variables réponses, puis à construire un modèle de régression distinct pour chaque output en suivant l'ordre de la chaîne sélectionnée. C'est-à-dire, supposons que la chaîne $C = (Y_1, Y_2, \dots, Y_d)$ est sélectionnée, le premier modèle modélise seulement Y_1 en fonction des indicateurs. Ensuite, les modèles suivants pour $Y_i, s.t.\ i > 1$ sont entraînés sur les ensembles de données transformés $D_i^* = \left\{\left(x_i^{*\,(1)}, y_i^{(1)}\right), \dots, \left(x_i^{*\,(n)}, y_i^{(n)}\right)\right\}$, tel que $x_i^{*\,(l)} = \left(x_1^{(l)}, \dots, x_m^{(l)}, y_1^{(l)}, \dots, y_{i-1}^{(l)}\right)$ est un vecteur d'entrée transformé constitué du vecteur d'entrée d'origine de l'ensemble d'apprentissage avec les valeurs de toutes les variables réponses précédentes de la chaîne (Borchani et al., 2015).

Une illustration graphique de cette approche est présentée dans la figure ci-dessous, en considérant aussi le cas où $d = 3$.

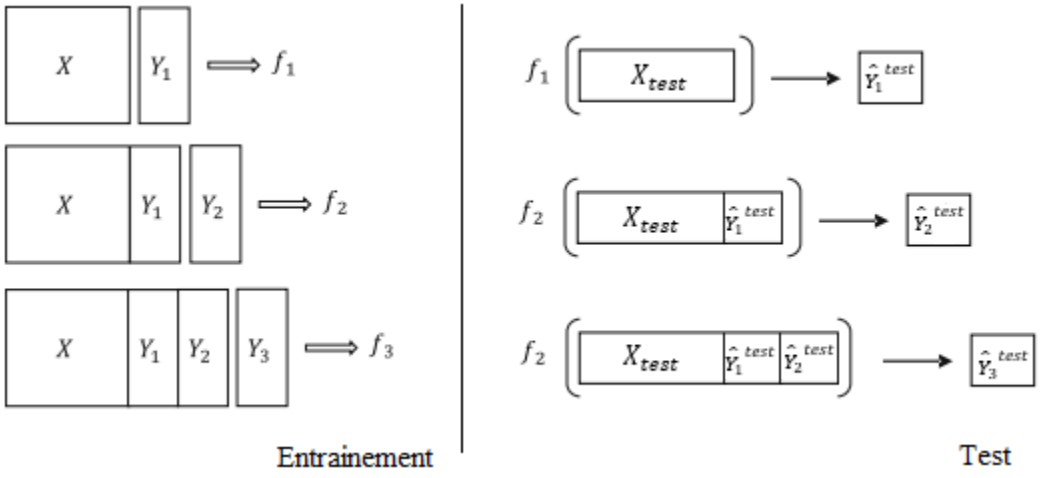


Figure 2.4 Illustration graphique de l'approche RC (Demirel et al., 2019)

Dans la phase d'entraînement, l'ajustement d'un modèle f_1 est effectuée pour la première variable réponse Y_1 en se basant sur les indicateurs X. Ensuite, un nouveau modèle f_2 est ajusté pour la variable réponse Y_2 en utilisant de nouveaux indicateurs qui est le résultat d'une concaténation des indicateurs de base X et les valeurs de Y_1 . De même, le modèle f_3 est construit en utilisant la variable Y_3 et des données concaténées X, Y_1 et Y_2.

Dans la phase de test, la prédiction des valeurs de la variable réponse est effectuée en se basant sur le modèle construit f_1. Ensuite, la variable prédite $\hat{Y}_1^{test}$ est ajoutée aux données d'entrée X_{test} pour prédire Y_2, en utilisant le modèle f_2. Enfin, les deux variables $\hat{Y}_1^{test}$ et $\hat{Y}_2^{test}$ sont concaténées aux indicateurs initiaux afin d'obtenir $\hat{Y}_3^{test}$, en utilisant f_3.

2.4.2.2 Méthodes de régression

Le MTR est un méta-apprenant, un ensemble de séquences d'apprentissage, pouvant utiliser différents estimateurs. Les algorithmes de régression que nous allons appliquer dans le chapitre 4 à notre ensemble de données sont les suivants :

- **La régression linéaire** effectue une modélisation linéaire entre une ou plusieurs variables explicatives $X = (X_1, X_2, \dots, X_p)$ et une variable réponse continue y. Le modèle est représenté par l'équation suivante :

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

$\beta_0, \beta_1, \ldots, \beta_p$ sont les coefficients de régression de ce modèle et ε est un terme d'erreur aléatoire. Afin de trouver la relation entre les variables explicatives et la variable réponse, les coefficients de régression sont généralement estimés en utilisant l'approche des moindres carrés, où la somme des carrés des résidus est minimisée (Yan & Su, 2009).

- **Les k plus proches voisins** ou encore « k-Nearest Neighbors k-NN » est un algorithme d'apprentissage supervisé utilisé pour les problèmes de régression. Pour prédire la variable réponse associée à une nouvelle observation entrée, l'algorithme k-NN sélectionne d'abord les k variables réponses dont les données d'entrée sont les plus proches de la nouvelle observation. Ensuite, le k-NN détermine la valeur de la variable réponse à prédire en calculant la moyenne des k variables réponses sélectionnées. Afin de choisir les k observations les plus proches, cet algorithme utilise des mesures de distance telles que la distance euclidienne, la distance de Minkowski ou la distance de Manhattan (Kramer, 2013).

- **Le boosting de gradient** ou encore « Gradient Boosting GB » est une technique d'apprentissage d'ensemble, dont leurs prédictions sont obtenues suite à une combinaison de plusieurs modèles d'apprentissages faibles, comme l'arbre de décision. L'algorithme GB se base sur le boosting comme un problème d'optimisation où l'objectif est de minimiser une fonction de coût en ajoutant des modèles d'apprentissage. Le boosting génère une séquence de modèles, où chaque nouveau modèle ajouté dans la séquence permet de mieux effectuer la prédiction que la précédente en affectant plus de poids aux observations mal ajustées. Le GB est composé de trois éléments principaux notamment une fonction de coût à optimiser, un apprenant faible pour faire des prédictions, et une stratégie pour ajouter un apprenant faible afin de minimiser la fonction de coût. La fonction de coût dépend du problème à résoudre. Pour la régression, par exemple, l'erreur quadratique moyenne est habituellement utilisée. Les arbres de décision sont utilisés comme modèles faibles dans l'amplification de gradient et la stratégie employée ajoute un arbre à la fois, et les arbres précédents ne sont pas modifiés. Une procédure de descente de gradient est utilisée pour minimiser la fonction de coût lorsque des arbres sont ajoutés (Friedman, 2002).

- **La forêt d'arbres décisionnels** ou encore « Random Forest RF » est aussi une technique d'apprentissage d'ensemble, se basant sur le bagging. Le bagging crée des arbres individuels et attribue un poids égal à tous les arbres, contrairement au boosting dans lequel les nouveaux arbres sont influencés par les performances des précédents et sont attribués un poids en fonction de leurs performances. L'algorithme du forêt d'arbres décisionnels fonctionne ainsi en effectuant un apprentissage sur plusieurs arbres de décision entraînés sur différents échantillons de données et en produisant la moyenne des valeurs prédites par tous les arbres (Breiman, 2001).

Ces algorithmes de régression sont souvent utilisés auprès des praticiens et des chercheurs et sont généralement connus pour leurs hauts niveaux de performance (Amaral et al., 2019; Kumar & Sahu, 2021; Mathew et al., 2017). Chacun de ces algorithmes choisis a ses propres caractéristiques. Le modèle de régression linéaire est connu pour être facilement interprétable. Cependant, pour que ce modèle soit valide, des hypothèses doivent être respectées notamment le linéaire dans les coefficients de régression, la normalité des résidus ou encore l'homoscédasticité des résidus (Amaral et al., 2019). L'algorithme k-NN, à l'opposé de l'algorithme LR, ne nécessite pas de relation linéaire entre les variables explicatives et la variable réponse, offrant une approche plus flexible. Le k-NN considère seulement un hyperparamètre, le nombre de voisins k. Toutefois, cet algorithme ne fonctionne pas très bien sur les ensembles de données avec un grand nombre d'indicateurs, ce qui est connu comme le problème de la malédiction de la dimension (Kramer, 2013). Comme le k-NN, les algorithmes de GB et de RF n'ont pas à respecter les hypothèses de régression linéaire. Ils peuvent aussi gérer une énorme quantité de données avec une dimensionnalité élevée d'indicateurs. D'autre part, ces algorithmes ne sont pas sensibles aux valeurs aberrantes, contrairement à la régression linéaire et au k-NN. Ces algorithmes sont connus pour leurs flexibilités, ce qui les rend capables d'apprendre et de modéliser des relations non-linéaires et complexes. Néanmoins, ils sont plus difficiles à interpréter et à visualiser. L'algorithme GB considère particulièrement un plus grand nombre d'hyperparamètres notamment le nombre d'arbres, la profondeur des arbres et le taux d'apprentissage. Si ces derniers ne sont pas réglés correctement, un surapprentissage est possible. En revanche, il y a généralement seulement deux hyperparamètres dans le RF, qui sont le nombre d'arbres et le nombre d'indicateurs à sélectionner à chaque nœud, ce qui le rend moins vulnérable au surapprentissage que le GB. Cela dit, étant

donné que les arbres sont ajoutés en optimisant une fonction objectif, le GB peut être beaucoup plus flexible que le RF. En effet, l'algorithme GB peut être utilisé pour résoudre presque toutes les fonctions objectifs en se basant sur la descente de gradient (Callens et al., 2020).

Afin d'améliorer encore plus les performances d'apprentissage, des méthodes d'ensembles peuvent être utilisées. Celles-ci prennent les variables prédites par plusieurs modèles d'apprentissages conceptuellement différents et créent un méta-modèle les intégrant tous. Ces variables prédites peuvent être ainsi combinées pour produire une nouvelle variable. Un modèle d'ensemble de régression consiste à permettre à tous les modèles de faire une prédiction et ensuite à calculer la moyenne de toutes ces prédictions pour obtenir une nouvelle variable réponse. Ce modèle d'ensemble équilibre les faiblesses individuelles des modèles qui le constituent (Grmanová et al., 2016).

CHAPITRE 3 CLASSIFICATION

Afin de répondre au premier objectif défini dans la Section 1.5 qui consiste à détecter la non-qualité dans le produit fini durant le processus de sciage, une modélisation basée sur la classification de chacun des 3 types de non-qualité en fonction des indicateurs est effectuée. Chacune des variables réponses continues correspondantes aux quantités des planches non conformes est ainsi transformée en variables catégoriques, plus précisément en forme binaire, en se basant sur des seuils de dichotomisation déterminés selon une approche statistique. Cette transformation permet de faciliter l'analyse de données et la modélisation des différents types de non-qualités (Brauer, 2002).

Pour ce faire, nous fournirons une explication générale de la méthodologie suivie pour modéliser les différents types de non-qualités par classification. Ensuite, la technique de classification LAD sera utilisée en se basant sur le logiciel cbmLAD, afin de générer des règles de regroupement représentant chaque type de non-qualité du produit fini notamment le délignement, le déchiquetage et le rejet.

3.1 Méthodologie

Le processus de modélisation par classification de la non-qualité des planches est présenté dans la Figure 3.1. Ce processus comprend principalement six étapes : la dichotomisation des variables réponses, l'extraction des caractéristiques, la sélection des caractéristiques, le traitement des données manquantes, le développement du modèle de classification et l'évaluation de ce modèle.

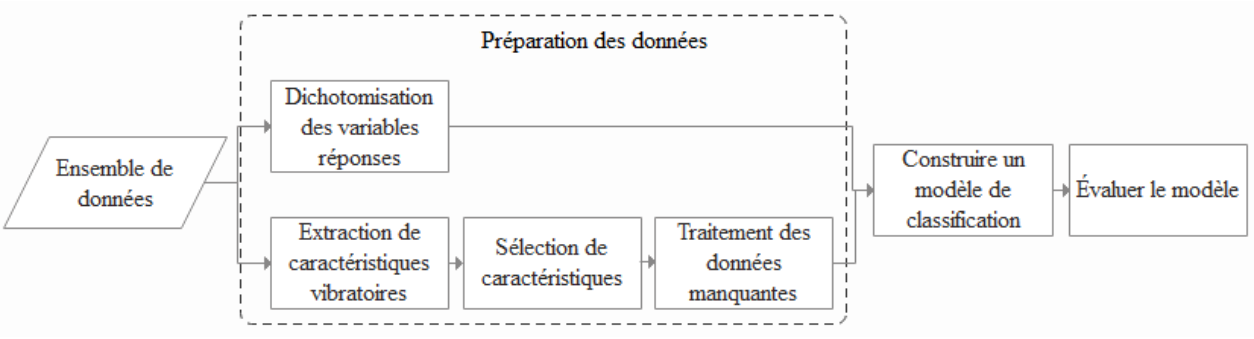


Figure 3.1 Étapes d'analyse de données par classification

L’objectif de cette section est d’utiliser la technique de classification LAD présenté dans le chapitre précédent afin de diagnostiquer les différents types de mauvaise qualité trouvé dans le produit fini. Pour construire ces modèles de classification, il est nécessaire de transformer les variables réponses en variables catégoriques. Ensuite, afin d’analyser l’effet des vibrations sur la qualité des planches, une extraction et une sélection des caractéristiques statistiques sont effectuées à partir des signaux vibratoires.

Une fois les données préparées, les modèles de classification caractérisant la non-qualité des planches peuvent être construits en se basant sur la technique LAD. Le logiciel de classification correspondant utilisé pour modéliser les 3 types de non-qualité des planches est appelé cbmLAD. Ce dernier a été introduit pour la première fois par (Salamanca & Yacout, 2007), où il a été adapté pour traiter l'application du diagnostic des défauts dans les roulements. Ce logiciel contient deux approches pour la classification, le « One versus One » (OvO) qui suppose qu'il existe un séparateur entre chaque deux classes et le « One versus All » (Grmanová et al.) qui suppose l'existence d'un seul séparateur entre une classe et toutes les autres classes. Dans notre cas d’études, nous utiliserons l’approche OvA, étant donné que les deux approches donnent les mêmes résultats ayant un problème de classification de deux classes.

Une évaluation de ce modèle est enfin effectuée en recourant à des métriques basées sur une comparaison entre les valeurs prédites à partir du modèle de classification avec les valeurs réelles trouvées dans l'ensemble de données d'origine.

3.1.1 Extraction des caractéristiques vibratoires

Dans cette étape, des caractéristiques statistiques du domaine temporel et du domaine fréquentiel sont extraites des signaux vibratoires issus à la fois des deux capteurs M1H et M2H, placés sur le moteur de l’équipement de coupe de bois VSS. Le langage de programmation Python a été utilisé pour préparer ces données en se basant sur les techniques d’extraction de caractéristiques présentées dans le chapitre précédent.

Ainsi, 11 caractéristiques statistiques dont 6 qui représentent les signaux vibratoires du domaine temporel et 5 qui représentent les signaux du domaine fréquentiel sont obtenues, pour chacun des deux capteurs, résultant ainsi à un total de 22 caractéristiques vibratoires.

3.1.2 Sélection des caractéristiques

Lorsque toutes les caractéristiques extraites sont utilisées pour la modélisation en se basant sur LAD, il est possible que la précision du modèle d'apprentissage diminue et que le temps de calcul soit augmenté en raison de la redondance ou de la non-pertinence de certaines caractéristiques. Afin d'améliorer les performances du modèle développé, certaines caractéristiques significatives fournissant des informations importantes pour expliquer la non-qualité des planches doivent être sélectionnées, et les caractéristiques non pertinentes ou redondantes doivent être supprimées (Nayana & Geethanjali, 2017). Une sélection des caractéristiques en se basant sur la méthode de régression pas à pas est ainsi effectuée dans cette section. Les caractéristiques sélectionnées sont utilisées également dans la régression dans le Chapitre 4.

Il existe deux approches principales pour la sélection de caractéristiques, l'approche de filtrage et l'approche d'emballage. L'approche de filtrage consiste à conserver uniquement le sous-ensemble des caractéristiques pertinentes, en se basant sur une mesure univariée qui reflète le pouvoir discriminant de chaque caractéristique (Kumari & Swarnkar, 2011). Par exemple, (Ragab, A. et al., 2019) emploient une des techniques de filtrage appelée « Compensation Distance Evaluation Technique » (CDET) sur des données vibratoires avant de procéder à l'implémentation de LAD. Dans ce cas, la mesure univariée utilisée est le score de CDET. Cette approche présente des avantages au niveau de son efficacité calculatoire. Cependant, elle ignore les dépendances entre les caractéristiques, ce qui peut entraîner de faibles performances. C'est-à-dire, une variable qui n'est pas utile, toute seule peut l'être lorsqu'elle est combinée avec d'autres.

La deuxième approche d'emballage est plus coûteuse en temps de calcul. En revanche, elle est généralement plus performante. Elle consiste à introduire les caractéristiques dans l'algorithme d'apprentissage et, en fonction de la performance du modèle, des caractéristiques sont ajoutées ou bien éliminées, comme le résume la Figure 3.2 (Kumari & Swarnkar, 2011). Il s'agit d'un processus itératif, nécessitant des calculs souvent lourds. Les méthodes d'emballage effectuent une recherche plus complète de l'espace d'ensemble de caractéristiques, tout en considérant leurs dépendances, afin de trouver un sous-ensemble optimal pour le modèle d'apprentissage à développer. Elles se basent également sur la performance du modèle construit comme critère de sélection de

caractéristiques. L'approche d'emballage est donc généralement considérée comme étant meilleure que celle de filtrage.

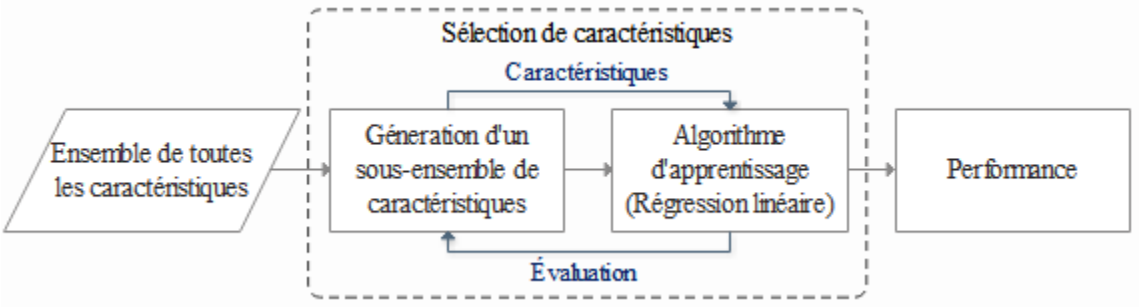


Figure 3.2 Approche d'emballage pour la sélection des caractéristiques

Dans notre cas d'études, la sélection de caractéristiques est effectuée par approche d'emballage en utilisant la méthode de régression pas à pas. Cette méthode est utilisée depuis longtemps et constitue la procédure standard dans certains logiciels statistiques (Cai et al., 2010). Les caractéristiques peuvent être sélectionnées en avant ou en arrière. La sélection vers l'avant commence sans variable et celle qui fournit la meilleure performance du modèle, entre dans le modèle. La procédure est répétée en introduisant, à chaque fois, à l'ensemble de caractéristiques sélectionnées, la variable qui améliore le mieux le modèle jusqu'à ce qu'un ajout d'une nouvelle variable n'améliore plus les performances du modèle. L'élimination vers l'arrière suit la même idée, mais fonctionne dans le sens inverse. Au lieu de commencer sans aucune variable et d'en ajouter à chaque itération, cette méthode commence par toutes les variables du modèle et supprime à chaque fois celle qui dégrade la performance du modèle. Ces deux méthodes aboutissent à un sous-ensemble de variables qui seront utilisées pour développer le modèle d'apprentissage automatique (Cai et al., 2010). La méthode utilisée, dans notre cas, est la régression pas à pas, basée sur l'élimination bidirectionnelle qui est une combinaison de la sélection avant et arrière. À chaque étape de la procédure, la méthode de la régression pas-à pas examine à la fois si une nouvelle variable doit être ajoutée, et si une des variables déjà incluses doit être éliminée.

Cette méthode de sélection de caractéristiques est mise en œuvre à l'aide du modèle de régression linéaire, comme le montre la Figure 3.2, avec validation croisée afin d'identifier le nombre optimal de caractéristiques. La validation croisée est un outil puissant qui permet une meilleure utilisation des données, fournissant beaucoup plus d'informations sur les performances du modèle

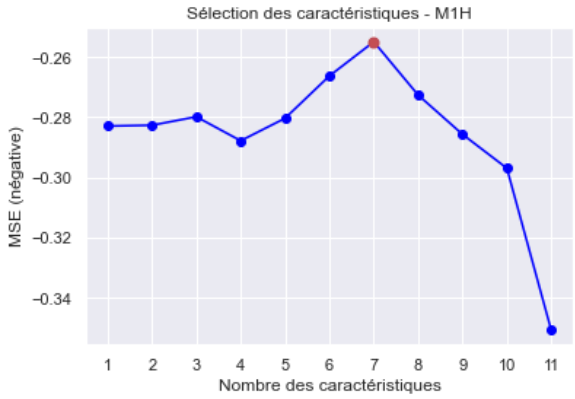
d’apprentissage. Habituellement, l'ensemble de données est divisé en deux ensembles, ensemble d'apprentissage et ensemble de test. Cependant, cela pourrait engendrer une perte de données coûteuses qui pourraient être utilisées pour développer des modèles mieux fondés (Spycher et al., 2004). Une solution possible consiste à utiliser les méthodes de validation croisée, particulièrement la technique k blocs « k-fold cross-validation ». En utilisant cette technique, l'ensemble de données serait utilisé à la fois pour l’entraînement et la validation du modèle. Ce dernier présentera non seulement un biais pas très important, mais aussi une faible variance (Harrell, 2001). La technique k-fold consiste à diviser l’ensemble de données en k groupes de tailles approximativement égales. À tour de rôle, un groupe est utilisé comme ensemble test, le reste des données ($k-1$ groupes) est utilisé comme ensemble d’entraînement. Le processus est répété k fois. Les valeurs recommandées de k varient entre 3 et 10 (Mashudi et al., 2021; Rahman & Akter, 2020). Dans notre cas d’étude, la valeur de k est fixée à 3. La métrique choisie pour évaluer la performance du modèle de régression est l’erreur quadratique moyenne MSE « Mean Squared Error ». Le MSE représente la différence quadratique moyenne entre les valeurs réelles de la variable réponse et celles prédites par le modèle. Cette métrique est idéale pour garantir que le modèle entraîné n'a pas de prédictions aberrantes avec d'énormes erreurs, car le MSE accorde une plus grande importance à ces erreurs en raison de la partie quadratique. À la i^{ème} répétition, $i = 1,2,3$, le MSE_i est calculé. Une valeur finale est obtenue du MSE, comme suit.

$$MSE = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

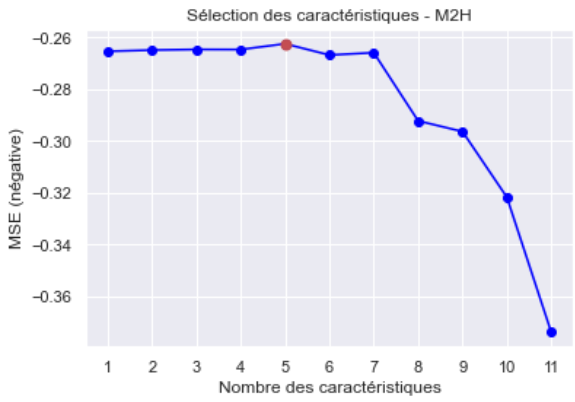
Le meilleur ensemble de caractéristiques est trouvé en minimisant la valeur du MSE. Une fois les caractéristiques sélectionnées, l’algorithme d’apprentissage peut être construit pour modéliser les 3 types de non-qualité des planches.

Dans ce qui suit, la méthode de régression pas à pas basée sur la validation croisée est appliquée sur les variables précédemment extraites qui représentent les signaux vibrations issus de chacun des deux capteurs M1H et M2H. Étant donné que cette méthode dépend du modèle d’apprentissage à construire, la méthode de sélection est appliquée pour chacun des 3 modèles correspondants aux types de non-qualité.

Les résultats obtenus de cette méthode sont présentés dans les Figures 3.3, 3.4 et 3.5. Chacune de ces figures illustre un graphique des valeurs MSE relatives au modèle de régression d'une des 3 variables de non-qualité en fonction du nombre de caractéristiques sélectionnées correspondant aux signaux vibratoires délivrés par les capteurs. Ce graphique permet de déterminer le nombre optimal de caractéristiques à sélectionner. Il est aussi accompagné d'un tableau qui expose les valeurs MSE_i calculées à chaque itération de la validation croisée. Par exemple, en regardant la Figure 3.3, les performances culminent à 7 caractéristiques vibratoires de M1H avec une valeur de MSE égale à 0,255, tandis que 5 caractéristiques vibratoires de M2H sont sélectionnées avec une valeur de MSE égale à 0,262. Notons que les signes négatifs de MSE représentent simplement les valeurs de MSE multipliées avec -1 juste pour suivre la convention de la bibliothèque de Python, Scikit-learn. L'idée est que minimiser la MSE équivaut à maximiser la MSE négative.

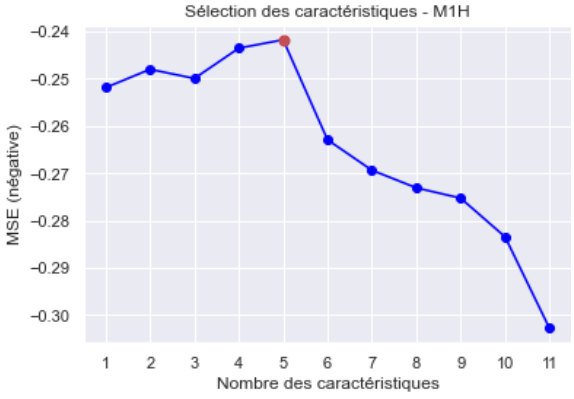


	MSE_i	MSE	Features name
1	[-0.333, -0.168, -0.348]	-0.283	(CF - M1H,)
2	[-0.332, -0.168, -0.348]	-0.283	(RMS spectral - M1H, CF - M1H)
3	[-0.329, -0.171, -0.339]	-0.28	(RMS spectral - M1H, RMS - M1H, CF - M1H)
4	[-0.342, -0.182, -0.339]	-0.288	(RMS spectral - M1H, RMS - M1H, CF - M1H, Ku -...
5	[-0.358, -0.139, -0.344]	-0.28	(RMS spectral - M1H, PeakValue - M1H, RMS - M1...
6	[-0.34, -0.154, -0.305]	-0.266	(RMS spectral - M1H, PeakValue - M1H, RMS - M1...
7	[-0.332, -0.19, -0.243]	-0.255	(RMS spectral - M1H, PeakValue - M1H, RMS - M1...
8	[-0.377, -0.213, -0.228]	-0.273	(RMS spectral - M1H, SS - M1H, PeakValue - M1H...
9	[-0.413, -0.213, -0.231]	-0.286	(RMS spectral - M1H, Variance spectral- M1H, S...
10	[-0.426, -0.237, -0.229]	-0.297	(Moyenne - M1H, RMS spectral - M1H, Variance s...
11	[-0.464, -0.333, -0.255]	-0.351	(Moyenne - M1H, RMS spectral - M1H, Variance s...

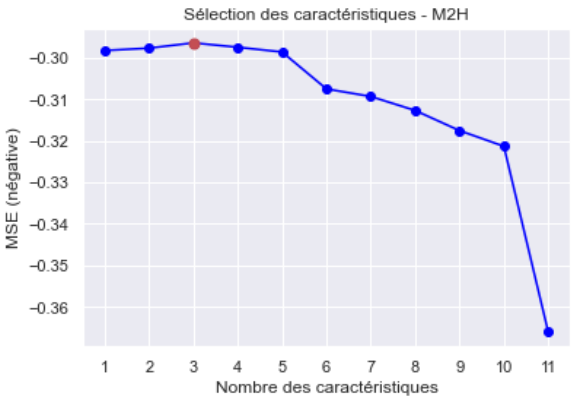


	MSE_i	MSE	Features name
1	[-0.306, -0.162, -0.328]	-0.265	(KS - M2H,)
2	[-0.309, -0.157, -0.329]	-0.265	(KS - M2H, RMS -M2H)
3	[-0.309, -0.158, -0.327]	-0.265	(RMS spectral - M2H, KS - M2H, RMS -M2H)
4	[-0.307, -0.161, -0.325]	-0.265	(Moyenne - M2H, RMS spectral - M2H, KS - M2H, ...
5	[-0.308, -0.163, -0.317]	-0.262	(Moyenne - M2H, RMS spectral - M2H, KS - M2H, ...
6	[-0.306, -0.175, -0.32]	-0.267	(Moyenne - M2H, RMS spectral - M2H, KS - M2H, ...
7	[-0.303, -0.176, -0.319]	-0.266	(Moyenne - M2H, RMS spectral - M2H, KS - M2H, ...
8	[-0.303, -0.253, -0.32]	-0.292	(Moyenne - M2H, RMS spectral - M2H, KS - M2H, ...
9	[-0.294, -0.315, -0.28]	-0.296	(Moyenne - M2H, RMS spectral - M2H, KS - M2H, ...
10	[-0.362, -0.323, -0.28]	-0.322	(Moyenne - M2H, RMS spectral - M2H, KS - M2H, ...
11	[-0.372, -0.457, -0.292]	-0.374	(Moyenne - M2H, RMS spectral - M2H, Variance s...

Figure 3.3 Sélection des caractéristiques pour modéliser le délignement

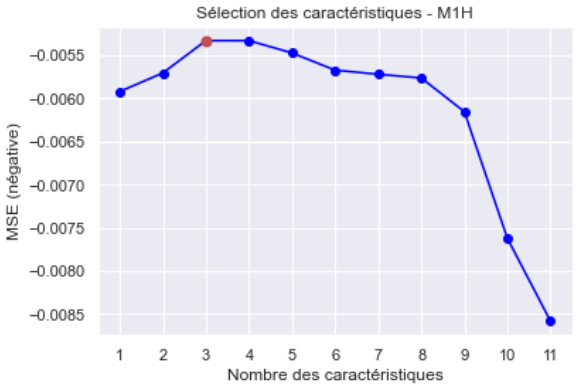


	MSE_i	MSE	Features name
1	[-0.428, -0.151, -0.177]	-0.252	(RMS spectral - M1H,)
2	[-0.425, -0.151, -0.168]	-0.248	(RMS spectral - M1H, PeakValue - M1H)
3	[-0.393, -0.156, -0.201]	-0.25	(Moyenne - M1H, RMS spectral - M1H, PeakValue ...
4	[-0.398, -0.141, -0.192]	-0.244	(RMS spectral - M1H, PeakValue - M1H, Variance...
5	[-0.396, -0.14, -0.189]	-0.242	(Moyenne - M1H, RMS spectral - M1H, PeakValue ...
6	[-0.387, -0.196, -0.206]	-0.263	(Moyenne - M1H, RMS spectral - M1H, SS - M1H, ...
7	[-0.358, -0.199, -0.252]	-0.269	(Moyenne - M1H, RMS spectral - M1H, SS - M1H, ...
8	[-0.36, -0.208, -0.252]	-0.273	(Moyenne - M1H, RMS spectral - M1H, KS - M1H, ...
9	[-0.36, -0.212, -0.254]	-0.275	(Moyenne - M1H, RMS spectral - M1H, KS - M1H, ...
10	[-0.358, -0.215, -0.277]	-0.283	(Moyenne - M1H, RMS spectral - M1H, Variance s...
11	[-0.39, -0.237, -0.282]	-0.303	(Moyenne - M1H, RMS spectral - M1H, Variance s...

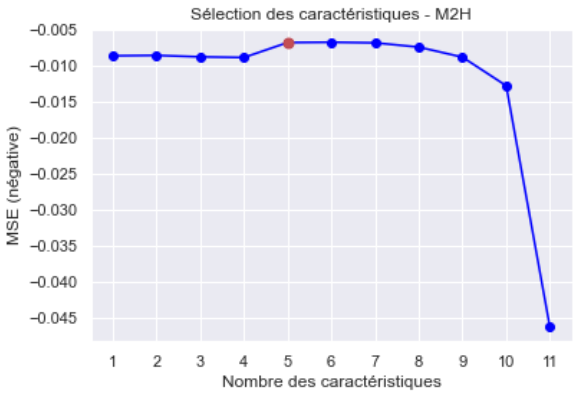


	MSE_i	MSE	Features name
1	[-0.433, -0.248, -0.214]	-0.298	(SS - M2H,)
2	[-0.431, -0.249, -0.213]	-0.298	(Moyenne - M2H, SS - M2H)
3	[-0.429, -0.248, -0.213]	-0.296	(Moyenne - M2H, Variance spectral - M2H, SS - ...
4	[-0.429, -0.251, -0.212]	-0.298	(Moyenne - M2H, RMS spectral - M2H, Variance s...
5	[-0.428, -0.255, -0.212]	-0.299	(Moyenne - M2H, RMS spectral - M2H, Variance s...
6	[-0.403, -0.237, -0.283]	-0.308	(Moyenne - M2H, RMS spectral - M2H, Variance s...
7	[-0.397, -0.245, -0.286]	-0.309	(Moyenne - M2H, RMS spectral - M2H, Variance s...
8	[-0.397, -0.248, -0.294]	-0.313	(Moyenne - M2H, RMS spectral - M2H, Variance s...
9	[-0.396, -0.259, -0.298]	-0.318	(Moyenne - M2H, RMS spectral - M2H, Variance s...
10	[-0.396, -0.268, -0.3]	-0.321	(Moyenne - M2H, RMS spectral - M2H, Variance s...
11	[-0.449, -0.331, -0.319]	-0.366	(Moyenne - M2H, RMS spectral - M2H, Variance s...

Figure 3.4 Sélection des caractéristiques pour modéliser le déchiquetage



	MSE_i	MSE	Features name
1	[-0.013, -0.001, -0.004]	-0.006	(RMS - M1H,)
2	[-0.012, -0.001, -0.004]	-0.006	(RMS - M1H, Variance - M1H)
3	[-0.012, -0.001, -0.004]	-0.005	(RMS - M1H, Variance - M1H, CF - M1H)
4	[-0.012, -0.001, -0.004]	-0.005	(Variance spectral- M1H, RMS - M1H, Variance -...
5	[-0.012, -0.001, -0.004]	-0.005	(Moyenne - M1H, RMS spectral - M1H, RMS - M1H,...
6	[-0.012, -0.001, -0.004]	-0.006	(RMS spectral - M1H, KS - M1H, SS - M1H, PeakV...
7	[-0.013, -0.001, -0.003]	-0.006	(RMS spectral - M1H, Variance spectral- M1H, K...
8	[-0.013, -0.001, -0.004]	-0.006	(Moyenne - M1H, RMS spectral - M1H, Variance s...
9	[-0.013, -0.002, -0.004]	-0.006	(Moyenne - M1H, RMS spectral - M1H, Variance s...
10	[-0.012, -0.003, -0.007]	-0.008	(Moyenne - M1H, RMS spectral - M1H, Variance s...
11	[-0.013, -0.003, -0.009]	-0.009	(Moyenne - M1H, RMS spectral - M1H, Variance s...



	MSE_i	MSE	Features name
1	[-0.019, -0.001, -0.006]	-0.009	(RMS -M2H,)
2	[-0.019, -0.001, -0.006]	-0.009	(Moyenne - M2H, RMS -M2H)
3	[-0.02, -0.001, -0.005]	-0.009	(Moyenne - M2H, RMS -M2H, Variance - M2H)
4	[-0.02, -0.002, -0.005]	-0.009	(Moyenne - M2H, RMS -M2H, Variance - M2H, CF -...
5	[-0.013, -0.003, -0.004]	-0.007	(Moyenne - M2H, KS - M2H, PeakValue -M2H, CF -...
6	[-0.013, -0.003, -0.005]	-0.007	(Moyenne - M2H, KS - M2H, SS - M2H, PeakValue ...
7	[-0.013, -0.003, -0.005]	-0.007	(Moyenne - M2H, RMS spectral - M2H, KS - M2H, ...
8	[-0.013, -0.005, -0.005]	-0.007	(Moyenne - M2H, Variance spectral - M2H, KS - ...
9	[-0.014, -0.008, -0.005]	-0.009	(Moyenne - M2H, RMS spectral - M2H, Variance s...
10	[-0.014, -0.021, -0.004]	-0.013	(Moyenne - M2H, RMS spectral - M2H, Variance s...
11	[-0.017, -0.117, -0.004]	-0.046	(Moyenne - M2H, RMS spectral - M2H, Variance s...

Figure 3.5 Sélection des caractéristiques pour modéliser le rejet

Le Tableau 3.1 énumère les caractéristiques sélectionnées par la méthode de régression pas à pas pour chacune des 3 variables de non-qualité.

Tableau 3.1 Liste de caractéristiques sélectionnées

Capteur	Délignement	Déchiquetage	Rejet
M1H	• PeakValue - M1H • RMS - M1H • Variance - M1H • FC - M1H • Kurtosis - M1H • Skewness - M1H • RMS spectral - M1H	• Moyenne - M1H • RMS spectral- M1H • PeakValue - M1H • Variance - M1H • FC - M1H	• RMS - M1H • Variance - M1H • FC - M1H
M2H	• Moyenne - M2H • RMS spectral- M2H • KS - M2H • RMS - M2H • Variance -M2H	• Moyenne - M2H • Variance spectral - M2H • SS - M2H	• Moyenne - M2H • KS - M2H • PeakValue - M2H • FC - M2H • Kurtosis - M2H

3.1.3 Traitement des données manquantes

Une fois les variables vibratoires statistiques sont sélectionnées pour chaque capteur, elles sont rassemblées avec les autres variables de production et de maintenance. Des valeurs manquantes sont détectées dans les données de vibration. En effet, les caractéristiques correspondantes au capteur M1H contiennent 28 valeurs manquantes, soit 36 % de tout l'ensemble de données. Les caractéristiques correspondantes au capteur M2H contiennent également 24 valeurs manquantes, soit 31 %. Pour le traitement de ces valeurs manquantes, la méthode d'imputation des k plus proches voisins k-NN a été choisie.

L'imputation k-NN est une méthode d'imputation non paramétrique qui est flexible à la fois en données continues et en données discrètes. (Rawal et al., 2017) montrent que les méthodes d'imputation non paramétriques sont plus efficaces que ceux qui sont paramétriques comme

l'algorithme d'espérance-maximisation pour des données de taille moyenne. (Pujianto et al., 2019) montrent également que la précision des résultats d'imputation k-NN se rapproche de la précision des données complètes.

L'imputation k-NN détermine la valeur moyenne des k voisins les plus proches qui sont les plus similaires à l'entité avec la valeur manquante. La moyenne calculée est ensuite imputée dans la valeur manquante. Les k voisins les plus proches ou les plus similaires sont sélectionnés en utilisant la métrique de similarité ou de distance (Troyanskaya et al., 2001). Dans cette étude, la métrique de distance sélectionnée est la distance euclidienne et le nombre de voisins k a été fixé à 5. La sélection de la taille du voisinage k joue un rôle important dans l'obtention d'une bonne performance de k-NN. Cependant, comme le souligne (Loukopoulos et al., 2017), aucune règle globale n'est définie pour déterminer ce k optimal. Dans le présent rapport, des expériences préliminaires comme celle réalisée par (Thanh Noi & Kappas, 2018), avec différentes valeurs de k entre 1 et 20, sont menées. Ensuite, la valeur k qui a donné la valeur la plus faible de MSE est sélectionnée.

3.1.4 Dichotomisation des variables réponses

Afin d'effectuer la modélisation des différents types de non-qualité des planches de bois par classification, une dichotomisation des variables réponses continue en variables catégoriques est nécessaire. Il est bien connu théoriquement que dans certaines situations, lorsque la dichotomisation a lieu, certaines informations contenues dans les données d'origine sont perdues. Cette perte d'informations peut entraîner une perte d'efficacité dans l'estimation et une perte de puissance dans les tests d'hypothèses (Cohen, 1983; Taylor et al., 2006). Cependant, dans certains cas, la dichotomisation simplifie la modélisation et l'analyse des données. En effet, l'inclusion des variables continues pourrait conduire à des difficultés d'interprétation des résultats (MacCallum et al., 2002). De plus, des études ont soutenu que lorsque le nombre d'observations est petit, ce qui est notre cas, l'utilisation des variables catégoriques augmente la robustesse des modèles (DeCoster et al., 2011). Il existe également un besoin d'étiqueter les jours comme ayant ou non une production importante d'une certaine caractéristique de non-qualité pour déterminer les variables qui expliquent le mieux l'apparition de chacune de ces caractéristiques.

La dichotomisation est répandue dans plusieurs domaines. Par exemple, dans la recherche médicale, certaines mesures cliniques, telles que la pression artérielle ou le taux d'hémoglobine, ont généralement des seuils conventionnels couramment utilisés par les médecins pour établir un diagnostic (Zhang et al., 2000). Ces mesures sont systématiquement dichotomisées dans l'analyse des données. Dans des études du traitement et de gestion de la qualité des eaux, (Esterby, 1989) utilise la méthode binomiale pour obtenir une variable binaire Z à partir d'une variable continue Y, définie comme étant la concentration du paramètre de qualité. Cette méthode consiste à définir une limite supérieure d'une concentration acceptable L à partir d'un nombre d'observations de la variable Y et à utiliser cette limite pour dichotomiser la variable continue en deux groupes, comme suit :

$$P(Z = 1) = P(Y > L)$$

$$P(Z = 0) = P(Y <= L)$$

Dans cet article, l'auteur a fixé L comme étant le 95^{e} centile de la variable Y. Cette méthode binomiale a été également utilisée par (Warn & Matthews, 1984) pour évaluer la conformité des effluents. Dans les études en sciences sociales et en psychologie, (Brauer, 2002) dichotomise les variables de réponse continues en se basant sur la médiane afin de diviser les participants de l'étude en deux groupes, tandis que (Stevanovic et al., 2019) utilisent la moyenne au lieu de la médiane.

Tout bien considéré, il existe principalement deux approches pour déterminer le seuil pour la dichotomisation. Dans de nombreuses situations, il existe des seuils reconnus qui sont souvent utilisés dans des études antérieures. Par exemple, le seuil pour dichotomiser les taux d'hémoglobine des patients est habituellement fixé à la limite supérieure de l'intervalle de référence chez les individus sains (MacCallum et al., 2002; Zhang et al., 2000). Ce type de seuil, indépendant des données de l'échantillon, est souvent considéré comme constant.

D'autres fois, en l'absence d'un seuil connu, une approche courante consiste à déterminer une mesure statistique de la variable continue étudiée tel que le 95e centile, la médiane ou la moyenne de l'échantillon. Ensuite, cette mesure est utilisée comme seuil pour la dichotomisation de la variable continue. Notre cas d'études couvre ce type de seuil. Le choix de ce dernier dépend

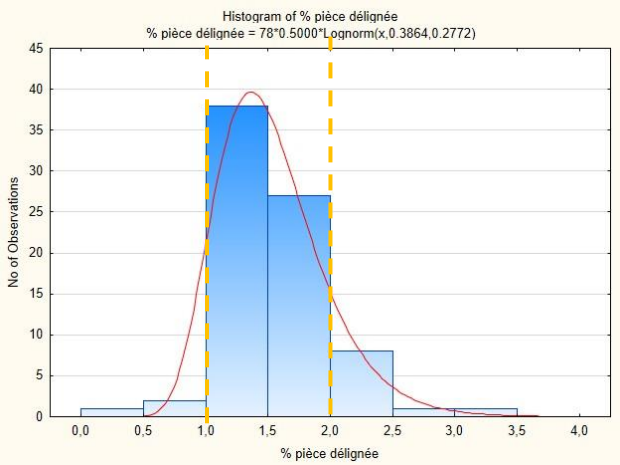
généralement de l'ensemble de données et peut être très complexe. Ainsi, il est important de bien comprendre la variation et la répartition des valeurs de chaque variable à transformer.

Pour ce faire, des mesures statistiques des variables étudiées sont déterminées. En se référant aux travaux précédents (Brauer, 2002; Stevanovic et al., 2019; Warn & Matthews, 1984), les deux mesures les plus utilisées pour dichotomiser les variables continues en variables binaires sont la moyenne et la médiane.

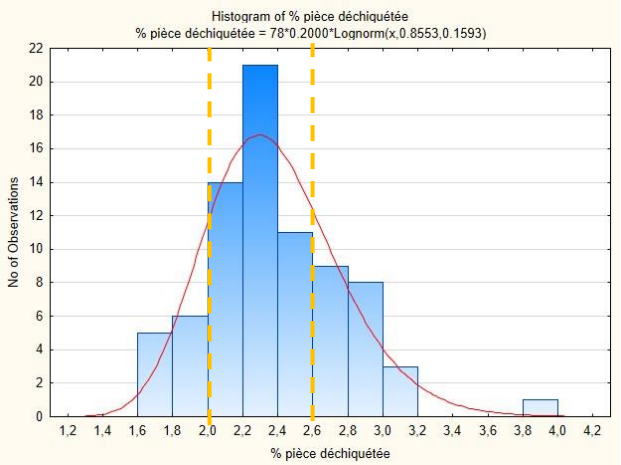
Nous ferons également recours à la visualisation des données à travers des histogrammes. En effet, un aperçu visuel de ces variables nous fournira des renseignements significatifs qui nous permettent de fixer un seuil convenable pour la dichotomisation telles que les valeurs minimum et maximum des données, leurs variabilités, ou encore leurs distributions. Cette approche a été proposée dans (Ciupke, 2005). En analysant des histogrammes des variables continues, il a été constaté que les valeurs ne sont pas réparties uniformément, s'apercevant que certaines valeurs se produisent plus fréquemment que d'autres. C'est pourquoi, l'auteur de cet article a suggéré d'utiliser ces propriétés pour déterminer les seuils de dichotomisation.

Pour chacune des 3 variables de non-qualité, un histogramme est illustré avec la fonction de densité théorique de la distribution correspondante (Figure 3.6). Chaque bande verticale de l'histogramme représente un intervalle de valeurs relatives à la variable continue étudiée. La hauteur de chaque bande indique le nombre d'observations qui se trouvent dans cet intervalle.

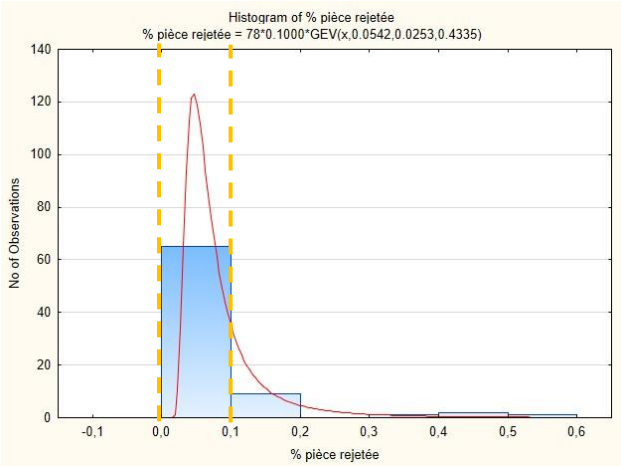
Pour chaque graphique, le seuil de dichotomisation est mis en évidence dans la Figure 3.6. En analysant, par exemple, la variable qui correspond à la quantité en pourcentage des planches délignées, nous remarquons que la quantité produite typique est comprise entre 1 et 2 %, car c'est dans cet intervalle que se situe la majorité des observations. Deux seuils sont ainsi fixés à 1 % et 2%. De même pour la variable relative à la quantité des planches déchiquetée, les seuils correspondants sont de valeurs 2 et 2,6 %, étant donné que la plupart des observations se situent dans cet intervalle. Suivant le même résonnement, en représentant l'histogramme relatif à la variable « Pourcentage des planches rejetées », la majorité des observations se situent dans l'intervalle 0 et 0,1%. Bien évidemment, aucune observation ne se trouve en dessous de 0. Dans ce cas, nous ne pourrons pas fixer les seuils en utilisant cette approche.



(a) Planches délignées



(b) Planches déchiquetées



(c) Planches rejetées

Figure 3.6 Détermination du seuil de dichotomisation en se basant sur les histogrammes

La Figure 3.6 représente également la fonction de densité de la distribution qui s'ajuste le mieux à chaque variable. Comme le montre le graphique, les deux variables correspondantes aux quantités des planches délignées et déchiquetées, suivent la loi log-normale, tandis que la troisième variable de non-qualité suit la loi d'extremum généralisée GEV « Generalized Extreme Value ».

Tout bien considéré, la transformation de chacune des variables réponses se base sur 3 différentes méthodes de détermination des seuils de dichotomisation : la moyenne, la médiane et les seuils identifiés à partir de la distribution fréquentielle. Pour chacune de ces variables continues, les valeurs de ces seuils sont présentées dans le Tableau 3.2.

Tableau 3.2 Seuils de dichotomisation pour les variables de non-qualité

Seuils / Variables	Moyenne	Médiane	Seuils basés sur la distribution
Pièce délignée [%]	1,53	1,45	1 et 2
Pièce déchiquetée [%]	2,38	2,33	2 et 2,6
Pièce rejetée [%]	0,90	0,06	

Pour chacune de ces variables, deux catégories sont obtenues en se basant sur un des 3 seuils considérés, tel que présenté dans le Tableau 3.3. Considérant par exemple un seuil défini comme étant la moyenne utilisée pour transformer la variable réponse correspondante à la quantité en pourcentage de planches délignées produite par jour, la première catégorie, représentée par 2, correspond à la production d'une quantité de planches délignée dépassant 1,53 %, constituant une production journalière importante des planches délignées, tandis que la deuxième catégorie, représentée par 1, correspond à la production d'une quantité considérée faible de planches délignée (inférieure à 1,53 %). Si les valeurs du seuil sont déterminées en utilisant l'approche de distribution fréquentielle, la première catégorie obtenue de la dichotomisation correspond aux quantités inférieures à 1 %, tandis que la deuxième catégorie obtenue de la dichotomisation correspond aux quantités de planches délignées supérieures à 2 %.

Tableau 3.3 Dichotomisation des variables réponses continues en variables binaires

Seuil	**Variable originale**	**Variable transformée (binaire)**	**Nombre d'observations**
Moyenne	Pièce délignée [%]	$Délignement = \begin{cases} 2\ si\ Pièce\ délignée\ [\%] > 1{,}53 \\ 1\ si\ Pièce\ délignée\ [\%] \leq 1{,}53 \end{cases}$	$37\ si\ class = 2$ $41\ si\ class = 1$
	Pièce déchiquetée [%]	$Déchiquetage = \begin{cases} 2\ si\ pièce\ déchiquetée\ [\%] > 2{,}38 \\ 1\ si\ pièce\ déchiquetée\ [\%] \leq 2{,}38 \end{cases}$	$37\ si\ classe = 2$ $41\ si\ classe = 1$
	Pièce rejetée [%]	$Rejet = \begin{cases} 2\ si\ pièce\ rejetée\ [\%] > 0{,}09 \\ 1\ si\ pièce\ rejetée\ [\%] \leq 0{,}09 \end{cases}$	$17\ si\ classe = 2$ $61\ si\ classe = 1$

Tableau 3.3 Dichotomisation des variables réponses continues en variables binaires (suite et fin)

Seuil	Variable originale	Variable transformée (binaire)	Nombre d'observations
Médiane	Pièce délignée [%]	$Délignement = \begin{cases} 2\ si\ Pièce\ délignée\ [\%] > 1{,}45 \\ 1\ si\ Pièce\ délignée\ [\%] \leq 1{,}45 \end{cases}$	$39\ si\ classe = 2$ $39\ si\ classe = 1$
	Pièce déchiquetée [%]	$Déchiquetage = \begin{cases} 2\ si\ Pièce\ déchiquetée\ [\%] > 2{,}33 \\ 1\ si\ Pièce\ déchiquetée\ [\%] \leq 2{,}33 \end{cases}$	$39\ si\ classe = 2$ $39\ si\ classe = 1$
	Pièce rejetée [%]	$Rejet = \begin{cases} 2\ si\ Pièce\ rejetée\ [\%] > 0{,}06 \\ 1\ si\ Pièce\ rejetée\ [\%] \leq 0{,}06 \end{cases}$	$39\ si\ classe = 2$ $39\ si\ classe = 1$
Basé sur la distribution	Pièce délignée [%]	$Délignement = \begin{cases} 2\ si\ Pièce\ délignée\ [\%] > 2 \\ 1\ si\ Pièce\ délignée\ [\%] \leq 1 \end{cases}$	$10\ si\ classe = 2$ $3\ si\ classe = 1$
	Pièce déchiquetée [%]	$Déchiquetage = \begin{cases} 2\ si\ Pièce\ déchiquetée\ [\%] > 2{,}6 \\ 1\ si\ Pièce\ déchiquetée\ [\%] \leq 2 \end{cases}$	$21\ si\ classe = 2$ $11\ si\ class = 1$

Le tableau 3.3 présente également le nombre d'observations correspondantes à chaque classe après la dichotomisation. Notons que le nombre d'observations total de l'ensemble de données qui est de 78 ne change pas dans le cas où le seuil est fixé comme étant la moyenne ou la médiane. Le compromis général entre la perte d'information et l'apport de cette classification est donc toujours intact. Cependant, si les seuils de dichotomisation sont déterminés en utilisant la méthode de distribution fréquentielle, plus que la moitié des observations sont perdues. En considérant cette méthode, on note également que le ratio des observations de la classe 1 par rapport à l'ensemble des observations est très faible, contrairement au cas où le seuil est défini comme étant la médiane. En effet, le nombre d'observations correspondant à la classe 2 est toujours égal à celui correspondant à la classe 1, lorsque la médiane est utilisée. Dans le cas où la valeur du seuil est la moyenne, un déséquilibre de classe est aussi observé pour la variable relative à la quantité des planches rejetées. Cela dit, un des avantages de la technique de classification LAD consiste à tenir compte du problème de déséquilibre des classes afin d'éviter de biaiser la modélisation.

Pour chaque variable réponse, trois modèles de classifications peuvent être obtenus, considérant ces différentes valeurs de seuils. Le choix du meilleur seuil se basera ensuite sur les performances du modèle de classification développé.

Chacune de ces variables réponses obtenues est potentiellement influencée par des indicateurs correspondants à la quantité des billes entrantes, à la quantité des différentes dimensions de planches produites, aux interventions de maintenances effectuées sur l'équipement de coupe VSS et aux quantités d'énergies vibratoires, comme mentionné dans le Chapitre 1. Ainsi, pour mieux comprendre les interactions entre ces indicateurs et les variables binaires obtenues, des modèles de classification caractérisant la non-qualité des planches (délignement, déchiquetage et rejet) avec les indicateurs correspondants, identifiés dans le Tableau 1.3 sont construits à l'aide de la technique de classification LAD.

Avant de procéder à la modélisation, il est nécessaire d'extraire tout d'abord les caractéristiques statistiques des signaux vibratoires et de prétraiter les échantillons de données.

3.1.5 Évaluation du modèle de classification

Une fois les données préparées, chacune des variables réponses obtenues de la dichotomisation est regroupée avec les indicateurs correspondants, identifiés dans le Tableau 1.3 comme étant des variables explicatives influençant la production d'un des 3 types de non-qualité. Pour chaque seuil de dichotomisation fixé, trois différents échantillons de données sont obtenus. Chaque échantillon est divisé en deux ensembles : ensemble d'apprentissage qui représente 70 % de tous les données et le reste constitue l'ensemble de test. De nombreux chercheurs utilisent un ratio de 70/30 dans la séparation des données, ayant trouvé que ce ratio présente de meilleures performances (Kurdthongmee & Suwannarat, 2019; Nguyen et al., 2021; van Blokland et al., 2021). Pour chaque variable réponse, plusieurs modèles de classifications sont développés, en utilisant différentes combinaisons des indicateurs correspondants. Chaque modèle est construit à l'aide des données d'entraînement. Ensuite, le modèle prédit la variable réponse étudiée en utilisant les indicateurs de l'ensemble test. Pour l'évaluation du modèle de classification construit en utilisant le logiciel cbmLAD, nous utilisons la précision globale, définie comme étant le rapport du nombre de prédictions correctes et du nombre total d'observations évaluées dans l'ensemble de données de test :

$$Précision\ globale = \frac{Nombre\ de\ prédictions\ correctes}{Nombre\ total\ d'observations\ d'entrée} = \frac{VP + VN}{VP + FP + FN + VN + PNC + FNC}$$

Où, VP représente le nombre de vrais positifs, VN est le nombre de vrais négatifs, FP représente le nombre de faux positifs et FN est le nombre de faux négatifs. Ces derniers sont généralement présentés dans la matrice de corrélation, comme le montre le Tableau 3.4. Une des caractéristiques de la technique LAD est le fait que toutes les observations sont couvertes par des patterns tant que les observations n'apportent aucune contradiction. La contradiction signifie avoir une classe avec une observation décrite avec les mêmes valeurs de variables qu'une observation d'une autre classe, introduisant ainsi le PNC qui est le nombre de positive non classé et le FNC qui est le nombre de négative non classé.

Tableau 3.4 Matrice de confusion

Prédites / Réelles	1	0	Non classé
1	VP	FN	PNC
0	FP	VN	FNC

Les patterns générés pour chaque classe du modèle de classification LAD ainsi que leurs poids corresponds seront également examinés.

3.2 Modélisation du délignement par classification

Pour trouver un modèle précis permettant de modéliser le délignement en se basant sur le logiciel cbmLAD, nous avons utilisé différentes combinaisons d'indicateurs. Cela nous permet de déterminer le modèle le plus significatif et le plus précis pour une utilisation future.

Dans un premier temps, nous avons adapté un modèle de classification aux données observées sur le pourcentage des planches délignées en fonction de tous les indicateurs influents identifiés dans le chapitre 1, comme présenté dans la Figure 3.7.

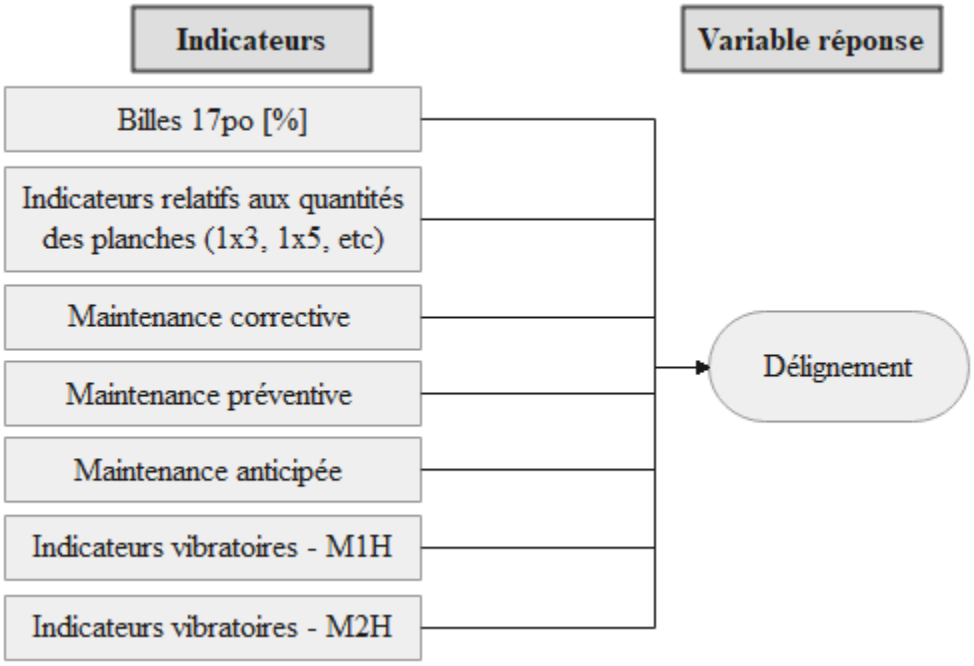


Figure 3.7 Modélisation du délignement des planches en fonction de tous les indicateurs

Selon la Figure 3.7, les caractéristiques sélectionnées précédemment en se basant sur la méthode de régression pas à pas qui représentent les signaux vibratoires issus des deux capteurs M1H et M2H sont considérées parmi ces indicateurs.

Le Tableau 3.5 résume tous les résultats des modèles de classification développés pour les différentes combinaisons d'indicateurs et les différentes valeurs des seuils fixées. Notons que l'indicateur relatif à la matière première seul n'est pas capable de capturer le phénomène de non-qualité. Cet indicateur est insuffisant pour la modélisation du délignement. Il est de même pour les indicateurs de maintenance. Aussi, le Tableau 3.5 indique que la précision du modèle LAD peut atteindre une valeur de 80%, dans le cas où le seuil est basé sur la distribution. Cependant, en utilisant ce seuil, plus que la moitié d'observations ont été perdues lors de la dichotomisation.

En examinant toutes les valeurs de précision pour les 3 différents seuils, les meilleurs résultats du modèle de classification LAD sont généralement obtenus en utilisant la variable réponse dichotomisée en se basant sur la médiane. En effet, dans ce cas, les valeurs de la précision du modèle sont presque toujours supérieures à celles des modèles LAD lorsque la dichotomisation est basée sur, soit la moyenne, soit la distribution. Ceci peut être dû au fait qu'en adoptant la médiane, les données sont parfaitement équilibrées entre les deux classes.

Tableau 3.5 Précisions des modèles de classification utilisés pour l'analyse du délignement

Indicateurs	Moyenne	Médiane	Seuil basé sur la distribution
Quantités des produits finis (PF) (%)	69,23	65,38	60
Indicateurs vibratoires – M1H (%)	61,54	61,53	60
Indicateurs vibratoires – M2H (%)	53,85	61,54	60
Indicateurs vibratoires - M1H+M2H (%)	42,31	53,85	80
Bille + PF (%)	57,69	57,69	40
PF+ Maintenance (%)	53,85	65,38	40
PF+M1H (%)	61,54	65,38	60
PF+M2H (%)	65,38	**72**	60
PF+M1H+M2H (%)	69,23	69,23	60
Tous les indicateurs (%)	53,85	50	50

Si le seuil de dichotomisation est défini comme étant la médiane, la valeur de la précision la plus élevée atteinte est de 72%, obtenue lorsque seulement les indicateurs relatifs aux quantités des planches et les indicateurs vibratoires du capteur M2H sont utilisés comme variables d'entrée. La matrice de confusion correspondante est représentée dans le tableau suivant.

Tableau 3.6 Matrice de confusion correspondante au délignement

Prédites / Réelles	1	0	Non classé
1	9	2	2
0	1	9	2

Les patterns générés par cbmLAD qui forment les règles de décision du modèle LAD en utilisant la médiane comme seuil sont présentés dans les Tableaux 3.6 et 3.7. Le Tableau 3.6 présente les patterns de la classe 2 correspondant à une production journalière importante des planches délignées, où la quantité de ces planches dépasse 1,53%, comme présentés dans le Tableau 3.3. Et, le Tableau 3.7 présente les patterns de la classe 1, qui correspondent à la production d'une quantité faible de planches délignée. Ces patterns peuvent être interprétés en termes d'indicateurs de l'ensemble de données. Par exemple, en examinant le Tableau 3.6, le pattern 2 de la classe 2 peut être interprété comme suit: la présence d'un niveau d'énergie vibratoire supérieur à 89,37 représenté par l'indicateur « RMS spectral – M2H » et une production d'une quantité de planches de dimension 5x4po supérieur à 7,22 indiquent que la quantité des planches délignées dépasse 1,53%. Cela peut être utile aux techniciens, car elle aide à comprendre les raisons de l'apparition de ce type de non-qualité. Suite à cela, un diagnostic des équipements de coupes pourrait être ainsi réalisé ou un contrôle de production d'un certain type de planches pourrait être effectué.

Tableau 3.7 Les patterns de la classe 2 correspondant aux problèmes du délignement

		Pattern 1	**Pattern 2**	**Pattern 3**
	Poids du pattern	0,396	0,292	0,312
Indicateurs	Planches 1x3	$< 4{,}47$		
	Planches 2x3		$< 2{,}63$	
	Planches 2x4	$< 28{,}31$	$< 34{,}41$	
	Planches 2x6	$< 38{,}52$	$< 37{,}76$	
	Planches 2x8			$> 4{,}05$
	Planches 2x10			$> 0{,}91$
	Planches 3x3	$< 2{,}41$		
	Planches 5x4		$> 7{,}22$	
	RMS spectral-M2H		$> 89{,}37$	
	RMS - M2H		$< 1{,}43$	$< 1{,}43$

Tableau 3.8 Les patterns de la classe 1 correspondant aux problèmes du délignement

		Pattern 1	**Pattern 2**	**Pattern 3**
	Poids du pattern	0,375	0.339	0.286
Indicateurs	Planches 1x3		$> 4{,}32$	
	Planches 1x4	$> 4{,}92$	$> 4{,}92$	
	Planches 1x6	$< 1{,}545$	$< 1{,}55$	$< 1{,}55$
	Planches 2x4	$> 28{,}31$		
	Planches 2x10			$< 0{,}83$
	Planches 5x4			$< 11{,}08$
	Moyenne - M2H	$< 0{,}03$	$< 0{,}03$	
	RMS spectral-M2H	$< 319{,}53$		
	RMS - M2H	$> 0{,}08$	$> 0{,}77$	$> 0{,}77$

3.3 Modélisation du déchiquetage par classification

Tel qu'effectué dans la Section 3.9, différents modèles de classification sont développés et évalués afin de sélectionner par la suite le meilleur modèle qui permet d'expliquer le déchiquetage. La figure suivante présente les indicateurs qui peuvent potentiellement simuler le déchiquetage des planches produites. Dans ce cas, les indicateurs correspondent aux quantités des planches de différentes dimensions, aux nombres d'interventions de maintenance, et aux quantités d'énergies vibratoires du capteur M1H et M2H.

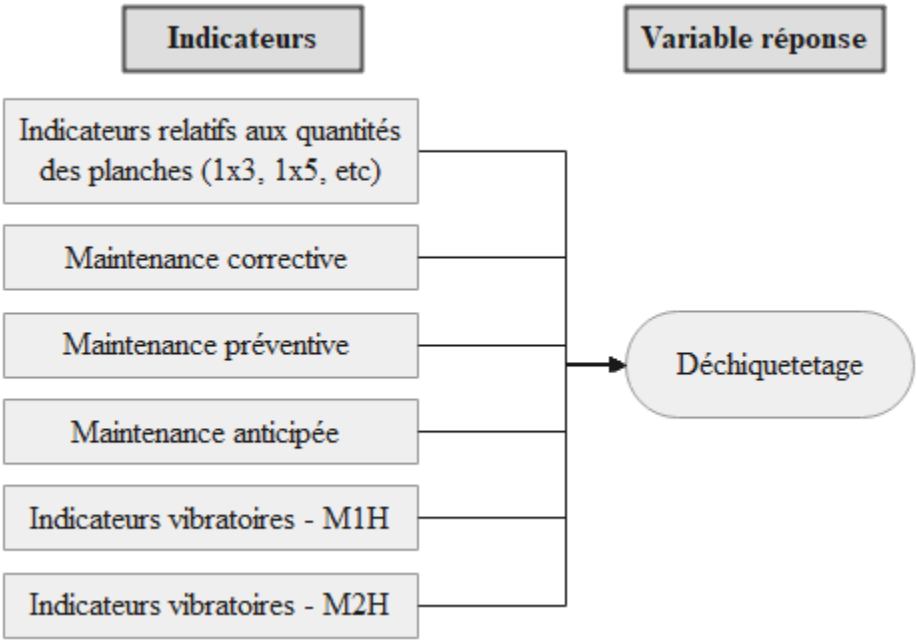


Figure 3.8 Modélisation du déchiquetage des planches en fonction de tous les indicateurs

Le Tableau 3.8 présente les résultats de la modélisation du rejet en utilisant LAD pour les différentes combinaisons d'indicateurs et les différents seuils de dichotomisation. Notons que les variables correspondantes aux nombres d'interventions de maintenance effectuées sur l'équipement de coupe ne sont pas suffisantes pour expliquer la variable réponse. Ces indicateurs sont donc considérés non significatifs dans ce modèle.

Le meilleur modèle LAD est obtenu en utilisant la médiane comme étant le seuil de dichotomisation, ayant des valeurs de précision toujours supérieures à celles des autres modèles ou le seuil est basé sur soit la moyenne soit la distribution. Le tableau 3.8 révèle aussi une précision comprise entre 46,15 et 65,38 %, lorsque le seuil est défini comme étant la moyenne, tandis qu'en se basant sur la moyenne ou sur la distribution, les valeurs de la précision des modèles varient respectivement entre 34,62 et 60 % ou entre 28,57 et 57,14 %.

Tableau 3.9 Comparaison de tous les modèles de classification utilisés pour l'analyse du déchiquetage

Indicateurs	Moyenne	Médiane	Seuil basé sur la distribution
PF (%)	42,31	53,85	28,57
Indicateurs vibratoires – M1H (%)	53,85	57,69	42,86
Indicateurs vibratoires – M2H (%)	60	60	42,86
Indicateurs vibratoires - M1H+M2H (%)	53,85	61,54	42,86
PF + Maintenance (%)	34,62	46,15	42,86
PF + M1H (%)	42,31	61,54	42,86
PF + M2H (%)	60	**65,38**	57,14
PF + M1H + M2H (%)	42,31	53,85	42,86
Tous les indicateurs (%)	50	57,69	42,86

Si la valeur du seuil de dichotomisation est définie comme étant la médiane, le modèle de classification ayant la meilleure performance a été obtenu en utilisant les caractéristiques relatives aux signaux vibratoires issus du capteur M2H et les indicateurs relatifs aux quantités des planches, avec une précision de 65,38 %. La matrice de confusion correspondante à ce modèle est représentée dans le tableau 3.10.

Tableau 3.10 Matrice de confusion correspondante au déchiquetage

Prédites / Réelles	1	0	Non classé
1	6	0	3
0	4	11	2

Les Tableaux 3.11 et 3.12 présentent les patterns générés du modèle correspondant, ainsi que leur poids. Les patterns de la classe 2 correspondant à une production journalière importante des planches déchiquetées sont présentés dans le Tableau 3.11, tandis que les patterns de la classe 1 sont présentés dans le Tableau 3.12.

Tableau 3.11 Les patterns de la classe 2 correspondant aux problèmes du déchiquetage

		Pattern 1	**Pattern 2**	**Pattern 3**	**Pattern 4**	**Pattern 5**
	Poids du pattern	0,071	0,257	0,214	0,214	0,243
Indicateurs	Planches 1x6				$> 1{,}22$	
	Planches 2x3			$> 1{,}268$		
	Planches 2x4	$> 25{,}76$	$> 25{,}76$	$> 25{,}76$	$> 25{,}76$	$> 27{,}17$
	Planches 2x6		$< 39{,}09$			$< 39{,}09$
	Planches 2x10		$< 1{,}143$	$< 1{,}14$	$> 0{,}37$	$> 0{,}37$
	Planches 3x3	$> 2{,}87$				
	Planches 5x4	$< 14{,}37$	$< 14{,}37$	$< 14{,}37$	$< 14{,}37$	$< 14{,}37$
	Variance spectrale – M2H			$> 3{,}99$		
	SS – M2H	$< 20{,}93$	$< 16{,}49$	$< 20{,}93$	$< 16{,}49$	$< 26{,}76$

Tableau 3.12 Les patterns de la classe 1 correspondant aux problèmes du déchiquetage

		Pattern 1	**Pattern 2**	**Pattern 3**	**Pattern 4**	**Pattern 5**
	Poids du pattern	0,241	0,259	0,189	0,121	0,189
Indicateurs	Planches 1x3					$< 4{,}76$
	Planches 1x6	$> 1{,}26$		$> 1{,}34$		
	Planches 2x3			$< 1{,}75$		
	Planches 2x4		$< 27{,}62$	$< 32{,}86$		
	Planches 2x8				$< 11{,}90$	$< 12{,}49$
	Planches 2x10			$< 1{,}68$		$> 1{,}68$
	Planches 3x3		$< 2{,}87$	$< 2{,}87$		$> 2{,}15$
	Planches 5x4	$> 8{,}86$	$> 4{,}14$			$> 4{,}14$
	Moyenne - M2H		$< 6{,}63$	$> 2{,}11$		
	Variance spectrale – M2H	$< 9{,}14$		$< 9{,}14$	$< 3{,}87$	$< 9{,}14$
	SS – M2H	$> 12{,}75$	$> 7{,}64$	$> 11{,}46$		$> 7{,}64$

3.4 Modélisation du rejet par classification

Cette section présente les modèles de classification LAD du troisième type de non-qualité des planches, le rejet. La variable réponse correspondante dépend de la quantité des billes et des planches, ainsi que des quantités vibratoires délivrées par les deux accéléromètres M1H et M2H, comme le montre la figure ci-dessous.

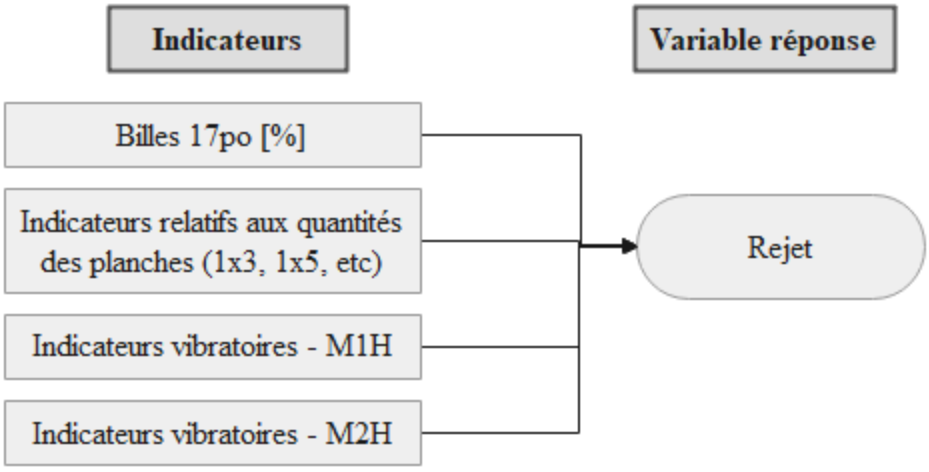


Figure 3.9 Modélisation du rejet des planches en fonction de tous les indicateurs

Les résultats des modèles de classification pour les deux valeurs du seuil fixées et pour les différentes combinaisons d'indicateurs utilisées sont résumés dans le tableau suivant. Ce dernier montre qu'en utilisant la médiane les performances des modèles de classification sont meilleures que ceux obtenus en utilisant la moyenne en termes de précision. Lorsque la valeur du seuil est définie comme étant la moyenne, les valeurs de précision varient entre 50 et 69 %, tandis qu'en utilisant la médiane, les valeurs de précision sont entre 69 % et 80 %.

Tableau 3.13 Comparaison de tous les modèles de classification utilisés pour l'analyse du rejet

Indicateurs	Moyenne	Médiane
Quantités des planches (PF) (%)	69,23	76,92
Indicateurs vibratoires – M1H (%)	52	80
Indicateurs vibratoires – M2H (%)	53,85	73,08
Indicateurs vibratoires - M1H+M2H (%)	65,38	73,08
Bille + PF (%)	50	73,08
PF + M1H (%)	57,69	73,08
PF + M2H (%)	50	69.23
PF + M1H + M2H (%)	53,85	73,0
Tous les indicateurs (%)	50	**80,77**

Dans le cas où le seuil est la médiane, la valeur de précision la plus élevée est de 80,77%, obtenue en utilisant tous les indicateurs présentés dans la Figure 3.3. La matrice de confusion correspondante à ce modèle est présentée dans le tableau 3.14.

Tableau 3.14 Matrice de confusion correspondante au rejet

Prédites / Réelles	1	0	Non classé
1	1	0	2
0	2	20	1

Les patterns générés par cbmLAD qui forment les règles de décision du modèle correspondant sont présentés dans les Tableaux 3.15 et 3.16. Le Tableau 3.15 présente les patterns de la classe 2, tandis que Tableau 3.16 présente les patterns de la classe 1 qui correspondent à une quantité de planches rejetées inférieure à 0,06%.

Tableau 3.15 Les patterns de la classe 2 correspondant aux problèmes du rejet

		Pattern 1	**Pattern 2**
	Poids du pattern	0,524	0,476
Indicateurs	Planches 1x3	> 4	> 4
	Planches 1x6	$> 1{,}28$	$> 1{,}19$
	Planches 2x3		$< 1{,}9$
	Planches 2x10		$> 0{,}48$
	Planches 3x3	$< 2{,}94$	$< 2{,}94$
	Variance – M1H	$< 0{,}15$	$< 0{,}15$
	FC – M1H	$> 0{,}1$	$> 0{,}1$
	Ku – M2H	$< 2{,}95$	$< 2{,}95$
	KS – M2H	$> 157{,}2$	

Tableau 3.16 Les patterns de la classe 1 correspondant aux problèmes du rejet

		Pattern 1	**Pattern 2**	**Pattern 3**	**Pattern 4**
	Poids du pattern	0,245	0,245	0,226	0,284
Indicateurs	Bille 17po [%]	$< 91{,}725$	$< 93{,}255$	$< 93{,}255$	$< 93{,}255$
	Planches 1x3				$< 5{,}005$
	Planches 1x4	$> 4{,}95$		$> 5{,}215$	$> 4{,}95$
	Planches 1x6		$< 1{,}7$	$< 1{,}385$	
	Planches 2x6	$> 32{,}425$			
	Planches 2x10		$> 0{,}54$		
	Planches 3x3		$> 2{,}02$		
	Planches 5x4				$< 11{,}90$
	Variance – M1H		$< 6{,}63$		
	Ku – M2H	$< 19360{,}3$	$< 19360{,}3$	$< 19360{,}3$	$< 3{,}87$

CHAPITRE 4 ANALYSE DE RÉGRESSION

Ayant répondu au premier objectif en procédant à la modélisation par classification dans le Chapitre 3, le Chapitre 4 a pour but de présenter et d'analyser les résultats des modèles de régression développés pour expliquer les différents types de non-qualité des planches produites lors du débitage de billes. L'influence de chaque variable d'entrée sur les variations des variables réponses est également étudiée en se basant sur l'analyse de sensibilité.

Avant d'entamer l'analyse de régression, nous présenterons la méthodologie suivie pour modéliser les différents types de non-qualités. Nous expliquerons ainsi les hypothèses à respecter pour construire des modèles de régression linéaires et les métriques à examiner pour valider la robustesse de ces modèles.

4.1 Méthodologie

Le but de la modélisation par régression est d'expliquer les causes de la production des produits finis de mauvaise qualité et d'identifier les variables affectant chaque caractéristique de non-qualité dans les planches. Pour ce faire, l'ensemble des données constituant les indicateurs présentés dans le Chapitre 1 et les variables réponses continues correspondantes seront exploitées. Chacune de ces variables présente la quantité en pourcentage d'un type de non-qualité.

Le processus de modélisation par régression de ces variables est présenté dans la Figure 4.1. Ce processus comprend principalement quatre étapes : la préparation des données comme effectuer dans la section précédente, la vérification des hypothèses de régression si le modèle construit est la régression linéaire, le développement du modèle de régression et l'évaluation de ce modèle.

Afin d'analyser l'effet des vibrations sur les variables réponses du modèle de régression, les caractéristiques statistiques extraites à partir des signaux vibratoires et sélectionnées en se basant sur la méthode de régression pas à pas dans la Section 3.1 sont utilisées comme indicateurs.

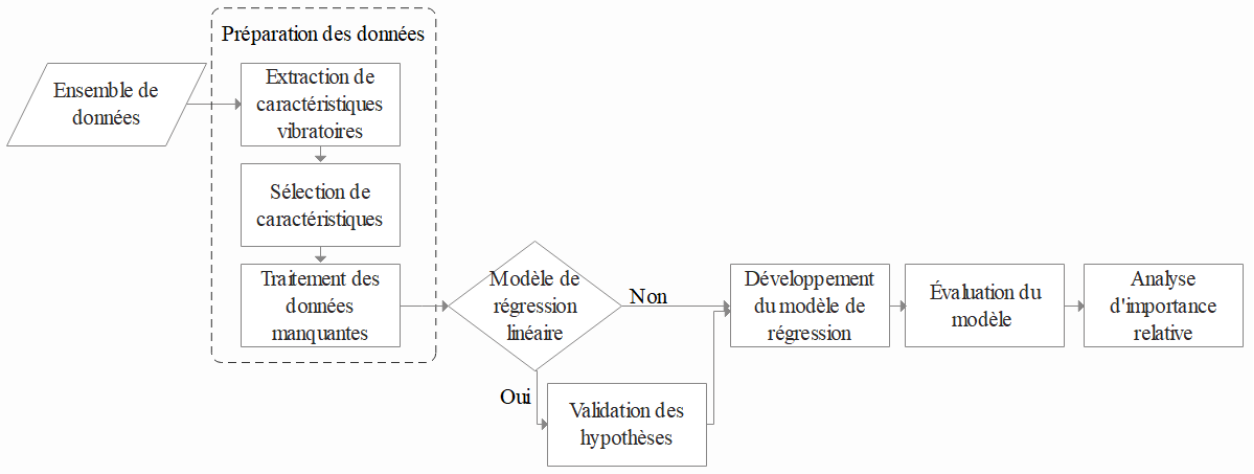


Figure 4.1 Étapes d'analyse de données par régression

Lors de l'application de la régression linéaire, il est important d'évaluer dans quelle mesure le modèle correspond aux données et que les données répondent aux hypothèses du modèle. Pour déterminer si un modèle de régression linéaire est approprié, les quatre hypothèses suivantes doivent être vérifiées, selon (Montgomery, 2017).

- L'indépendance dans les variables réponses,
- La normalité où les résidus sont distribués normalement,
- La linéarité où la relation entre la variable réponse et les indicateurs est une droite,
- Et, l'homoscédasticité-variabilité où la variable réponse a une variance constante.

Comme expliqué dans (Montgomery, 2017), si les hypothèses sur les erreurs sont satisfaites, les hypothèses sur la variable de réponse sont validées. La première hypothèse, l'indépendance dans la variable réponse, dépend de la manière dont les données ont été recueillies. Montgomery trace le graphique des résidus dans l'ordre temporel de la collecte de données pour détecter toute dépendance dans la variable réponse. Idéalement, les résidus devraient être distribués de manière aléatoire autour de zéro. Pour étudier la deuxième hypothèse, (Montgomery, 2017) suggère de construire un graphique des résidus sur échelle de probabilité gaussienne. La normalité des résidus est validée si le graphique montre des points correspondant aux résidus qui sont alignés. Pour vérifier la linéarité et l'homoscédasticité, (Montgomery, 2017) recommande de tracer un graphique des résidus standardisés en fonction des valeurs prédites par le modèle construit. La variabilité des

résidus ne doit en aucun cas dépendre des valeurs prédites. Si ces deux hypothèses sont satisfaites, les résidus devraient être répartis de manière aléatoire autour de zéro sur l'ensemble du graphique. Le tableau 4.1 présente les hypothèses ainsi que les graphiques correspondants utilisés pour leur validation.

Tableau 4.1 Hypothèses de régression et graphiques correspondants

Hypothèses	**Graphiques**
Indépendance	Graphique des résidus en fonction de l'ordre d'exécution
Normalité	Graphique des résidus sur échelle de probabilité gaussienne
Linéarité	Graphique des résidus standardisés en fonction des valeurs prédites
Homoscédasticité	

Le modèle de régression est construit à l'aide de l'ensemble de données d'entraînement qui représente, dans ce cas, 70 % de tout l'ensemble. Le modèle fait des prédictions de la variable réponse en utilisant les entrées de l'ensemble de données de test.

Étant donné que 3 variables réponses sont considérées dans cette étude, nous avons affaire à un problème de régression à multi-sorties. Comme présenté dans la Section 2.4.2.1, deux approches seront utilisées afin de résoudre ce type de problème. Dans notre cas, chacune des variables réponses a ces propres variables expliquées qui lui correspondent. Celles-ci sont représentées dans les Figures 3.8, 3.9 et 3.10. Ainsi, différents ensembles d'indicateurs sont considérés pour chaque modèle comme le montre la Figure 4.2. Nous considérons, dans notre cas d'études, que le Y_1 représente la quantité des planches délignées en pourcentage, Y_2 la quantité des planches déchiquetées en pourcentage et Y_3 représente la quantité des planches rejetées en pourcentage.

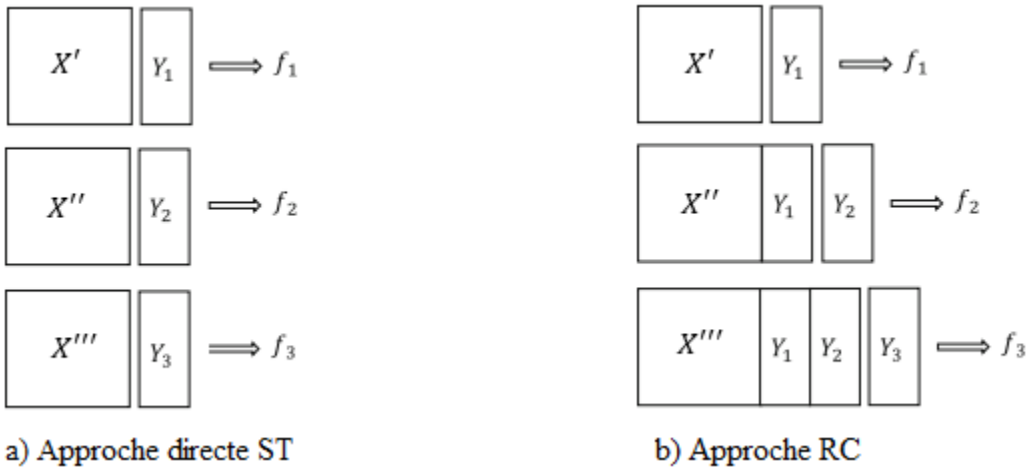


Figure 4.2 Approches ST et RC en phase d'entraînement dans ce cas d'études

Pour modéliser les variables de non-qualité, nous allons tout d'abord nous baser sur l'approche directe à sortie unique ST « Single Target », où les modèles de régression suivants sont appliqués en utilisant Python:

- La régression linéaire. Dans ce cas, les quatre hypothèses sont à vérifier. Et, pour déterminer le modèle de régression linéaire le plus ajusté aux données de la variable réponse étudiée, différentes combinaisons d'indicateurs sont utilisées.
- Modèle d'ensemble basé sur les 3 modèles régression suivants : les k plus proches voisins k-NN, le boosting de gradient GB et la forêt d'arbres décisionnels RF. Ce modèle d'ensemble calcule la moyenne des prédictions résultantes de ces 3 modèles pour former une prédiction finale, comme expliquer dans la Section 2.4.2.2.

Compte tenu des performances de ces modèles implémentés, un modèle d'ensemble basé sur l'approche de régression à chaîne RC est proposé.

Pour mesurer la qualité des modèles développée, le coefficient de détermination R^2 est examiné. De nombreux auteurs discutent de l'importance et de la façon de calculer cette valeur. (Hogg & Ledolter, 1992) définissent ce coefficient comme étant la proportion de variation de la variable réponse qui est expliquée par le modèle de régression. La valeur de R^2 varie entre 0 et 1. Si elle est proche de 0, le modèle de régression ne correspond pas bien aux données et qu'il n'y a pas de

relation entre les indicateurs et la variable étudiée. Et, si cette valeur est proche de 1, le modèle développé s'ajuste bien aux données. Ce coefficient est défini comme suit :

$$R^2 = \frac{\sum(Y - \bar{Y})^2 - \sum\left(Y - \hat{Y}\right)^2}{\sum(Y - \bar{Y})^2}$$

$\bar{Y}$ représente la valeur moyenne de la variable réponse et $\hat{Y}$ correspond à la variable prédite par le modèle de régression.

Pour équilibrer l'effet du nombre des indicateurs sur le coefficient de détermination, le $R^2_{\text{ajusté}}$ est aussi examiné, défini comme suit :

$$R^2_{ajusté} = 1 - \frac{(1 - R^2) - (n - 1)}{n - p - 1}$$

n est le nombre d'observations de l'ensemble test et p est le nombre des indicateurs utilisés.

Enfin, une analyse de sensibilité a été réalisée pour évaluer l'importance des variables dans les modèles ajustés.

4.2 Approche directe basée sur la régression à sortie unique

4.2.1 Analyse basée sur la régression linéaire

4.2.1.1 Analyse du délignement des planches

Cette section présente l'analyse de régression linéaire qui permet de construire le meilleur modèle d'ajustement de la quantité des planches délignées en pourcentage en fonction des mêmes indicateurs utilisés dans la Section 3.2.

Pour déterminer le modèle de régression le plus ajusté aux données de cette variable réponse, différentes combinaisons d'indicateurs sont utilisées. Ainsi, les étapes suivantes sont principalement suivies :

- Ajuster les données à un modèle de régression multiple qui considère tous les indicateurs.
- Effectuer une analyse de régression linéaire des planches délignées avec une catégorie d'indicateurs à la fois, soit la quantité des billes en pourcentage, soit les indicateurs relatifs

aux quantités des planches de bois produites, soit les variables relatives aux interventions de maintenance ou les indicateurs représentant les signaux vibratoires.

4.2.1.1.1 Modélisation du délignement des planches par régression multiple

Dans un premier temps, nous tenterons d'adapter une régression linéaire multiple aux données observées sur le pourcentage des planches délignées en fonction de tous les indicateurs identifiés comme étant influents. Pour pouvoir utiliser ce modèle, les quatre hypothèses expliquées précédemment doivent être validées :

i. Hypothèse d'indépendance dans la variable réponse qui correspond dans ce cas au pourcentage des planches délignées. La Figure 4.3 représente le graphique des résidus de la quantité des planches délignées en pourcentage ajustée par rapport à l'ordre d'exécution de la collecte de données. Notons que la ligne traversant les points est la ligne la mieux ajustée pour le résidu. Ce graphique montre des résidus distribués de manière aléatoire, validant cette hypothèse.

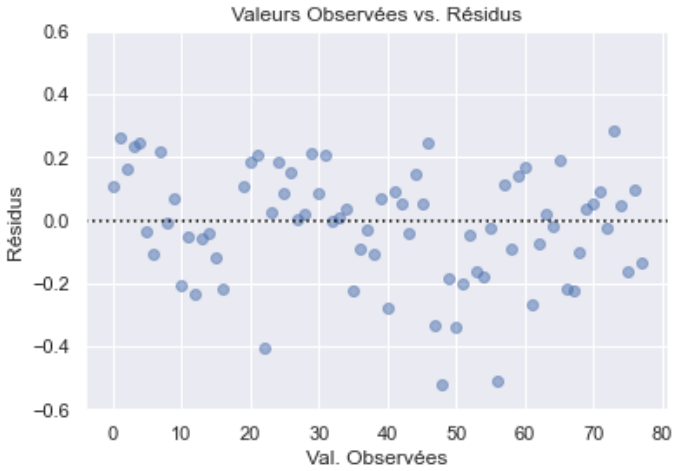


Figure 4.3 Graphique des résidus en fonction de l'ordre d'exécution issu de la modélisation du délignement en fonction de tous les indicateurs

ii. L'hypothèse de normalité des résidus est également valide, étant donné que les points du graphique des résidus sur échelle de probabilité gaussienne forment une ligne (Figure 4.4). Ce graphique illustre aussi une ligne rouge de référence qui nous permet de détecter

facilement les valeurs aberrantes. Aucun écart important par rapport à la ligne droite rouge n'est observé à part deux points. Les erreurs aléatoires sont donc normalement distribuées.

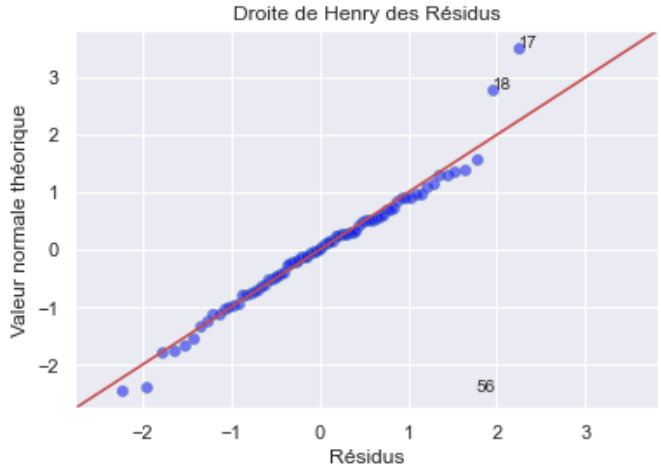


Figure 4.4 Graphique des résidus sur échelle de probabilité gaussienne issu de la modélisation du délignement en fonction de tous les indicateurs

iii. Les hypothèses de linéarité et d'homoscédasticité sont valides. En effet, le graphique des résidus versus les prédictions (Figure 4.5) montre une variance du résidu constante et des valeurs du résidu qui sont dispersées aléatoirement autour de zéro.

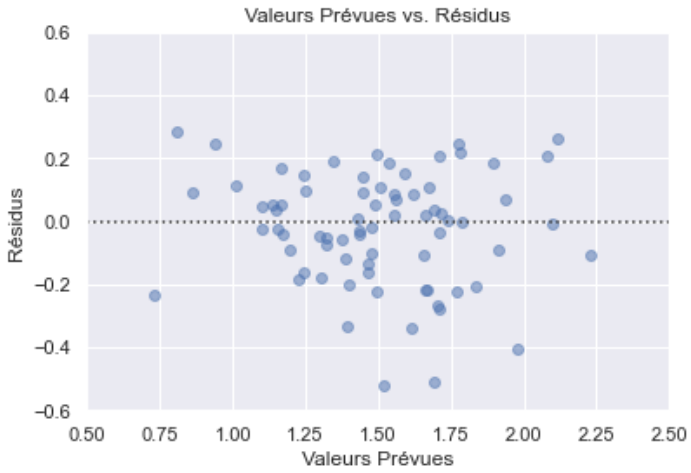


Figure 4.5 Graphique des résidus versus les prédictions issues de la modélisation du délignement en fonction de tous les indicateurs

Ayant vérifié les quatre hypothèses de régression, le modèle peut être construit en utilisant les données d'entraînement. Pour évaluer la performance de ce modèle, nous allons nous baser sur les coefficients de détermination R^2 et $R^2_{ajusté}$. Ces derniers sont présentés dans le Tableau 4.2.

Tableau 4.2 Résultats de la régression linéaire en utilisant tous les indicateurs

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	79,13	57,42	50,65	21,33

D'après le Tableau 4.2, nous remarquons que le coefficient de détermination relatif aux données d'entraînement est acceptable (proches de 1), tandis que le coefficient de détermination ajusté est faible. Les valeurs de R^2 relatifs aux données test sont également faibles. Par conséquent, nous pouvons affirmer que la variable relative à la quantité des planches délignées ne peut pas être représentée par tous les indicateurs. Nous allons donc tenter d'ajuster le modèle avec un sous-ensemble de ces indicateurs.

4.2.1.1.2 Délignement des planches en fonction de la quantité des billes

Dans cette section, nous analysons la quantité d'une certaine dimension des matières premières comme le seul effet sur le pourcentage de planches délignées. Ce processus est la régression linéaire simple, car nous avons affaire à un seul indicateur.

Le Tableau 4.3 montre les coefficients de détermination résultants de l'évaluation de ce modèle de régression.

Tableau 4.3 Résultats de la régression linéaire en utilisant l'indicateur relatif à la quantité de billes

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	13,54	11,82	-13,87	-18,61

Les valeurs de $R^2_{ajusté}$ de l'ensemble d'entraînement et de l'ensemble de test sont respectivement à 11,82% et -18,61%. Ayant des valeurs très faibles, ce modèle ne correspond pas aux données, car plus de 80% de la variabilité n'est pas expliquée par ce modèle.

4.2.1.1.3 Délignement des planches en fonction de la quantité des planches

Pour modéliser la quantité des planches délignée en utilisant la régression, nous allons utiliser cette fois seulement les variables relatives aux quantités des différentes dimensions des planches, qui sont au nombre de dix. Le Tableau 4.4 présente les résultats de cette modélisation.

Tableau 4.4 Résultats de la régression linéaire en utilisant les indicateurs relatifs à la quantité de planches

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	74,69	68,52	70,68	51,14

Les résultats montrent une bonne concordance entre les valeurs expérimentales et prédites puisque le coefficient de détermination calculé est de 74,69 % pour les données d'entraînement et de 70,68% pour les données test. La plupart de la variabilité est expliquée par ce modèle.

Tel qu'expliqué dans la Section 4.1, ce modèle de régression doit suivre les quatre hypothèses avant de l'utiliser pour modéliser les planches délignées.

i. Hypothèse d'indépendance dans la variable réponse est valide, en représentant le graphique des résidus de la quantité des planches délignées en pourcentage par rapport à l'ordre d'exécution de la collecte des données (Figure 4.6). En effet, ce graphique ne montre pas de relation systématique entre le résidu et l'ordre d'observation.

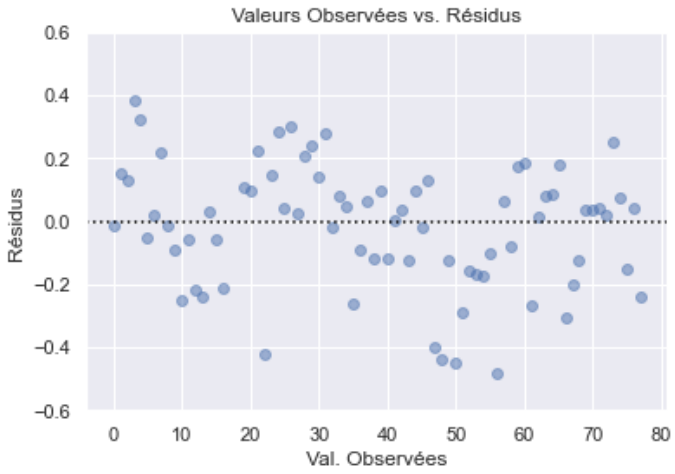


Figure 4.6 Graphique des résidus en fonction de l'ordre d'exécution issu de la modélisation du délignement en fonction de la quantité des planches

ii. Le graphique des résidus sur échelle de probabilité gaussienne tracé dans la Figure 4.7 montre un nuage de points qui forment une ligne droite, validant l'hypothèse de normalité des résidus.

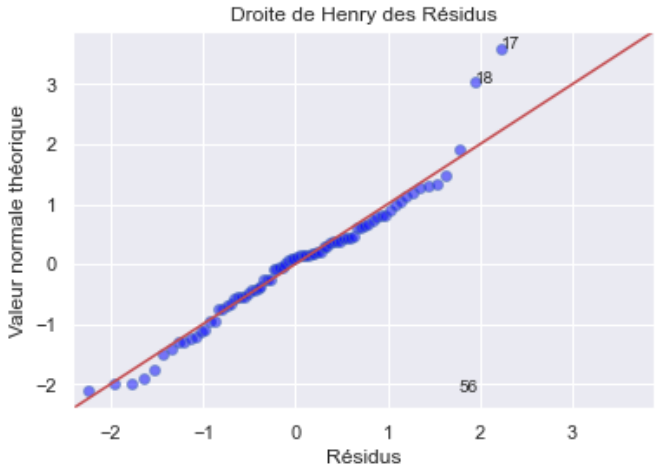


Figure 4.7 Graphique des résidus en fonction de l'ordre d'exécution issu de la modélisation du délignement en fonction de la quantité des planches

iii. Le graphique des valeurs prédites par rapport aux résidus standardisés indique une variance qui est constante et prouve également la validité de l'homoscédasticité. En effet, le nuage de points ne forme aucune tendance ou structure particulière et les points fluctuent aléatoirement autour de zéro.

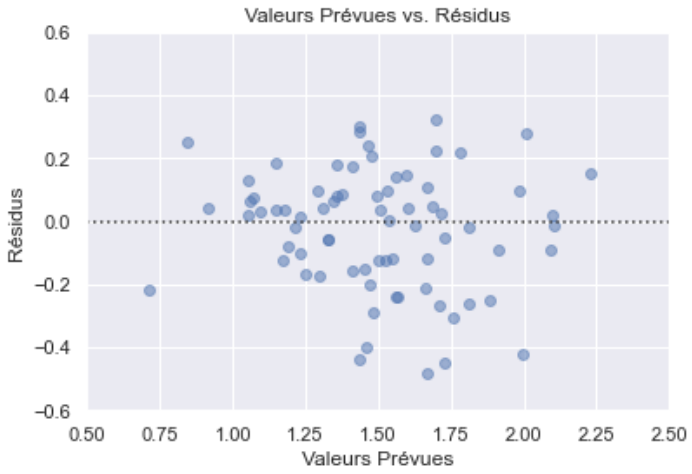


Figure 4.8 Graphique des résidus versus les prédictions issues de la modélisation du délignement en fonction de la quantité des planches

4.2.1.1.4 Délignement des planches en fonction de la maintenance

Cette section présente une analyse des interventions de maintenance corrective, anticipée et préventive comme étant les seuls influents sur le pourcentage de planches délignées. Le Tableau 4.5 présente les coefficients de détermination du modèle de régression correspondant.

Tableau 4.5 Résultats de la régression linéaire en utilisant les indicateurs relatifs à la maintenance

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	4	-2	10,28	-1,96

Ce modèle proposé ne peut pas être utilisé, car les valeurs des coefficients de détermination sont extrêmement faibles. Les indicateurs relatifs aux interventions de maintenance seuls sont insuffisants pour expliquer la quantité des planches délignées.

4.2.1.1.5 Délignement des planches en fonction de la vibration

Une modélisation par régression du délignement en fonction des caractéristiques statistiques extraites des signaux vibratoires délivrés par les capteurs M1H et M2H est effectuée. Le Tableau 4.6 présente les résultats de cette modélisation, en utilisant soit les indicateurs relatifs au capteur M1H soit les indicateurs relatifs au capteur M2H, ou bien tous les indicateurs représentant tous les signaux capturés.

Tableau 4.6 Résultats de la régression linéaire en utilisant les indicateurs relatifs à la vibration

	Entraînement		**Test**	
Indicateurs	$\mathbf{R^2}$	$\mathbf{R^2_{ajusté}}$	$\mathbf{R^2}$	$\mathbf{R^2_{ajusté}}$
Indicateurs de vibration M1H	19,81	7,05	22,1	-8,9
Indicateurs de vibration M2H	8,53	-1,41	7.48	-15,65
Indicateurs de vibration M1H et M2H	26,38	3,72	11,14	-16,21

Les résultats montrent des valeurs R^2 faibles qui varient entre 8,53 et 26,38 %. Ainsi, l'adéquation entre chacun de ces 3 modèles et les données observées est mauvaise. Les caractéristiques vibratoires seules ne sont pas suffisantes pour expliquer la variable réponse.

4.2.1.1.6 Comparaison des modèles de régression pour modéliser le délignement

Pour trouver un modèle précis permettant de modéliser la quantité des planches délignées en pourcentage, nous avons utilisé différentes combinaisons d'indicateurs plusieurs méthodes. Le Tableau 4.7 explore le modèle le plus précis avec les meilleures performances.

Tableau 4.7 Comparaison de tous les modèles de régression pour l'analyse du délignement

		Entraînement		Test	
Indicateurs	**Validité**	**R^2**	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Bille 17po [%]	Non valide : indicateurs non significatifs	13,54	11,82	-13,87	-18,61
Maintenance	Non valide : indicateurs non significatifs	4	-2	10,28	-1,96
Dimensions du produit fini (PF)	Valide : hypothèses satisfaites	74,69	68,52	70,68	51,14
PF + M1H	Non valide : R^2 test faible	77,65	66,47	54,82	32,23
PF + M2H	Valide : hypothèses satisfaites	**75,45**	**65,22**	**72,55**	**61,11**
PF + M1H + M2H	Non valide : R^2 test faible	78,41	62,03	59,14	28,14
MP+PF+M1H	Non valide : R^2 test faible	77,69	65,53	53,02	27,4
MP+PF+M2H	Valide: hypothèses satisfaites	75,45	64,23	72,74	60,27
Tous les indicateurs	Non valide : R^2 test faible	79,13	57,42	50,65	21,33

En examinant les valeurs de R^2 et de $R^2_{ajusté}$, nous remarquons que la meilleure performance du modèle de régression est obtenue en utilisant les indicateurs correspondants aux quantités de planches et les indicateurs vibratoires délivrés par le capteur M2H. En effet, dans ce cas, la valeur de R^2 de l'ensemble d'entraînement est de 75,45 % et celle de l'ensemble de test est de 72,55%. Les valeurs de $R^2_{ajusté}$ correspondantes (65,22 % et 61,11 %) sont légèrement supérieures à celles obtenues en utilisant seulement les indicateurs relatifs aux quantités des planches (68,52 et 63,53%). Par conséquent, la variabilité est mieux expliquée, de sorte que ce modèle de régression s'ajuste mieux aux données.

4.2.1.1.7 Importance des indicateurs pour modéliser le délignement

Pour déterminer la contribution de chaque indicateur au modèle de régression de la quantité des planches délignées, leurs importances relatives « relative significance » ont pu être évaluées en comparant leurs coefficients de régression standardisés, comme effectué dans (Prabhakar et al., 2006). La Figure 4.9 illustre ainsi l'influence de chaque variable explicative.

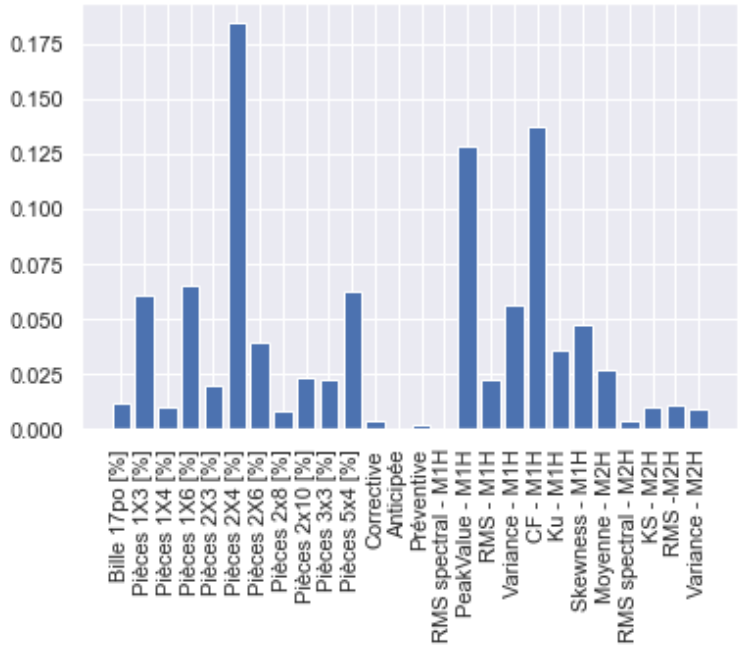


Figure 4.9 Importance relative des indicateurs pour modéliser la quantité des pièces délignées

Les indicateurs les plus significatifs correspondent aux signaux vibratoires et aux dimensions des planches. Les causes probables du délignement des planches sont ainsi liées à l'équipement de coupe VSS ainsi qu'au choix du type de planche à produire.

4.2.1.2 Analyse du déchiquetage des planches

Cette partie présente l'analyse de régression pour modéliser la quantité en pourcentage des planches déchiquetées en fonction des indicateurs utilisés dans la Section 3.3.

Dans les sections qui suivent, différents modèles de régression des planches déchiquetées seront développés, tout d'abord, en fonction de tous les indicateurs correspondants, par la suite en fonction d'une catégorie d'indicateurs à la fois, soit les indicateurs relatifs aux quantités des planches de

bois produites, soit les variables relatives aux interventions de maintenance corrective ou bien les indicateurs vibratoires.

4.2.1.2.1 Modélisation du déchiquetage des planches avec tous les indicateurs

Cette section présente l'analyse de régression linéaire de la quantité des planches déchiquetées en utilisant tous les indicateurs.

Pour utiliser ce modèle, les quatre hypothèses expliquées précédemment doivent être validées :

i. L'hypothèse d'indépendance de la variable réponse est valide. En effet, le graphique des résidus de la quantité des planches délignées en pourcentage ajustée par rapport à l'ordre d'exécution de la collecte de données montre des résidus distribués de manière aléatoire (Figure 4.10).

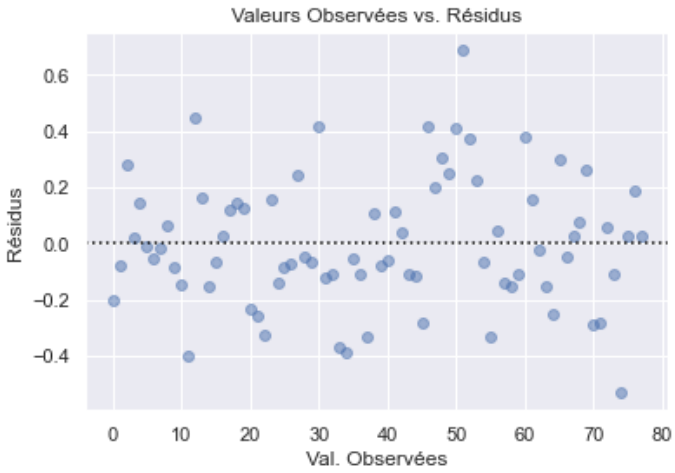


Figure 4.10 Graphique des résidus en fonction de l'ordre d'exécution

ii. L'hypothèse de normalité des résidus est également valide, étant donné que presque tous les points du graphique des résidus sur échelle de probabilité gaussienne forment une ligne,

comme le montre la Figure 4.11. Cela dit, nous observons deux observations qui peuvent être aberrantes.

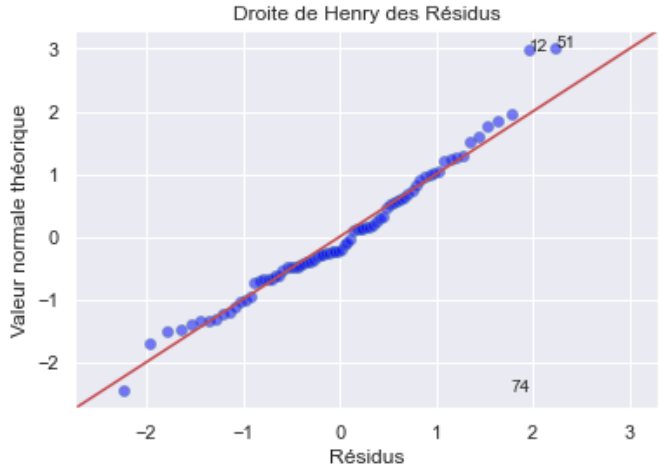


Figure 4.11 Graphiques des résidus sur échelle de probabilité gaussienne

iii. Les hypothèses de linéarité et de homoscédasticité sont non valides. En effet, le graphique des résidus versus les prédictions (Figure 4.12) montre des valeurs du résidu qui ne fluctuent pas aléatoirement autour de zéro.

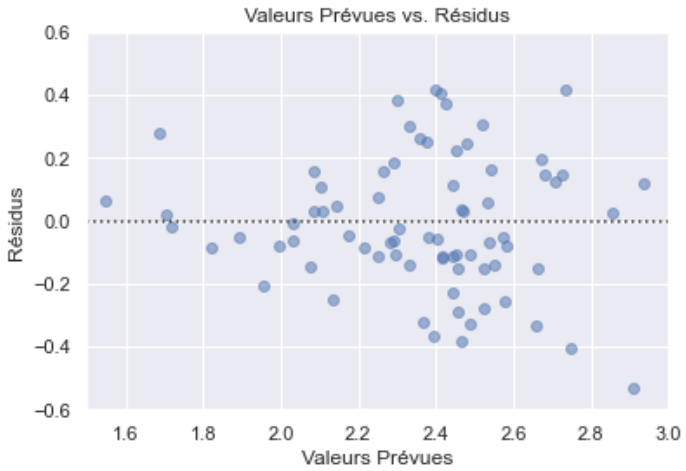


Figure 4.12 Graphique des résidus versus les prédictions

Pour évaluer la performance de ce modèle, nous allons nous baser sur les coefficients de détermination R^2 et $R^2_{ajusté}$. Ces derniers sont présentés dans le Tableau 4.8.

Tableau 4.8 Coefficient de détermination de la régression des planches déchiquetées

	Données d'entraînement		**Données de test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	65,98	42,16	24,02	-29,16

Les coefficients de détermination de l'ensemble test sont faibles révélant un mauvais ajustement du modèle aux données, car plus de 70% de la variabilité n'est pas expliquée par ce modèle.

4.2.1.2.2 Déchiquetage des planches en fonction de la quantité des planches

Pour modéliser la quantité des planches déchiquetée en utilisant la régression, nous allons utiliser cette fois seulement les variables relatives aux quantités des différentes dimensions des planches de bois. Le Tableau 4.9 présente les résultats de cette modélisation.

Tableau 4.9 Résultats de la régression linéaire en utilisant les indicateurs relatifs à la quantité des planches

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	46,89	33,93	25,44	7,25

En examinant les coefficients de déterminations R^2 et $R^2_{ajusté}$ de ce modèle, nous remarquons que les valeurs sont faibles indiquant un mauvais ajustement du modèle aux données. Ces indicateurs sont ainsi insuffisants pour expliquer la variable réponse.

4.2.1.2.3 Déchiquetage des planches en fonction de la vibration

Dans cette section, une modélisation par régression du déchiquetage en fonction des caractéristiques statistiques extraites des signaux vibratoires délivrés par les capteurs M1H et M2H est effectuée. Le Tableau 4.10 montre les coefficients de détermination résultants de l'évaluation de ce modèle de régression.

Tableau 4.10 Résultats de la régression linéaire en utilisant les indicateurs vibratoires

	Entraînement		**Test**	
Indicateurs	**R^2**	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Indicateurs de vibration M1H	20,46	11,81	29,5	21,84
Indicateurs de vibration M2H	8,57	2,85	3,99	-2,01
Indicateurs de vibration M1H et M2H	31,02	18,19	33,57	21, 21

Le tableau montre des valeurs $R^2_{ajusté}$ faibles qui varient entre -2,01 et 21,84%. Ainsi, l'adéquation entre chacun de ces 3 modèles et les données observées est mauvaise. Les caractéristiques vibratoires seules ne sont pas suffisantes pour expliquer la variable réponse.

4.2.1.2.4 Comparaison des modèles de régression pour modéliser le déchiquetage

Les résultats de l'évaluation du modèle de régression sont présentés dans le Tableau 4.11 pour conclure le meilleur modèle de régression capable d'expliquer la quantité des planches déchiquetées.

Tableau 4.11 Comparaison de tous les modèles de régression utilisés pour l'analyse du déchiquetage

		Entraînement		**Test**	
Indicateurs	**Validité**	**R^2**	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Dimensions des PF	Non valide : R^2 test faibles	46,89	33,93	25,44	7,25
Maintenance	Non valide : R^2 faibles	4,04	-1,96	-6,1	-12,73
PF + M1H	Non valide : R^2 test faibles	49,39	28,3	43,01	19,27
PF + M2H	Non valide : R^2 test faibles	54,87	39,43	29,66	5,59
PF + M1H + M2H	Non valide : R^2 faibles	**60,65**	**52,66**	**54,92**	**45,76**
Tous les indicateurs	Non valide : R^2 test faibles	65,98	42,16	24,02	-29,16

Le Tableau 4.11 montre que la meilleure performance du modèle de régression est obtenue en utilisant les indicateurs correspondants aux quantités de planches et les indicateurs vibratoires. Cependant, dans ce cas, la valeur de $R^2_{ajusté}$ de l'ensemble d'entraînement est de 52,66% et celle de l'ensemble de test est de 45,76%. Par conséquent, ce modèle de régression ne s'ajuste pas bien aux données.

4.2.1.2.5 Importance des indicateurs pour modéliser le déchiquetage

Comme dans la Section 4.2.7, pour déterminer la contribution de chaque indicateur au modèle de régression de la quantité des planches déchiquetées, leurs importances relatives ont pu être évaluées en comparant leurs coefficients de régression standardisés, comme montrés dans la Figure 4.13.

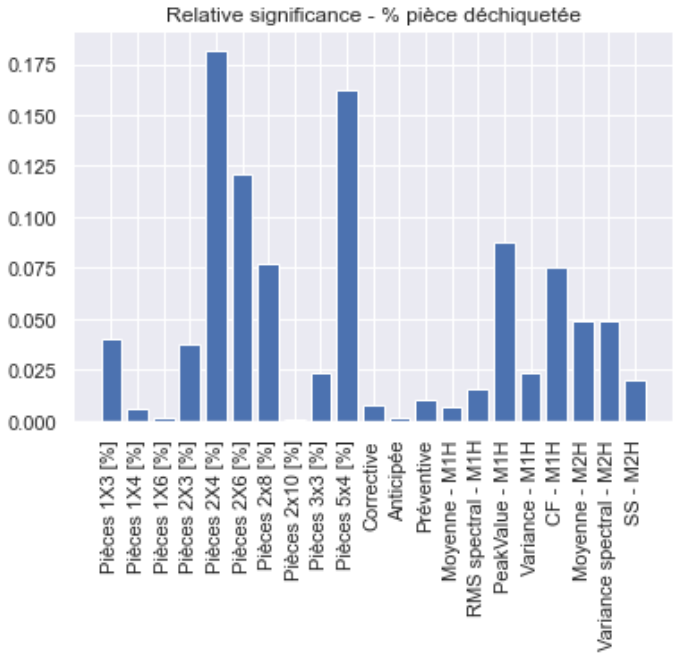


Figure 4.13 Importance relative des indicateurs pour modéliser la quantité des pièces déchiquetées

Les indicateurs les plus significatifs correspondent aux signaux vibratoires et aux dimensions des planches. Les causes probables du déchiquetage des planches sont ainsi liées à l'équipement de coupe VSS ainsi qu'au choix du type de planche à produire.

4.2.1.3 Analyse du rejet des planches

Cette section présentera l'analyse de régression de la quantité des planches rejetée qui dépend de la quantité de billes et de planches ainsi que les caractéristiques vibratoires.

4.2.1.3.1 Modélisation du rejet des planches avec tous les indicateurs

Cette section présente l'analyse de régression linéaire de la quantité des planches rejetées en utilisant tous les indicateurs correspondants. Comme expliqué dans la section précédente, le modèle de régression doit suivre les 4 hypothèses avant de l'utiliser.

i. Hypothèse d'indépendance dans la variable des planches rejetées est valide. En effet, la Figure 4.14 montre des observations indépendantes, car les résidus sont distribués de manière aléatoire.

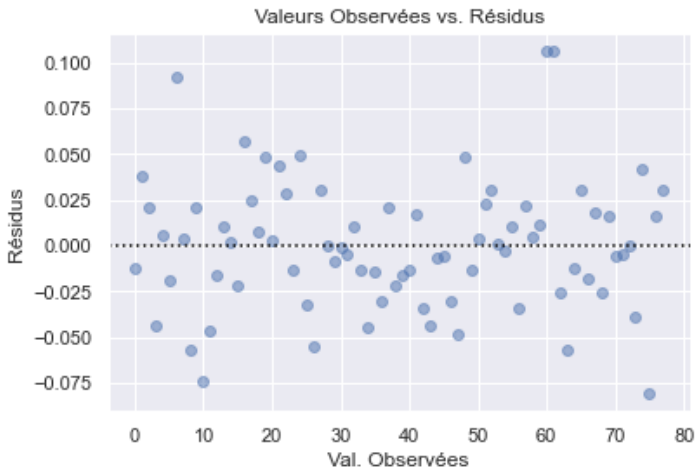


Figure 4.14 Graphique des résidus en fonction de l'ordre d'exécution

ii. L'hypothèse de normalité est valide, vu que la plupart des points des résidus sont alignés sur l'échelle gaussienne dans le graphique des résidus sur échelle de probabilité gaussienne (Figure 4.15). Cependant, nous observons quelques points aberrants ayant un écart par rapport à la droite de régression.

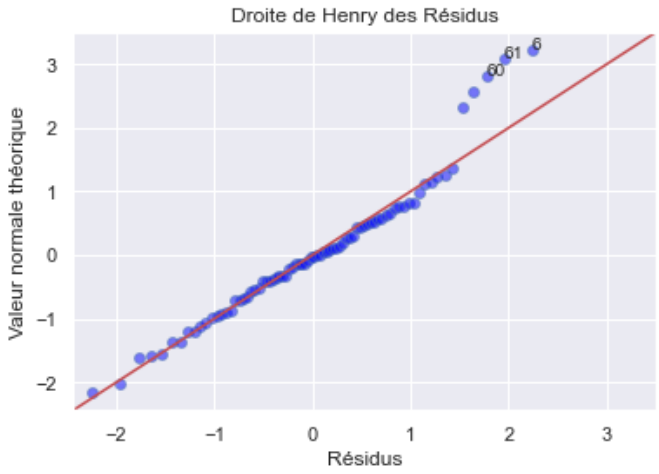


Figure 4.15 Graphiques des résidus sur échelle de probabilité gaussienne

iii. Le graphique illustré dans la Figure 4.16 confirme que les deux dernières hypothèses (la linéarité et l'homoscédasticité) ne sont pas respectées. En effet, les valeurs ne fluctuent pas aléatoirement autour de zéro. La plupart des valeurs sont concentrées autour de 0 et 0,2.

Figure 4.16 Graphiques des résidus versus les prédictions

Pour évaluer la performance de ce modèle, nous allons nous baser sur les coefficients de détermination R^2 et $R^2_{ajusté}$. Ces derniers sont présentés dans le Tableau 4.12. Le coefficient de détermination de ce modèle est de 81,81%. Ce modèle correspond bien aux données. Il existe une bonne concordance entre les valeurs expérimentales et prédites puisque le coefficient de détermination ajusté calculé est de 75,85 %. Pour l'ensemble d'entraînement, la valeur de R^2 et $R^2_{ajusté}$ sont respectivement de 69,93 et 60,08, ce qui est acceptable.

Tableau 4.12 Coefficient de détermination de la régression des planches rejetées

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	81,81	75,85	69,93	60,08

Pour valider ce modèle de régression linéaire, toutes les 4 hypothèses doivent être respectées. Ce qui n'est pas le cas comme le montre le tableau suivant. En effet, la seule hypothèse validée est l'indépendance de la variable réponse relative à la quantité des planches rejetée. Dans ce cas, le modèle de régression linéaire ne convient pas à ces données pour modéliser la quantité du pourcentage de ce type de planches.

Tableau 4.13 Résumé des hypothèses de la régression des planches rejetées

Hypothèse	**Graphique**	**Validité de l'hypothèse**
Indépendance	Graphique des résidus en fonction de l'ordre d'exécution	Oui
Normalité	Graphique des résidus sur échelle de probabilité gaussienne	Non
Linéarité	Graphique des résidus versus prédictions	Non
Homoscédasticité		Non

4.2.1.3.2 Modélisation du rejet des planches en fonction de la vibration

Dans cette section, une modélisation par régression du rejet en fonction des caractéristiques statistiques extraites des signaux vibratoires délivrés par les capteurs M1H et M2H est effectuée. Le Tableau 4.14 montre les coefficients de détermination résultants de l'évaluation de ce modèle de régression.

Tableau 4.14 Résultats de la régression linéaire en utilisant les indicateurs vibratoires

	Entraînement		**Test**	
Indicateurs	**R^2**	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Indicateurs de vibration M1H	34,65	32	42,74	40,42
Indicateurs de vibration M2H	39,12	34,9	38,63	34,37
Indicateurs de vibration M1H et M2H	62,44	58,08	60,35	55,75

Le tableau montre que les meilleures performances du modèle de régression sont obtenues lorsque toutes les caractéristiques vibratoires sont utilisées. Dans ce cas, les valeurs $R^2_{ajusté}$ sont tout de même considérées faibles, de 58,08% pour l'ensemble d'entraînement et de 55,75% pour l'ensemble de test. Ainsi, l'adéquation de chacun de ces 3 modèles avec les données observées est mauvaise. Les caractéristiques vibratoires seules ne sont pas suffisantes pour expliquer la variable réponse.

4.2.1.3.3 Comparaison des modèles de régression pour modéliser le rejet

Pour trouver un modèle précis permettant de modéliser la quantité des planches rejetées en pourcentage, nous avons utilisé différentes combinaisons d'indicateurs plusieurs méthodes. Le Tableau 4.15 explore le modèle le plus précis avec les meilleures performances.

Le tableau montre que la meilleure performance du modèle de régression est obtenue en utilisant tous les indicateurs avec une valeur de $R^2_{ajusté}$ de l'ensemble d'entraînement de 81,81% et celle de l'ensemble de test de 75,85%. Ces valeurs indiquent que le modèle s'ajuste bien aux données. Cependant, dans ce cas, les hypothèses de régression correspondantes ne sont pas valides.

Tableau 4.15 Comparaison de tous les modèles de régression utilisés pour l'analyse du rejet

		Entraînement		Test	
Indicateurs	**Validité**	**R^2**	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Matière première MP	Non valide : R^2 test faible	41,73	40,97	24,63	23,64
Dimensions du PF	Non valide : R^2 test faible	50,79	43,45	-3,12	-18,51
MP + PF	Non valide : R^2 test faible	60,99	54,49	7,88	-7,48
PF + M1H	Non valide : R^2 test faible	58,08	49,56	14,0	-3,47
PF + M2H	Hypothèses non valides	72,56	65,92	61,4	52,06
MP + PF + M1H	Non valide : R^2 test faible	72,09	65,89	13,28	-5,99
MP + PF + M2H	Hypothèses non valides	77,12	71,12	67,25	58,66
Tous les indicateurs	Hypothèses non valides	**81,81**	**75,85**	**69,93**	**60,08**

4.2.1.3.4 Importance des indicateurs pour modéliser le déchiquetage

Afin déterminer l'importance relative de chaque indicateur au modèle de régression de la quantité des planches rejetées, la Figure 4.17 est présentée.

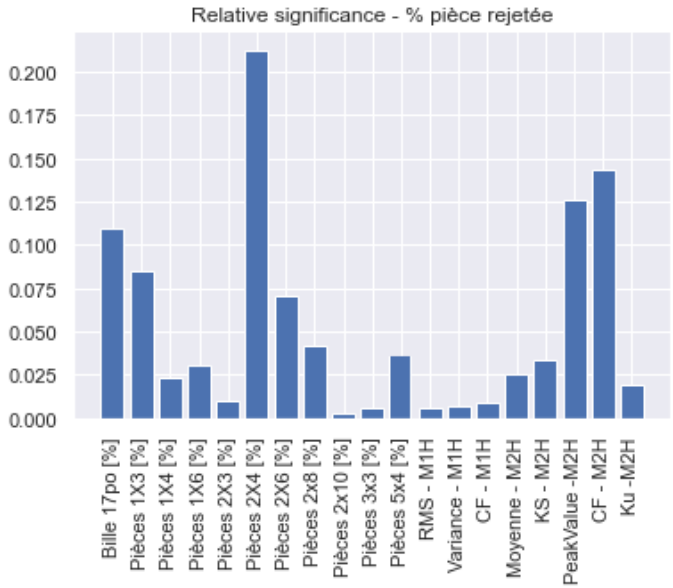


Figure 4.17 Importance relative des indicateurs pour modéliser la quantité des pièces rejetées

Les indicateurs les plus significatifs correspondent aux signaux vibratoires et aux dimensions des planches et des billes.

4.2.2 Analyse basée sur le modèle d'ensemble

Cette section présente les résultats de l'application du modèle d'ensemble de régression qui combinent les 3 modèles de k-NN, GB et RF. Ces résultats seront comparés au meilleur modèle de régression linéaire obtenue pour la même variable de non-qualité. Le modèle d'ensemble utilisera ainsi les mêmes variables explicatives que celui du meilleur modèle de régression linéaire. Dans le cas où la variable réponse est la quantité des planches délignées, seuls les indicateurs relatifs aux quantités de planches et aux signaux vibratoires délivrés par le capteur M2H sont utilisés. En effet, ces derniers sont les plus significatifs, résultant au modèle le plus performant, selon le tableau de comparaison dans la Section 4.2.1.1.6. Dans le cas où la variable réponse est la quantité des planches déchiquetées, les variables explicatives considérées sont les indicateurs relatifs aux quantités de planches et aux signaux vibratoires. Et si la variable réponse correspond à la quantité des planches rejetées, tous les indicateurs seront considérés, selon la Section 4.2.1.3.3. Le Tableau 4.16 présente le modèle k-NN correspondant à chacune des variables de non-qualité.

Tableau 4.16 Résultats du modèle k-NN

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	75,05	64,66	73,27	62,13
Planches déchiquetées [%]	68,24	61,79	27,64	12,94
Planches rejetées [%]	78,84	71,91	56,65	42,45

Les résultats montrent que le modèle k-NN est acceptable, dans le cas où la variable à expliquer correspond à la quantité des planches delignées, puisque le coefficient de détermination ajusté est de 64,66 % pour les données d'entraînement et de 62,13 % pour les données test. Ces valeurs sont meilleures que celles du modèle de régression linéaire en utilisant les données test. Si la quantité des planches déchiquetées est considérée comme étant la variable réponse, les valeurs de R^2 sont

plus faibles que celle de la régression linéaire en utilisant les données test. En revanche, le modèle k-NN est meilleur que la régression linéaire en regardant les résultats des données d'entraînement. En considérant le rejet, d'un autre côté, les coefficients de régression du k-NN sont plus faibles que ceux de la régression linéaire. Cependant, dans ce cas, ce modèle ne nécessite pas la validation des hypothèses présentées dans la Section 4.1.

Le Tableau 4.17 montre les coefficients de détermination résultants de l'évaluation du modèle Boosting de gradient (Xu et al.).

Tableau 4.17 Résultats du modèle GB

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	75,28	64,98	70,89	58,76
Planches déchiquetées [%]	45,18	34,04	47,03	36,27
Planches rejetées [%]	80,09	73,56	60,40	47,43

Les résultats du tableau ci-dessus montrent que les valeurs des coefficients de régression du modèle GB sont plus faibles que celles du modèle de régression linéaire pour les 3 différentes variables de non-qualité.

Le Tableau 4.18 présente les résultats du modèle de la forêt d'arbres décisionnels RF, en utilisant les mêmes indicateurs correspondant à chaque variable réponse.

Tableau 4.18 Résultats du modèle RF

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	86,01	80,18	59,71	42,92
Planches déchiquetées [%]	42,19	30,45	45,68	34,65
Planches rejetées [%]	78,28	71,16	57,03	42,95

De ce tableau, nous remarquons que les résultats sont meilleurs que ceux de la régression linéaire pour modéliser la quantité des planches délignées. Cependant, ceci n'est pas le cas si les autres variables réponses sont considérées.

Le Tableau 4.19 présente le modèle d'ensemble qui combine les 3 modèles de k-NN, GB et RF pour expliquer chacune des variables de non-qualité.

Tableau 4.19 Résultats du modèle d'ensemble de régression

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	82,45	75,14	76,65	66,93
Planches déchiquetées [%]	62,86	55,31	57,78	49,21
Planches rejetées [%]	79,28	72,49	58,62	45,07

En regardant le modèle relatif aux planches délignées, nous remarquons que les coefficients de détermination relatifs aux données d'entraînement et de test sont bons. Ce modèle correspond bien aux données. Il existe une bonne concordance entre les valeurs observées et prédites puisque le coefficient de détermination ajusté est de 75,14%. Nous notons également que les performances de ce modèle sont meilleures que celui de la régression linéaire, ayant une valeur de R^2 ajusté de 65,22%.

Le modèle d'ensemble correspondant à la quantité des planches déchiquetées montre aussi des résultats meilleurs que ceux de la régression linéaire, étant donné que le R^2 ajusté (55.31%) est plus élevé que celui de la régression linéaire qui est de 52.66. Cependant, les coefficients de détermination de ce modèle sont considérés tout de même faibles révélant un mauvais ajustement du modèle aux données.

En examinant le modèle d'ensemble relatif à la quantité des planches rejetées, nous observons que les coefficients de détermination correspondants sont plus faibles que ceux de la régression linéaire. Cela dit, les hypothèses du modèle de la régression linéaire ne sont pas valides. Les résultats de ce modèle d'ensemble sont ainsi considérés meilleurs, ayant un R^2 ajusté d'entraînement acceptable

de valeur 72,49%. Cependant, les coefficients de régression correspondants aux données test sont faibles, indiquant que ce modèle surajuste les données.

4.3 Approche de régression à chaîne

Cette section présente les résultats de l'application des 4 modèles de régression, la régression linéaire, le k-NN, le Boosting de gradient GB et la forêt d'arbre décisionnel RF, ainsi que le modèle d'ensemble de régression combinant les 3 derniers modèles, en se basant sur l'approche de régression à chaîne RC. Le Tableau 4.20 montre les résultats de l'évaluation du modèle k-NN basé sur l'approche RC.

Tableau 4.20 Résultats du modèle k-NN basé sur l'approche RC

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	75,05	64,66	73,27	62,13
Planches déchiquetées [%]	68,58	62,20	27,24	12,47
Planches rejetées [%]	79,68	72,07	58,37	42,76

Ce tableau montre des résultats meilleurs pour les deux modèles relatifs aux planches déchiquetées et rejetées en les comparant à ceux des modèles basés sur l'approche directe (ST), que ce soit pour l'ensemble d'entraînement ou l'ensemble de test. En effet, dans le cas où la variable réponse est les planches rejetées, la valeur de R^2 de l'ensemble d'entraînement est de 79,68%, comparée à une valeur de 78,84% et celle de l'ensemble de test est de 58,37%, comparée à une valeur de 56,65.

Le Tableau 4.21 montre les coefficients de détermination résultants de l'évaluation du modèle GB basé sur l'approche RC.

Tableau 4.21 Résultats du modèle GB basé sur l'approche RC

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	75,28	64,98	70,89	58,76
Planches déchiquetées [%]	46,87	36,08	53,06	43,52
Planches rejetées [%]	81,11	74,03	63,18	49,37

Considérant les variables relatives aux quantités des planches déchiquetées et rejetées, les résultats du modèle GB basé sur l'approche RC sont meilleurs par rapport à ceux du modèle basé sur l'approche ST.

Les résultats du modèle RF basé sur l'approche RC sont présentés dans le Tableau 4.22.

Tableau 4.22 Résultats du modèle RF basé sur l'approche RC

	Entraînement		**Test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	86,01	80,18	59,71	42,92
Planches déchiquetées [%]	42,19	30,45	45,68	34,65
Planches rejetées [%]	79,06	71,21	58,52	42,97

De ce tableau, nous remarquons que les coefficients sont meilleurs que ceux du modèle basé sur l'approche ST seulement si la variable réponse considérée est la quantité des planches rejetées. Dans les deux autres cas, les résultats ne changent pas.

Le Tableau 4.23 présente le modèle d'ensemble qui combine les 3 modèles de k-NN, GB et RF pour expliquer chacune des variables de non-qualité.

Tableau 4.23 Résultats du modèle d'ensemble de régression basé sur l'approche RC

	Données d'entraînement		**Données de test**	
	R^2	**$R^2_{ajusté}$**	**R^2**	**$R^2_{ajusté}$**
Planches délignées [%]	82,45	75,14	76,65	66,93
Planches déchiquetées [%]	63,84	56,49	58,70	50,31
Planches rejetées [%]	80,17	72,74	60,62	45,85

Ce tableau montre des résultats meilleurs pour les deux modèles relatifs aux planches déchiquetées et rejetées en les comparant à ceux des modèles basés sur l'approche directe (ST), que ce soit pour l'ensemble d'entraînement ou l'ensemble de test. En effet, dans le cas où la variable réponse est les planches déchiquetées, la valeur de R^2 de l'ensemble d'entraînement est de 56,49%, comparée à une valeur de 55,31% et celle de l'ensemble de test est de 50,31%, comparée à une valeur de 49,21. La quantité de planches délignées affecte ainsi le comportement de celle des planches déchiquetées. Dans le cas où la variable réponse est les planches rejetées, les valeurs de R^2 de l'ensemble d'entraînement et de test sont respectivement de 72,74% et de 50,31%, légèrement supérieures à celles des modèles basés sur l'approche ST de 72,49 et 45,07%. Par conséquent, la quantité des planches rejetées peut aussi dépendre de celles des autres types de non-qualité.

Dans cette section, nous nous étions aperçus lors de la validation des hypothèses de régression, qu'il avait présence de données aberrantes en illustrant le graphique des résidus sur échelle de probabilité gaussienne. A cet effet, il est possible d'effectuer une étude plus approfondie pour la détection des données aberrantes et leur traitement. Cela pourrait améliorer les performances des modèles de régression développés.

Nous remarquons également que le modèle d'ensemble de régression présente les meilleures performances par rapport aux autres modèles de régression, ayant les plus grandes valeurs de coefficients de détermination. L'approche de régression à chaine montre aussi qu'il existe une dépendance entre les variables de non-qualité. En effet, la production des planches délignées affecte celle des planches déchiquetées et rejetées.

CHAPITRE 5 CONCLUSION ET RECOMMANDATIONS

L'objectif de ce travail était de modéliser trois différents types de non-qualité présents sur les planches de bois lors du processus du sciage afin d'identifier les causes primaires de la mauvaise qualité de ces produits finis. Pour ce faire, des techniques de modélisation ont été développées pour expliquer trois variables de non-qualité correspondantes aux planches déchiquetées, délignées et rejetées en fonction de différents indicateurs qui caractérisent la quantité de billes de différents diamètres, la quantité des planches de différents dimensions, le nombre d'interventions de maintenance et les quantités d'énergie vibratoire produites par le moteur de l'équipement de sciage.

Avant de procéder à la modélisation, un prétraitement des données a été réalisé. En effet, afin d'analyser l'effet des vibrations sur la qualité des planches, une extraction et une sélection des caractéristiques statistiques ont été effectuées à partir des signaux vibratoires.

Une première partie de ce mémoire consistait à réaliser une analyse globale basée sur les modèles de classification pour détecter la non-qualité dans les planches. Un étiquetage des jours a été effectué comme ayant ou non une production importante d'un certain type de non-qualité. A cet effet, les variables réponses de non-qualité ont été transformées en variables catégoriques, en se basant sur des seuils de dichotomisation déterminés selon des méthodes statistiques. Dans cette partie, la technique de l'analyse logique des données LAD a été utilisée en se basant sur le logiciel cbmLAD.

Les résultats de LAD nous a permis de générer des règles de regroupement qui représentent chaque type de non-qualité du produit fini notamment le délignement, le déchiquetage et le rejet. Ces résultats montrent que la performance des modèles qui expliquent le délignement et le déchiquetage sont meilleures lorsque seulement les indicateurs relatifs aux quantités des planches et les indicateurs vibratoires du capteur M2H sont utilisés comme variables d'entrée. L'évaluation du modèle de classification qui explique le rejet présente quant à lui des résultats meilleurs lorsque tous les indicateurs sont utilisés.

Une deuxième partie d'analyse consiste à identifier les conditions sous lesquelles chaque type de non-qualité apparaît sur les planches, en se basant sur des modèles de régression à multi-sorties, pour expliquer les différents types de non-qualité trouvée dans les planches produites lors de la

coupe. Deux approches différentes ont été utilisées pour effectuer l'analyse de régression à multi-sorties : l'approche directe basée sur la régression à sortie unique et l'approche de régression à chaîne RC qui est capable de prendre en considération les dépendances entre les différents types de non-qualité. Une analyse d'importance a été également effectuée sur les variables entrantes afin de détecter celles qui contribuent significativement aux modèles de régression construits basés sur l'approche directe.

Les résultats de l'analyse d'importance confirment que les indicateurs les plus significatifs aux modèles qui expliquent le délignement et le déchiquetage correspondent aux signaux vibratoires et aux dimensions des planches, tandis que les indicateurs les plus significatifs au modèle qui explique le rejet correspondent aux signaux vibratoires et aux dimensions des planches et des billes. Un modèle d'ensemble basé sur l'approche RC a été proposé. Ce modèle présente de meilleures performances par rapport à tous les autres modèles de régression et montre qu'il existe une dépendance importante entre les types de non-qualité présents sur les planches.

Ce projet de recherche, nous a permis de bien appréhender les points suivants :

- Les deux analyses de classification et de régression montrent des résultats similaires, notamment en examinant les performances des différents modèles relatifs aux variables de non-qualités et aux facteurs qui influencent ces variables.
- Les indicateurs qui affectent le délignement et le déchiquetage sont liées à l'état de santé de l'équipement de coupe en analysant les signaux vibratoires et aux dimensions des planches.
- Les indicateurs qui influencent le rejet sont liés aux dimensions des planches et de la bille ainsi à l'état de santé de l'équipement de coupe.
- Il important de tenir en compte les dépendances entre les variables réponses dans la modélisation. Ceci est prouvé lors de l'utilisation de l'approche RC où nous trouvons que la quantité de planches délignées affecte celle des planches déchiquetées et rejetées.

Les résultats obtenus par ces techniques sont perfectibles pour autant qu'un accès à un plus grand volume de données en termes d'observations et de variables, tels que : la qualité des billes de bois,

le nombre de nœuds ou les fissures dans les billes, les réglages d'optimisation de coupe, l'état des scies et des capteurs, et les signaux vibratoires des autres équipements de sciages.

De plus, il serait possible d'utiliser d'autres méthodes de modélisation. En guise de recommandation pour les futurs travaux dans ce volet, nous suggérons de travailler sur :

- Le développement du modèle d'ensemble en se basant sur d'autres approches de résolution de problèmes de régression à multi-sorties comme la méthode de régression partielle de moindres carrés (PLS) « Partial Least Square » ou encore l'approche d'ensemble de régression à chaîne ERC qui consiste à utiliser un ensemble de RC constitué de toutes les chaînes possibles.
- L'utilisation des modèles alternatifs de régression comme la régression MARS « Multivariate Adaptive Regression Spline », ou la régression logistique pour comparer leurs performances avec ceux qui furent développés dans ce projet.
- Une étude de sensibilité locale et globale en utilisant des techniques plus avancées peut être effectuée afin de quantifier l'influence des facteurs sur les variables de non-qualité.
- L'implémentation du système de contrôle de processus statistique ou encore SPC « Statistical Process Control » afin de déterminer si les événements de non-qualité sont assignables ou inhérents.
- Une analyse de la capabilité et de la variabilité du processus en utilisant par exemple l'indicateur de capabilité process CPK.

RÉFÉRENCES

Ab-Samat, H., & Kamaruddin, S. (2014). Opportunistic maintenance (OM) as a new advancement in maintenance approaches: A review. *Journal of Quality in Maintenance Engineering*.

Abbasion, S., Rafsanjani, A., Farshidianfar, A., & Irani, N. (2007). Rolling element bearings multi-fault classification based on the wavelet denoising and support vector machine. *Mechanical Systems and Signal Processing, 21*(7), 2933-2945. https://doi.org/10.1016/j.ymssp.2007.02.003

Aguilar, R., Ramírez, E., Haach, V. G., & Pando, M. A. (2016). Vibration-based nondestructive testing as a practical tool for rapid concrete quality control. *Construction and Building Materials, 104*, 181-190. https://doi.org/10.1016/j.conbuildmat.2015.12.053

Amaral, J., Silva, J. R. C., de Andrade, D. S. M., Ferreira, L. T., Quirino, T. M., & Quirino, J. (26-30 Aout 2019). *Machine Learning Algorithms Applied to the Inference of the Flow Rate in a Non-intrusive Thermal Flow meter*. 2019 4th International Symposium on Instrumentation Systems, Circuits and Transducers (INSCIT), Sao Paulo, Brazil (p. 1-6). https://doi.org/10.1109/INSCIT.2019.8868345

Antoni, J., & Randall, R. B. (2006). The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines. *Mechanical systems and signal processing, 20*, 308-331. https://doi.org/10.1016/j.ymssp.2004.09.002

Borchani, H., Varando, G., Bielza, C., & Larranaga, P. (2015). A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining Knowledge Discovery, 5*, 216-233. https://doi.org/10.1002/widm.1157

Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A., Mayoraz, E., & Muchnik, I. (2000). An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering, 12*, 292-306. https://doi.org/10.1109/69.842268

Brauer, M. (2002). L'analyse des variables indépendantes continues et catégorielles: Alternatives à la dichotomisation. *L'année psychologique, 102*, 449-484. https://doi.org/10.3406/psy.2002.29602

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

Cai, Y., Chow, M.-Y., Lu, W., & Li, L. (2010). Statistical feature selection from massive data in distribution fault diagnosis. *IEEE Transactions on Power Systems, 25*, 642-648. https://doi.org/10.1109/TPWRS.2009.2036924

Callens, A., Morichon, D., Abadie, S., Delpey, M., & Liquet, B. (2020). Using Random forest and Gradient boosting trees to improve wave forecast at a specific location. *Applied Ocean Research, 104*, 102339. https://doi.org/10.1016/j.apor.2020.102339

Carnero, M. C., González-Palma, R., Almorza, D., Mayorga, P., & López-Escobar, C. (2010). Statistical quality control through overall vibration analysis. *Mechanical Systems and Signal Processing, 24*, 1138-1160. https://doi.org/10.1016/j.ymssp.2009.09.007

Ciupke, K. (2005). A comparative study on methods of reduction and selection of information in technical diagnostics. *Mechanical systems signal processing, 19*, 919-938. https://doi.org/10.1016/j.ymssp.2004.08.003

Cohen, J. (1983). The cost of dichotomization. *Applied psychological measurement, 7*, 249-253. https://doi.org/10.1177/014662168300700301

Côté, N. (2013). *Potentiel de récupération de composantes de fermes de toit dans les sciages flacheux produits en scierie: étude de cas* [Thèse de doctorat, Université Laval].

Demirel, K. C., Sahin, A., & Albey, E. (26-28 juillet 2019). *Ensemble Learning based on Regressor Chains: A Case on Quality Prediction.* 8th International Conference on Data Science, Technology and Applications, DATA 2019, Prague, Czech Republic. https://doi.org/10.5220/0007932802670274

Desfor, G. (2003). *Guide de façonnage et de mise en marché du bois*. Syndicat des producteurs de bois de la région de Montréal.

Dupuis, C., Gamache, M., & Pagé, J.-F. (2012). Logical analysis of data for estimating passenger show rates at Air Canada. *Journal of Air Transport Management, 18*, 78-81. https://doi.org/10.1016/j.jairtraman.2011.10.004

Dyer, D., & Stewart, R. (1978). Detection of rolling element bearing damage by statistical vibration analysis. *Journal of Mechanical Design*. https://doi.org/10.1115/1.3453905

Esterby, S. (1989). *Statistical Methods for the Assessment of Point Source Pollution*. Springer, Dordrecht.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis, 38*, 367-378.

Ghasemi, A., Yacout, S., & Ouali, M.-S. (2009). Evaluating the reliability function and the mean residual life for equipment with unobservable states. *IEEE Transactions on Reliability, 59*, 45-54. https://doi.org/10.1109/TR.2009.2034947

Grmanová, G., Laurinec, P., Rozinajová, V., Ezzeddine, A. B., Lucká, M., Lacko, P., . . . Návrat, P. (2016). Incremental ensemble learning for electricity load forecasting. *Acta Polytechnica Hungarica, 13*(2), 97-117.

Hammer, P. L., Kogan, A., & Lejeune, M. A. (2012). A logical analysis of banks' financial strength ratings. *Expert Systems with Applications, 39*, 7808-7821. https://doi.org/10.1016/j.eswa.2012.01.087

Haq, I. U., Gondal, I., Vamplew, P., & Brown, S. (28-30 Novembre 2018). *Categorical features transformation with compact one-hot encoder for fraud detection in distributed environment* Australasian Conference on Data Mining, Bathurst, NSW, Australia. https://doi.org/10.1007/978-981-13-6661-1_6

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Helmi, H., & Forouzantabar, A. (2019). Rolling bearing fault detection of electric motor using time domain and frequency domain features extraction and ANFIS. *IET Electric Power Applications, 13*, 662-669. https://doi.org/10.1049/iet-epa.2018.5274

Hogg, R. V., & Ledolter, J. (1992). *Applied statistics for engineers and physical scientists*. Pearson Higher Ed.

Iskra, P., & Hernández, R. E. (2012). Toward a process monitoring of CNC wood router. Sensor selection and surface roughness prediction. *Wood science and technology, 46*(1), 115-128. https://doi.org/10.1007/s00226-010-0378-7

Jackson, M. R., Parkin, R. M., & Brown, N. (2002). Waves on wood. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 216*(4), 475-497. https://doi.org/10.1243/0954405021520175

James, E., & Robinson, C. (22-25 octobre 2001). *Description of peakvue and illustration of its wide array of applications in fault detection and problem severity assessment* [Communication de conférence]. Emerson Process Management Reliability Conference.

Kleinosky, L. R., Yarnal, B., & Fisher, A. (2007). Vulnerability of Hampton Roads, Virginia to storm-surge flooding and sea-level rise. *Natural Hazards, 40*(1), 43-70. https://doi.org/10.1007/s11069-006-0004-z

Kramer, O. (2013). *Dimensionality reduction with unsupervised nearest neighbors*. Springer.

Kuljich, S., Hernández, R. E., & Blais, C. (2017). Effects of cutterhead diameter and log in feed position on surface quality of black spruce cants produced by a chipper-canter. *Wood and Fiber Science, 49*(3), 1-14.

Kumar, V., & Sahu, M. (2021). Evaluation of nine machine learning regression algorithms for calibration of low-cost PM2. 5 sensor. *Journal of Aerosol Science, 157*, 105809. https://doi.org/10.1016/j.jaerosci.2021.105809

Kumari, B., & Swarnkar, T. (2011). Filter versus wrapper feature subset selection in large dimensionality micro array: A review. *International Journal of Computer Science and Information Technologies, 2*(3), 1048-1053.

Kurdthongmee, W., & Suwannarat, K. (2019). *Locating Wood Pith in a Wood Stem Cross Sectional Image Using YOLO Object Detection.* 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI) (p. 1-6).

Lampinen, J., & Smolander, S. (31 octobre -2 novembre 1994). *Wood defect recognition with self-organizing feature selection* [Communication de conférence]. Intelligent Robots and Computer Vision XIII: Algorithms and Computer Vision, Boston, MA, USA. https://doi.org/10.1117/12.188910

Lebold, M., McClintic, K., Campbell, R., Byington, C., & Maynard, K. (2000). *Review of vibration analysis methods for gearbox diagnostics and prognostics.* Proceedings of the 54th meeting of the society for machinery failure prevention technology (vol. 634, p. 16). https://doi.org/10.1.1.462.9240

Lemaster, R. L., Lu, L., & Jackson, S. (2000). The use of process monitoring techniques on a CNC wood router. Part 2. Use of a vibration accelerometer to monitor tool wear and workpiece quality. *Forest products journal*, *50*(9), 59-64.

Lerch, A. (2012). *An introduction to audio content analysis: Applications in signal processing and music informatics*. Wiley-IEEE Press.

Li, B., Zhang, P.-l., Wang, Z.-j., Mi, S.-s., & Zhang, Y.-t. (2011). Gear fault detection using multi-scale morphological filters. *Measurement*, *44*(10), 2078-2089. https://doi.org/10.1016/j.measurement.2011.08.010

Loukopoulos, P., Zolkiewski, G., Bennett, I., Pilidis, P., Duan, F., & Mba, D. (2017). Dealing with missing data as it pertains of e-maintenance. *Journal of Quality in Maintenance Engineering*, *23*, 260-278. https://doi.org/10.1108/JQME-08-2016-0032

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods*, *7*, 19. https://doi.org/10.1037/1082-989X.7.1.19

Marenče, J., Šega, B., & Gornik Bučar, D. (2020). Monitoring the Quality and Quantity of Beechwood from Tree to Sawmill Product. *Croatian Journal of Forest Engineering*, *41*(1), 119-128. https://doi.org/10.5552/crojfe.2020.613

Mashudi, N. A., Rossli, S. A., Ahmad, N., & Noor, N. M. (1-3 mars 2021). *Comparison on Some Machine Learning Techniques in Breast Cancer Classification* [Communication de conférence]. 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), Virtual, Langkawi Island, Malaysia. https://doi.org/10.1109/IECBES48179.2021.9398837

Mathew, V., Toby, T., Singh, V., Rao, B. M., & Kumar, M. G. (20-21 Décembre 2017). *Prediction of Remaining Useful Lifetime (RUL) of turbofan engine using machine learning.* 2017 IEEE International Conference on Circuits and Systems (ICCS), Thiruvananthapuram, India.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John wiley & sons.

Mortada, M.-A., Yacout, S., & Lakis, A. (2014). Fault diagnosis in power transformers using multi-class logical analysis of data. *Journal of Intelligent Manufacturing*, *25*, 1429-1439. https://doi.org/10.1007/s10845-013-0750-1

Mortada, M. A., Yacout, S., & Lakis, A. (2011). Diagnosis of rotor bearings using logical analysis of data. *Journal of Quality in Maintenance Engineering*. https://doi.org/10.1108/13552511111180186

Nasir, V., & Cool, J. (2019). Optimal power consumption and surface quality in the circular sawing process of Douglas-fir wood. *European Journal of Wood and Wood Products*, *77*(4), 609-617. https://doi.org/10.1007/s00107-019-01412-z

Nayana, B., & Geethanjali, P. (2017). Analysis of statistical time-domain features effectiveness in identification of bearing faults from vibration signal. *IEEE Sensors Journal*, *17*, 5618-5625. https://doi.org/10.1109/JSEN.2017.2727638

Nguyen, Q. H., Ly, H.-B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., . . . Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering, 2021*.

Parent, B. (2010). *Ressources et industries forestières Portrait statistique édition 2010*. https://mffp.gouv.qc.ca/publications/forets/connaissances/stat_edition_complete/preface.pdf

Plante, T., Nejadpak, A., & Yang, C. X. (2-5 novembre 2015). *Faults detection and failures prediction using vibration analysis* [Communication de conférence]. 2015 IEEE AUTOTESTCON, National Harbor, MD, USA. https://doi.org/10.1109/AUTEST.2015.7356493

Prabhakar, Y. S., Gupta, M. K., Roy, N., & Venkateswarlu, Y. (2006). A high dimensional QSAR study on the aldose reductase inhibitory activity of some flavones: topological descriptors in modeling the activity. *Journal of chemical information*

modeling, *46*, 86-92. https://doi.org/10.1007/s00267-011-9764-7

Pujianto, U., Wibawa, A. P., & Akbar, M. I. (23-24 octobre 2019). *K-Nearest Neighbor (K-NN) based Missing Data Imputation* [Communication de conférence]. 2019 5th International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia. https://doi.org/10.1109/ICSITech46713.2019.8987530

Quesada-Pineda, H. J., & Arias, E. (2015). *Statistical Process Control: Applications and Examples for Forest Products Industries* [Mémoire de thèse, Virginia Polytechnic Institute and State University]. Publications, Virginia Cooperative Extension (VCE).

Rafiee, J., Tse, P., Harifi, A., & Sadeghi, M. (2009). A novel technique for selecting mother wavelet function using an intelli gent fault diagnosis system. *Expert Systems with Applications*, *36*(3), 4862-4875. https://doi.org/10.1016/j.eswa.2008.05.052

Ragab, A., Yacout, S., Ouali, M.-S., & Osman, H. (2019). Prognostics of multiple failure modes in rotating machinery using a pattern-based classifier and cumulative incidence functions. *Journal of Intelligent Manufacturing*, *30*, 255-274. https://doi.org/10.1007/s10845-016-1244-8

Ragab, A. R. A. (2014). *Fault Prognostics Using Logical Analysis of Data and Non-Parametric Reliability Estimation Methods* [Thèse de doctorat, École Polytechnique de Montréal].

Rahman, M. A., & Akter, Y. A. (6-7 mai 2020). *Multi-lingual Author Profiling: Predicting Gender and Age from Tweets!* [Communication de conférence]. International Conference on Image Processing and Capsule Networks, Bangkok, Thailand. https://doi.org/10.1007/978-3-030-51859-2_46

Rawal, S., Gupta, S., & Singh, S. (2017). Predicting missing values in a dataset: challenges and approaches. *International Journal of Recent Research Aspects, 4*(3), 34-38. https://www.ijrra.net/Vol4issue3/IJRRA-04-03-07.pdf

Ribarits, S., Carmond, P., Romilly, D., & Evans, P. (7-9 mai 2007). *The Effect of Tool Geometry and Machine Parameters on the Surface Finish of Machined MDF* [Communication de conférence]. 18th International Wood Machining Seminar, Vancouver, Canada.

Rudakov, N., Eerola, T., Lensu, L., Kälviäinen, H., & Haario, H. (9-12 Octobre 2018). *Detection of mechanical damages in sawn timber using convolutional neural networks.* 40th German Conference on Pattern Recognition, GCPR 2018. https://doi.org/10.1007/978-3-030-12939-2_9

Ryoo, H. S., & Jang, I.-Y. (2009). Milp approach to pattern generation in logical analysis of data. *Discrete Applied Mathematics, 157*, 749-761. https://doi.org/10.1016/j.dam.2008.07.005

Safizadeh, M. (2001). *Diagnostic des machines dans le plan temps-frequence* [Thèse de doctorat, École Polytechnique de Montréal]. ProQuest Dissertations and Theses Global. https://publications.polymtl.ca/8838/1/1999_Safizadeh.pdf

Salamanca, D., & Yacout, S. (2007). *Condition based maintenance with logical analysis of data.* 7e Congrès International de génie industriel, Québec, Canada.

Salewski, V., Bairlein, F., & Leisler, B. (2003). Niche partitioning of two Palearctic passerine migrants with Afrotropical residents in their West African winter quarters. *Behavioral Ecology*, *14*(4), 493-502. https://doi.org/10.1093/beheco/arg021

Smoljan, N., & Ohran, N. (2015). *Introduction of continuous and structured improvement methodology in sawmill industry: a case study* [Mémoire de maîtrise, Linnaeus University]. DiVA portal https://www.diva-portal.org

Spycher, S., Nendza, M., & Gasteiger, J. (2004). Comparison of different classification methods applied to a mode of toxic action data set. *QSAR Combinatorial Science, 23*, 779-791. https://doi.org/10.1002/qsar.200430877

Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., & Vlahavas, I. (2016). Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning journal, 104*, 55-98. https://doi.org/10.1007/s10994-016-5546-z

Stevanovic, M., Henttonen, P., Koskinen, E., Peräkylä, A., Nieminen von-Wendt, T., Sihvola, E., . . . Sams, M. (2019). Physiological responses to affiliation during conversation: Comparing neurotypical males and males with Asperger syndrome. *Plos One, 14*, 1-9. https://doi.org/10.1371/journal.pone.0222084

Subrahmanyam, M., & Sujatha, C. (1997). Using neural networks for the diagnosis of localized defects in ball bearings. *Tribology international, 30*(10), 739-752. https://doi.org/10.1016/S0301-679X(97)00056-X

Sun, W., Chen, J., & Li, J. (2007). Decision tree and PCA-based fault diagnosis of rotating machinery. *Mechanical Systems and Signal Processing, 21*(3), 1300-1317. https://doi.org/10.1016/j.ymssp.2006.06.010

Taylor, A. B., West, S. G., & Aiken, L. S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational psychological measurement, 66*, 228-239. https://doi.org/10.1177/0013164405278580

Thanh Noi, P., & Kappas, M. (2018). Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors, 18*, 18. https://doi.org/10.3390/s18010018

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., . . . Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*, 520-525. https://doi.org/10.1093/bioinformatics/17.6.520

Tumenjargal, B., Ishiguri, F., Aiso-Sanada, H., Takahashi, Y., Nezu, I., Baasan, B., . . . Yokota, S. (2019). Geographical variations of lumber quality of Larix sibirica naturally grown in five different provenances of Mongolia. *Journal of Wood Science, 65*(1), 1-9. https://doi.org/10.1186/s10086-019-1823-3

van Blokland, J., Nasir, V., Cool, J., Avramidis, S., & Adamopoulos, S. (2021). Machine learning-based prediction of surface checks and bending properties in weathered thermally modified timber. *Construction and Building Materials, 307*, 124996.

Warn, A., & Matthews, P. (1984). Calculation of the compliance of discharges with emission standards. *Water Science Technology, 16*, 183-196. https://doi.org/10.2166/wst.1984.0131

Wu, Z., & Lian, G. (19-24 juillet 2020). *A novel dynamically adjusted regressor chain for taxi demand prediction.* 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK. https://doi.org/10.1109/IJCNN48605.2020.9207160

Xu, Z., Li, Y., Wang, Z., & Xuan, J. (2016). A novel clustering method combining ART with Yu's norm for fault diagnosis of bearings. *Shock and Vibration, 2016*. https://doi.org/10.1155/2016/5468716

Yan, X., & Su, X. (2009). *Linear regression analysis: theory and computing*. World Scientific.

Young, T. M., Bond, B. H., & Wiedenbeck, J. (2007). Implementation of a real-time statistical process control system in hardwood sawmills. *Forest Products Journal, 57*(9), 54-62. https://pubag.nal.usda.gov/catalog/5892

Zhang, Q., Safford, M., Ottenweller, J., Hawley, G., Repke, D., Burgess, J. F., . . . Pogach, L. M. (2000). Performance status of health care facilities changes with risk adjustment of HbA1c. *Diabetes Care, 23*, 919-927. https://doi.org/10.2337/diacare.23.7.919

Zolotarev, F., Eerola, T., Lensu, L., Kälviäinen, H., Haario, H., Heikkinen, J., & Kauppi, T. (2019). *Timber tracing with multimodal encoder-decoder networks* International Conference on Computer Analysis of Images and Patterns, Salerno, Italy. https://doi.org/10.1007/978-3-030-29891-3_30

ANNEXE A DESCRIPTION STATISTIQUE DES INDICATEURS

Tableau A.1 Description statistique des indicateurs

Nom de la variable	Moyenne	Écart-type	Min	Max	Catégories
Nombre billes 25po	396	156	0	582	
Nombre billes 17po	19186	6969	3013	24521	
Nombre pièces classées	84720	30442	17027	109479	
Volume pièces classées	673.52	241.60	116.61	864.03	
Volume pièces 1x3 [Mpmp]	28.99	10.73	4.37	41.35	
Volume pièces 1x4 [Mpmp]	36.22	13.00	5.63	48.95	
Volume pièces 1x6 [Mpmp]	9.33	3.87	0.75	17.57	
Volume pièces 2x3 [mpmp]	11.51	6.85	0.46	38.67	
Volume pièces 2x4 [Mpmp]	193.28	72.36	21.05	295.85	
Volume pièces 2x6 [Mpmp]	231.38	89.78	30.18	365.51	
Volume pièces 2x8 [Mpmp]	69.12	35.28	1.92	133.27	
Volume pièces 2x10 [Mpmp]	5.58	2.98	0	14.23	
Volume pièces 3x3 [Mpmp]	14.91	7.08	0	26.59	
Corrective			0	1	{0,1}
Anticipée			0	1	{0,1}
Préventive			0	3	{0,1,2,3}
Nombre pièces délignées	1331	571	116	2352	
Nombre pièces déchiquetées	2077	786	368	3288	
Nombre pièces rejetées	77	84	5	561	

ANNEXE B MATRICE DE CORRÉLATION DES DONNÉES

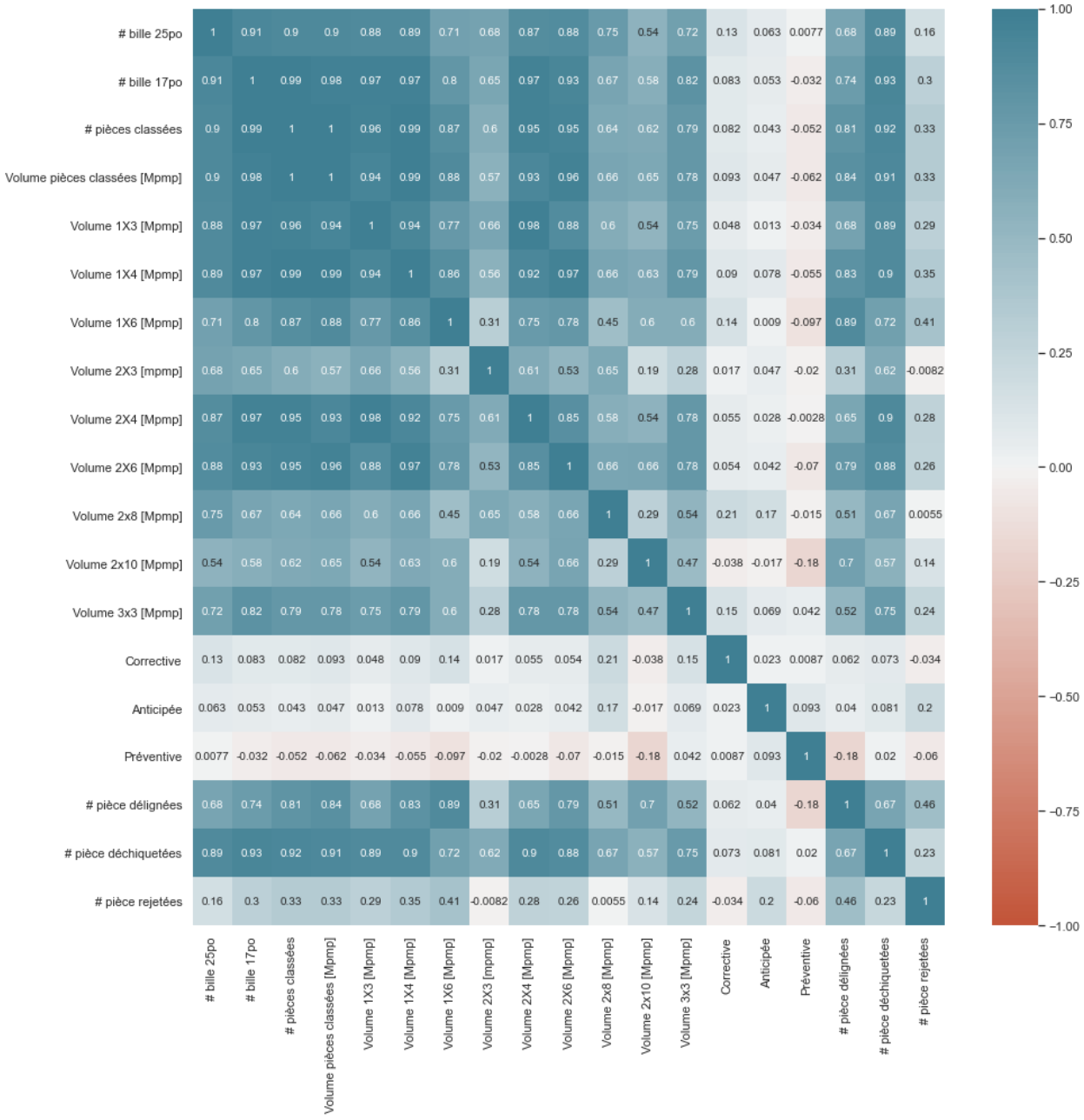


Figure B.1 Matrice de corrélation des données brutes

ANNEXE C MATRICE DE CORRÉLATION DES DONNÉES TRANSFORMÉES

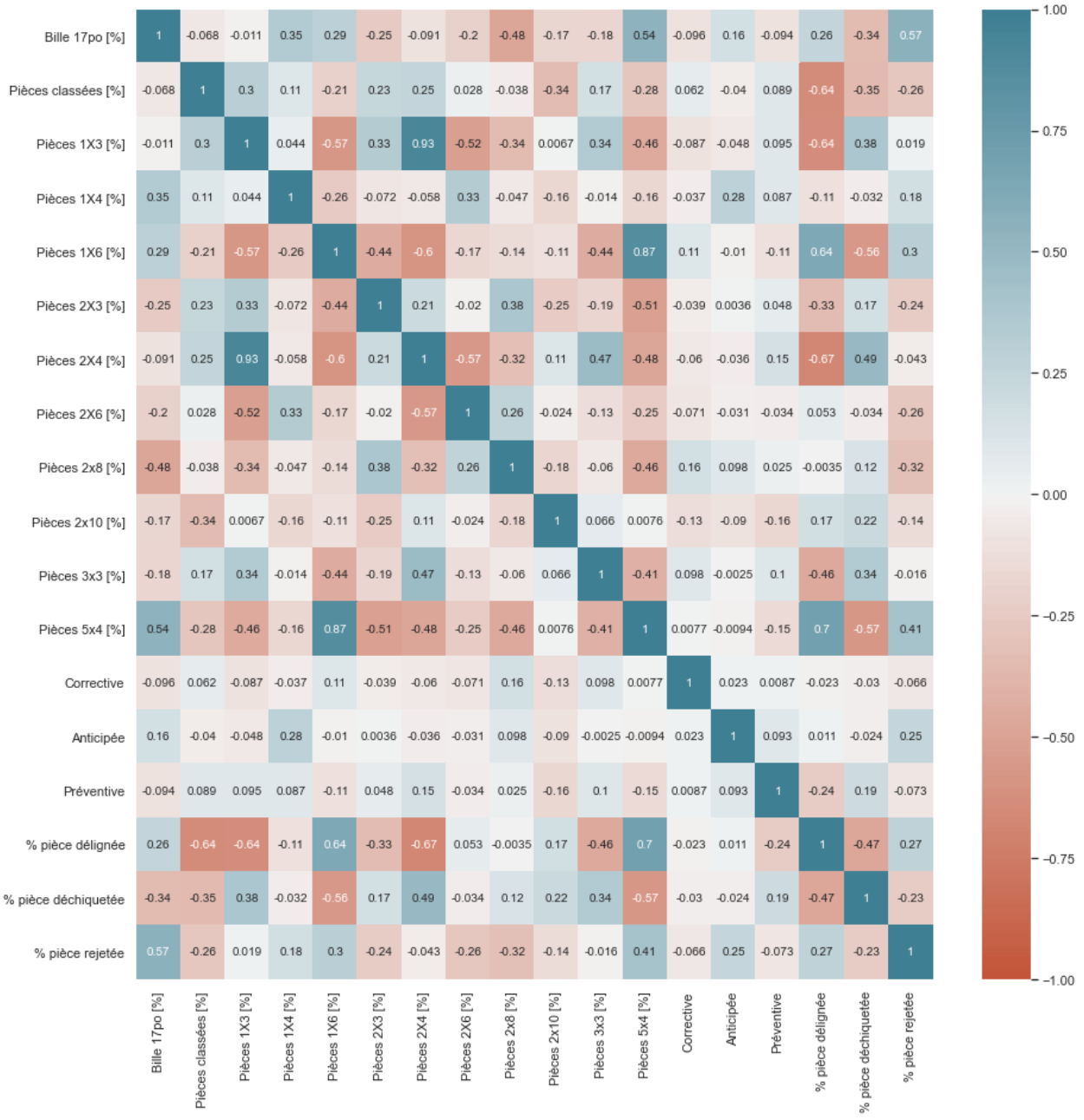


Figure C.2 Matrice de corrélation des données transformées