Titre: Title:	Développement d'un outil d'amélioration des performances industrielles basé sur les données : application dans l'industrie minière
Auteur: Author:	Martin Cuenot
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Cuenot, M. (2025). Développement d'un outil d'amélioration des performances industrielles basé sur les données : application dans l'industrie minière [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/70009/

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Développement d'un outil d'amélioration des performances industrielles basé sur les données : application dans l'industrie minière

MARTIN CUENOT
Département de mathématiques et de génie industriel

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

Octobre 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Développement d'un outil d'amélioration des performances industrielles basé sur les données : application dans l'industrie minière

présenté par **Martin CUENOT**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Camélia DADOUCHI, présidente
Bruno AGARD, membre et directeur de recherche
Michel GAMACHE, membre et codirecteur de recherche
Antoine-Souheil TAHAN, membre et codirecteur de recherche
Ambre DUPUIS, membre

REMERCIEMENTS

Je remercie les membres du jury pour leurs commentaires et leurs critiques constructives sur mon travail, qui m'ont permis d'améliorer tant le fond que la forme de ce document.

Je remercie particulièrement mes directeurs de recherche qui m'ont accompagné et guidé tout au long du projet. Merci Bruno, Michel et Antoine pour votre accueil, vos conseils et vos discussions lors de nos réunions régulières, c'était un réel plaisir de travailler avec vous.

Je remercie également les étudiants du département de Génie industriel et les membres du Laboratoire en Intelligence de Données avec qui j'ai eu le plaisir de partager un bureau et de très nombreux moments conviviaux qui ont participé à rendre mon environnement de travail agréable.

Je remercie le partenaire industriel et en particulier Millan Herce avec qui j'ai eu plaisir à échanger lors de nos différentes rencontres et qui a rendu possible ce projet.

Je remercie également tous les fonds de recherche qui ont participé directement ou indirectement au financement de ma recherche.

Je remercie enfin l'école Polytechnique Montréal, l'INSA de Lyon et leur partenariat. J'ai pu, grâce à cela, participer au programme de double diplôme et réaliser ma maîtrise sur un autre continent. J'ai ainsi pu découvrir une nouvelle culture et de nouvelles façons de travailler, et cela m'offre de nombreuses opportunités pour le futur.

RÉSUMÉ

Dans le contexte de l'Industrie 4.0, de nombreuses données peuvent être acquises sur le fonctionnement de machines ou de véhicules. Ces données présentent un grand nombre d'informations et l'obtention de ces informations est un enjeu important. Dans de nombreux secteurs d'activité, et plus particulièrement dans les entreprises minières, l'amélioration de la productivité et la réduction des coûts sont actuellement des priorités. L'exploitation des données pour la surveillance des activités, l'identification des leviers de performance et l'optimisation des opérations s'avère ainsi essentielle. Le projet présenté dans ce mémoire vise à développer un outil d'amélioration des performances industrielles en se basant sur des données de fonctionnement de machines lors de la réalisation répétée d'un processus.

Basée sur une méthodologie générale de valorisation de données largement utilisée dans l'industrie (CRISP-DM), nous développons un outil qui vise à fournir un cadre de comparaison et d'analyse des performances d'une activité industrielle et à détecter des anomalies de performances. Dans notre contexte, à partir de données de fonctionnement de différentes machines lors d'un même processus, des mesures de performances permettant de caractériser l'activité étudiée sont collectées et analysées. Ces indicateurs spécifiques permettant d'agréger différentes informations à partir des données sont calculés et projetés dans un espace d'analyse isoprobabiliste. L'analyse de la position des points dans l'espace de projection permet de comparer les performances de l'activité et ainsi de prendre des décisions pour améliorer les opérations. Les premiers bénéficiaires de l'outil proposé seront les utilisateurs opérationnels qui ont besoin de comprendre les leviers de la performance, de détecter des comportements anormaux et d'adapter leurs actions en conséquence.

Plus précisément, un ensemble de référence d'itérations de l'activité étudiée est choisi en fonction des besoins d'analyse. Cet ensemble de départ permet de créer un espace de projection pour l'ensemble des itérations à traiter afin de visualiser et d'analyser leur comportement par rapport à l'ensemble de référence choisi. L'ensemble de référence peut dépendre d'un contexte temporel ou bien d'un paramètre de départ de l'activité. La création de l'espace de projection met en jeu une normalisation via transformation quantile et une décorrélation des données grâce à la décomposition de Cholesky. L'espace de projection obtenu est centré sur $\mathbf{0}$ qui représente l'itération moyenne de l'ensemble de référence. Plus une itération analysée apparaît projetée loin du centre de l'espace de projection, plus elle se distingue du groupe de référence. La position d'une itération dans l'espace projeté indique également comment ses performances se situent par rapport à l'ensemble de référence. De par la construction

de l'espace, les distances mesurées correspondent à des seuils statistiques. Cela permet de mettre en place des limites de détection d'anomalies.

La méthode est appliquée à un cas industriel réel, dans le cas de montée de minerai réalisée par des camions en environnement minier souterrain. Cela permet de développer un outil à destination du gestionnaire de flotte afin de mieux comprendre l'influence de certains paramètres sur les performances des camions et d'en suivre l'évolution dans le temps, afin d'agir en cas de dérive constatée.

La mise en place de l'outil dans un contexte opérationnel minier permet de fournir des recommandations aux conducteurs des camions et au gestionnaire de la flotte. L'anticipation de défaillance en surveillant les baisses de performances semble également possible mais devra être confirmée dans de futures recherches avec des jeux de données plus riches. D'autres industries et activités pourraient bénéficier de la mise en place de la méthodologie proposée afin de valoriser les données récoltées et de mettre en place un outil de surveillance et d'amélioration des performances adaptable et interprétable.

ABSTRACT

In the context of Industry 4.0, a substantial amount of data can be collected on the operation of machines or vehicles. As the data set is voluminous, the challenge of utilising it effectively is considerable. This is particularly evident in numerous sectors of activity, with a particular emphasis on mining companies, where enhancing productivity and reducing costs are identified as key priorities. The utilisation of data for the purpose of monitoring activities, identifying performance indicators, and enhancing operational efficiency is, therefore, imperative. The project presented in this thesis aims to develop a tool for improving industrial performance based on machine operating data during the repeated execution of a process.

The tool is based on a general data valuation methodology that has been widely adopted within industry (CRISP-DM). In the context of this study, performance measurements that characterise the activity under investigation are collected and analysed from the operating data of different machines during the same process. These specific indicators, which facilitate the aggregation of disparate pieces of information from the data, are calculated and projected into an isoprobabilistic analysis space. The analysis of the position of points in the projection space facilitates a comparison of the performance of the activity, thus enabling the formulation of decisions to enhance operational efficiency. The intended primary beneficiaries of the proposed tool are operational users who require an understanding of the factors influencing performance, as well as the capability to detect anomalous behaviour and adapt their actions accordingly.

In more detail, a reference set of iterations of the activity under study is selected according to the analysis requirements. The initial set is employed to establish a projection space for all iterations to be processed, enabling the visualization and analysis of their behaviour in relation to the designated reference set. The reference set may be contingent upon a temporal context or an initial parameter of the activity. The creation of the projection space involves the implementation of a quantile transformation and the decorrelation of the data using Cholesky decomposition. The resulting projection space is centred on the origin, which represents the average iteration of the reference set. A greater distance from the centre of the projection space indicates a greater divergence from the reference group. The position of an iteration in the projected space also provides an indication of how its performance compares to the reference set. It is important to note that, due to the nature of the constructed space, the distances measured correspond to statistical thresholds. This facilitates the establishment of limits for the purpose of anomaly detection.

The method is applied to a real industrial case study, in the context of ore transport by trucks in an underground mining environment. This facilitates the development of a tool for fleet managers to enhance their comprehension of the impact of specific parameters on truck performance and to monitor alterations over time, enabling the implementation of remedial actions if any deviations are identified.

The implementation of the tool in an operational mining context makes it possible to provide recommendations to truck drivers and fleet managers. It is also apparent that the monitoring of performance declines may allow for the anticipation of failures; however, this hypothesis requires confirmation through the utilisation of more comprehensive data sets in subsequent research. The proposed methodology could be implemented in other industries and activities to leverage the data collected and establish an adaptable and interpretable tool for monitoring and enhancing performance.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

Acronyme	**Définition**
ANCOVA	Analyse de covariance (*Analysis of Covariance*)
ANOVA	Analyse de la variance (*Analysis of Variance*)
CBM	Maintenance conditionnelle (*Condition-Based Monitoring*)
CRISP-DM	Processus standard inter-industriel pour l'exploration de données (*CRoss Industry Standard Process for Data Mining*)
DL	Apprentissage profond (*Deep Learning*)
KDD	Découverte de connaissances dans les bases de données (*Knowledge Discovery in Databases*)
MANOVA	Analyse de la variance multivariée (*Multivariate Analysis of Variance*)
ML	Apprentissage automatique (*Machine Learning*)
MSPC	Contrôle statistique multivarié des procédés (*Multivariate Statistical Process Control*)
PCA	Analyse en composantes principales (*Principal Component Analysis*)
SEMMA	Échantillonner-Explorer-Modifier-Modéliser-Évaluer (*Sample-Explore-Modify-Model-Assess*)
SPC	Contrôle statistique des procédés (*Statistical Process Control*)
VAE	Auto-encodeur variationnel (*Variational Autoencoder*)
ZCA	Analyse en composantes à phase nulle (*Zero-phase Component Analysis*)

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

La valorisation de données industrielles est, aujourd'hui, un enjeu important pour de nombreuses firmes. La collecte et le stockage de données rendus possibles grâce aux technologies du numérique mettent à disposition des quantités importantes d'informations. Elles peuvent être utilisées pour comprendre les processus et l'utilisation de machines et de matériel, pour appuyer des décisions et pour améliorer l'efficacité d'une organisation. Ce mémoire présente les travaux réalisés avec un partenaire de l'industrie minière, dans le but de valoriser des données obtenues lors du fonctionnement de camions en environnement souterrain.

1.1 Industrie 4.0 et transition numérique

Depuis 2011, à la suite d'une initiative allemande, le terme 'Industrie 4.0' [Drath and Horch, 2014] évoque la quatrième révolution industrielle. Elle succède à la mécanisation et à la machine à vapeur à la fin du XVIII^{e} siècle, à l'électrification et à la production de masse au début du XX^{e} siècle, puis à l'automatisation rendue possible par l'électronique et l'informatique dans les années 1970.

L'Industrie 4.0 met en jeu différentes technologies et savoir-faire numériques pour permettre aux systèmes physiques de communiquer entre eux et avec les humains dans le but de coopérer et de décentraliser la prise de décision [Danjou *et al.*, 2017]. L'intelligence artificielle (IA), les données massives (*Big Data*), l'infonuagique (*Cloud-Computing*), l'Internet des objets (IoT), la cybersécurité, les jumeaux numériques, la réalité augmentée, les systèmes cyber-physiques et les machines autonomes sont autant de leviers [Rüßman *et al.*, 2015] qui permettent aux entreprises de mettre en place de nouvelles stratégies pour améliorer leurs processus, leurs produits et leurs services. Les systèmes acquièrent de nouvelles capacités, qu'il est possible de distinguer en quatre groupes [Danjou *et al.*, 2017] [Porter and Heppelmann, 2014].

- **Surveillance :** les données obtenues grâce aux capteurs et à la connectivité du système permettent d'en surveiller les performances, l'état et l'environnement extérieur afin d'aider à la prise de décision.
- **Contrôle :** des programmes ou des algorithmes exécutent des actions simples pour adapter le comportement du système en fonction de son état ou de son environnement.
- **Optimisation :** à partir d'un historique et d'algorithmes complexes, le système adapte ses paramètres et son fonctionnement pour optimiser ses performances.
- **Autonomie :** la surveillance, le contrôle et l'optimisation fonctionnent ensemble pour rendre le système autonome. L'environnement, les besoins et les préférences de l'uti-

lisateur sont pris en compte afin de garantir un fonctionnement optimal du système sans intervention extérieur.

1.2 Mines 4.0 et enjeux actuels du secteur minier

La plupart des industries se transforment avec l'Industrie 4.0. Dans le secteur minier, le terme 'Mining 4.0' [Lööw *et al.*, 2019] est parfois employé. Les principales avancées liées aux technologies numériques concernent la sécurité, la productivité et la protection de l'environnement. Les transformations numériques du secteur minier se déroulent dans un contexte contraint et complexe. Les pressions économiques sont fortes, les prix des métaux fluctuent et les marchés ne sont pas stables. Aussi, les coûts de production sont en hausse (+30% depuis 2019) et les réserves minérales deviennent moins riches et plus difficiles d'accès. À l'échelle mondiale, les profits des années 2023 et 2024 ont baissé de plus de 40% [PWC, 2024]. Ces difficultés poussent le secteur à se concentrer sur l'optimisation de la productivité des opérations et la réduction des coûts.

1.3 Mines au Canada et partenariat industriel

Au Canada, l'industrie minière est importante. Elle représente près de 8 % du PIB en 2022. Le secteur emploie près de 700 000 personnes directement ou indirectement et contribue au développement des infrastructures dans les régions isolées du Nord. Les minières québécoises intègrent de nombreuses technologies de l'Industrie 4.0 et collaborent avec les centres de recherche universitaires pour mettre en place des projets innovants afin d'améliorer leurs activités. Le partenaire industriel de ce projet est une entreprise minière qui exploite des mines souterraines au nord du Québec. Plusieurs projets liés à l'Industrie 4.0 y sont mis en place, tels que des systèmes de détection de proximité entre engins et opérateurs pour assurer la sécurité, ou des véhicules autonomes et leur pilotage à distance depuis la surface. Les véhicules circulant dans la mine sont équipés de nombreux capteurs permettant la récolte de données de fonctionnement.

L'activité principale du partenaire consiste à extraire et à augmenter la concentration du minerai issu d'une exploitation souterraine. Cela implique de creuser des puits et des galeries, de forer et de dynamiter la roche, puis de transporter le minerai jusqu'à la surface. Le minerai est ensuite broyé et traité, dans une usine spécialisée appelée concentrateur située à proximité de la mine, afin de séparer les éléments utiles du reste de la roche et d'obtenir un produit plus pur. Le partenaire industriel cherche à augmenter la productivité (diminuer le coût de ses opérations) afin de pouvoir obtenir du minerai avec un coût de revient par tonne de roche

extraite plus faible.

Les technologies de l'Industrie 4.0 permettent de rassembler de nombreuses données à propos des processus et activités industrielles. Bien que de plus en plus répandues, l'analyse et la valorisation de ces données ne sont pas systématiques et ne sont pas forcément destinées à une utilisation opérationnelle directe. Elles peuvent pourtant permettre d'améliorer les performances des opérations pour les entreprises et le secteur minier n'est pas le seul concerné ; les performances d'autres industries pourraient également bénéficier de méthodes de valorisation de données de fonctionnement. Nous proposons ainsi de développer un outil pour améliorer la performance d'une activité industrielle basé sur les données et d'appliquer la mise en place d'un tel outil dans le contexte minier. Cet outil est destiné à être utilisé par des utilisateurs opérationnels. L'utilisation de l'outil dans le cadre d'une activité industrielle minière peut permettre de mieux comprendre les leviers de la performance, de détecter des comportements anormaux et d'adapter les opérations pour contribuer à augmenter la productivité et à réduire les coûts d'opérations.

Dans la suite du mémoire, le Chapitre 2 présentera une revue de littérature associée à la problématique, le Chapitre 3 détaillera les objectifs de recherche, le Chapitre 4 décrira la méthodologie. Dans le Chapitre 5, le cas d'étude sera présenté et la méthodologie sera appliquée à une activité industrielle minière. Le Chapitre 6 présentera les conclusions du projet de recherche et les pistes pour de futurs travaux liés.

CHAPITRE 2 REVUE DE LITTÉRATURE

Comme détaillé dans l'introduction (Chapitre 1), l'objectif de recherche consiste à utiliser des données de fonctionnement pour proposer un outil de support à la performance à destination des utilisateurs opérationnels. La revue de littérature présente d'abord une méthode d'exploration de données et de découverte de connaissances dans la Section 2.1. Ensuite, elle explore comment les performances peuvent être supportées dans l'industrie (Section 2.2) ainsi que différentes approches statistiques et de traitement de données. Enfin, le lien entre anomalies de performances et maintenance est expliqué dans la Section 2.3.

2.1 Méthodologie d'exploration et de valorisation de données

L'accélération de l'accessibilité aux données, des capacités de stockage et de la puissance de calcul popularise la science des données au niveau des organisations depuis le début des années 2000 [Ahmad *et al.*, 2022]. La science des données se concentre sur l'extraction de connaissances à partir des données ou comment transformer les données brutes en informations utilisables et valorisables pour les organisations [Provost and Fawcett, 2013]. Le processus de découverte d'informations à partir des données est complexe. Il nécessite des compétences variées, tant au niveau de la compréhension du contexte et du problème industriel qu'au niveau du traitement des données, de la mise en production et du déploiement d'applications [Wirth and Hipp, 2000]. Pour assurer un processus robuste et efficace, plusieurs méthodologies standards ont été proposées telles que :

- la méthode Découverte de connaissances dans les bases de données (*Knowledge Discovery in Databases*) (KDD) [Fayyad *et al.*, 1996],
- la méthode Échantillonner-Explorer-Modifier-Modéliser-Évaluer (*Sample-Explore-Modify-Model-Assess*) (SEMMA) [SAS Institute Inc., 2003],
- le Processus standard inter-industries pour l'exploration de données (*CRoss Industry Standard Process for Data Mining*) (CRISP-DM) [Chapman *et al.*, 2000, Wirth and Hipp, 2000]

Dans une revue de littérature systématique réalisée en 2023 [Shimaoka *et al.*, 2024], différentes études basées sur la méthodologie CRISP-DM sont explorées. Il apparaît que la méthodologie développée au début des années 2000 par un consortium européen (IBM SPSS, Mercedes-Benz Group, OHRA, University of Stuttgart) est celle qui est la plus largement utilisée, dans sa forme initiale ou avec des adaptations propres au contexte de chaque organisation. La méthode CRISP-DM semble pertinente pour explorer et valoriser des données industrielles.

Les différentes étapes qui la composent sont décrites dans les paragraphes suivants à partir de ces articles : [Chapman *et al.*, 2000, Wirth and Hipp, 2000, Shimaoka *et al.*, 2024]. Elles sont également décrites dans la Figure 2.1.

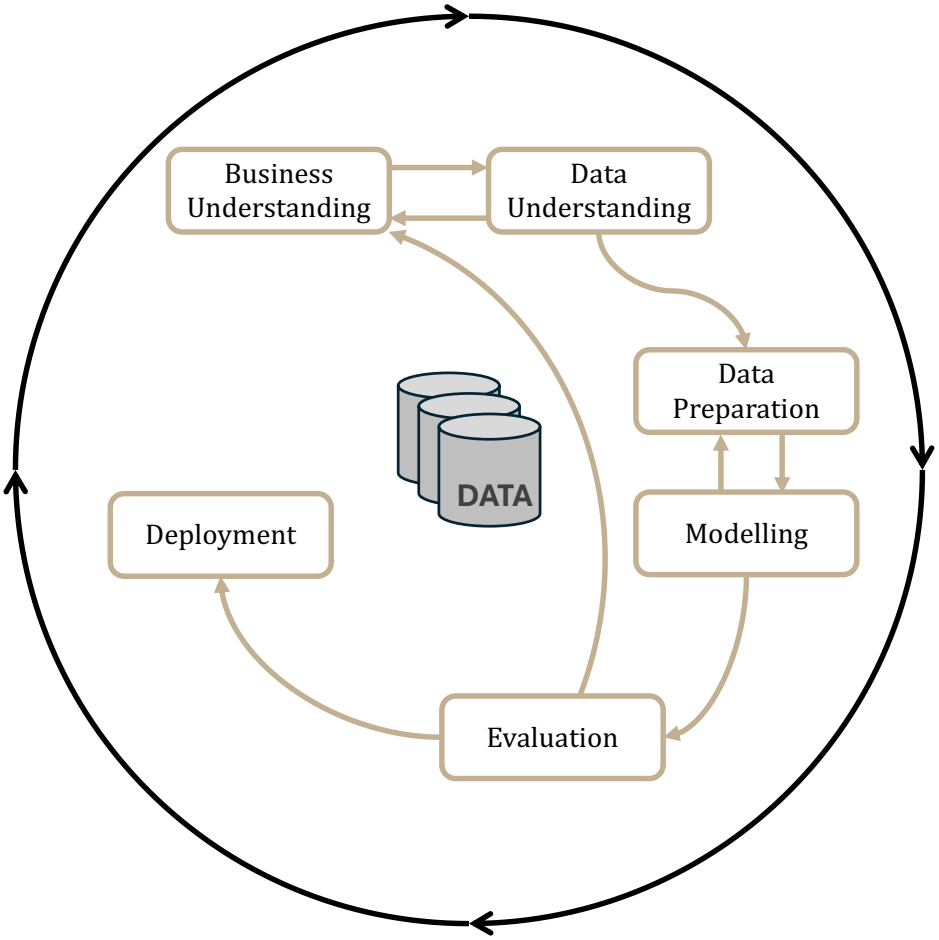

FIGURE 2.1 Cycle du processus CRISP-DM, adapté de [Wirth and Hipp, 2000]

Le processus proposé est conçu pour être appliqué dans n'importe quelle industrie selon une structure définie et reproductible. CRISP-DM décrit le cycle de vie d'un projet d'exploration et de valorisation de données selon les six phases suivantes.

— **Compréhension du contexte industriel :** il s'agit de comprendre les objectifs du coté de l'industrie ou de l'organisation. Le contexte, les spécificités et les demandes de l'entreprise ainsi que l'avis des experts du domaine sont utilisés pour définir un ou des objectifs industriels concrets et pour évaluer la situation actuelle par le biais de ces objectifs. Les objectifs industriels sont ensuite traduits en objectifs d'exploration de données qui permettront de répondre au besoin de l'organisation. Cette phase est importante car elle permet de s'assurer que le projet d'exploration et de valorisation

de données répond à un ou plusieurs problèmes concrets de l'organisation.

— **Compréhension des données :** il s'agit d'accéder aux données et de les explorer pour comprendre comment elles peuvent être utilisées. L'inventaire des données disponibles est réalisé, en décrivant la qualité et la quantité des données. Une phase d'exploration préliminaire à partir d'études et de visualisations simples permet de formuler des hypothèses et d'envisager les potentielles transformations à appliquer durant la phase suivante.

— **Préparation des données :** il s'agit de sélectionner et de transformer les données vers un format pouvant être utilisé lors de la phase de **Modélisation**. Cette phase est souvent la plus longue lors d'un projet de valorisation de données. Les phases de **Compréhension du contexte industriel** et de **Compréhension des données** permettent de cibler les modifications à effectuer et en conditionnent l'efficacité. Cette préparation peut impliquer la sélection des données pertinentes, le nettoyage des données, la construction de nouveaux attributs, l'intégration de différentes sources de données, le formatage, l'échantillonnage,...

— **Modélisation :** il s'agit d'utiliser les données préparées pour répondre à l'objectif d'exploration de données grâce à un ou plusieurs modèles. Plusieurs techniques de modélisations peuvent être utilisées et peuvent permettre de répondre à un même problème. La construction de modèles peut impliquer des tests et la recherche de paramètres optimaux.

— **Évaluation :** il s'agit d'évaluer les résultats de la modélisation par rapport aux objectifs d'exploration. Les résultats peuvent être un modèle final ou bien des conclusions issues du processus d'exploration. Il faut également décider si l'exploration de données permet de répondre à l'objectif industriel et si le projet peut être déployé à plus grande échelle ou s'il faut reprendre le processus et continuer l'exploration.

— **Déploiement :** Il s'agit, une fois les résultats jugés satisfaisants, de mettre à disposition le modèle ou les conclusions issues de l'analyse auprès des utilisateurs finaux. Cette phase de déploiement, qui consiste à intégrer le modèle ou le processus complet de découverte de connaissances dans l'environnement opérationnel, est généralement assurée par l'organisation ou les services informatiques et sort du périmètre direct de l'analyste de données.

2.2 Amélioration de la performance

Dans la première partie de la revue de littérature, nous décrivons une approche générale pour valoriser des données. Nous nous intéressons maintenant à l'utilisation de ces données pour supporter la performance industrielle. Comme défini par l'office québécois de la langue française, la performance correspond au résultat d'une activité, mesuré à l'aide d'indicateurs. Dans l'industrie, améliorer les performances en vue d'atteindre un certain niveau souhaité implique la mise en place de processus dédiés [Deming, 1982]. La norme ISO 9001 :2015 [International Organization for Standardization, 2015] décrit les étapes de ce processus d'amélioration continue et insiste sur deux besoins : réaliser un diagnostic de la situation existante et analyser les résultats des changements mis en place. Les décisionnaires doivent ainsi surveiller la performance, l'analyser et agir en conséquence. Ils ont besoin d'obtenir les informations nécessaires pour visualiser et expliquer les performances actuelles.

Pour atteindre l'objectif de ce projet, il convient de s'intéresser à l'analytique de données lors de la partie **Modélisation** de la méthodologie CRISP-DM. L'objectif n'est pas de résoudre un problème spécifique ou de réaliser une tâche en se basant sur des données historiques comme il est possible de le faire avec les méthodes de l'Apprentissage automatique (*Machine Learning*) (ML). Il consiste plutôt à enregistrer et afficher les connaissances extraites des données pour permettre à l'utilisateur d'interpréter les données et de se servir de l'outil pour soutenir des décisions [Mannila, 1996] et le processus d'amélioration continue.

2.2.1 Contrôle des processus

L'idée d'améliorer la performance est assez vague et peut faire référence à plusieurs thèmes. Dans la littérature, le Contrôle statistique des procédés (*Statistical Process Control*) (SPC) permet de surveiller les performances d'un processus industriel ou d'une activité pour s'assurer de rester dans un état de contrôle d'un point de vue statistique [Wetherill and Brown, 1991]. L'une des applications du SPC consiste à mettre en œuvre des cartes de contrôle. Ce sont des outils statistiques utilisés pour surveiller la stabilité d'un processus industriel au fil du temps. Elles représentent graphiquement une caractéristique mesurée vis-à-vis d'une valeur cible, généralement la moyenne, et comportent des limites de contrôle statistique autour de cette moyenne. Tant que les mesures se situent à l'intérieur de ces limites, le processus est jugé stable et sous contrôle. Lorsque des valeurs dépassent ces limites ou affichent des motifs inhabituels, cela peut signaler un déséquilibre dans le contrôle et demander une analyse des anomalies. Le suivi d'un processus par SPC permet de détecter des comportements anormaux et amène à un diagnostic pour comprendre les causes et conséquences d'un tel comportement

afin d'améliorer le processus.

Le Contrôle statistique multivarié des procédés (*Multivariate Statistical Process Control*) (MSPC) repose sur l'analyse simultanée de plusieurs variables interdépendantes, permettant ainsi de capter la structure globale des variations du processus [Martin *et al.*, 1998]. À la différence du SPC univarié qui considère chaque variable séparément, le MSPC permet d'examiner la direction des variations, c'est-à-dire la façon dont les variables évoluent ensemble, ce qui facilite l'interprétation et le diagnostic des anomalies. De plus, en intégrant les informations de nombreuses variables, le MSPC facilite la découverte des signaux faibles qui sont souvent masqués par le bruit lorsqu'on n'analyse qu'une seule variable à la fois [MacGregor, 1994].

Le SPC et le MSPC sont bien connus dans l'industrie des procédés chimiques et manufacturière depuis le milieu du XXe siècle et ont vu leurs pratiques évoluer avec l'essor du traitement numérique des données pendant les années 1990 et 2000. Ce sont des sujets bien documentés dans la littérature avec de nombreuses publications actuelles qui s'intéressent au développement de méthodes non paramétriques, où les distributions des variables suivies sont inconnues [Xue and Qiu, 2021, Mukherjee and Marozzi, 2022].

Projection de données

Dans le cadre du MSPC et plus globalement pour le contrôle de processus et l'analyse de données, en présence de plusieurs variables interdépendantes, les techniques de projection des données vers un espace latent sont courantes.

L'Analyse en composantes principales (*Principal Component Analysis*) (PCA) est une méthode statistique classique de réduction de dimension utilisée pour représenter des données multidimensionnelles dans un espace de dimension réduite tout en conservant un maximum d'information. Elle repose sur une transformation linéaire des variables d'origine vers des composantes principales. Celles-ci sont non corrélées (orthogonales entre elles) et classées selon la variance expliquée. Les premières composantes principales capturent la plus grande part de la variance totale des données, ce qui permet de résumer efficacement l'information contenue dans les données brutes. La PCA est très utilisée dans le cadre du MSPC.

L'analyse discriminante est une autre méthode qui projette les données vers un espace latent optimisé pour la séparation entre plusieurs groupes. Dans ce cas, l'objectif est de choisir les combinaisons linéaires de variables explicatives qui maximisent la séparation entre les groupes tout en minimisant la variance intra-groupe. Cette technique permet de visualiser les différences entre groupes dans un espace de dimension réduite et d'identifier les variables

qui contribuent le plus à leur discrimination.

Les espaces de projections sont ainsi créés pour répondre au besoin d'analyse spécifique à la problématique étudiée. Un autre type d'espace intéressant à mentionner sont les espaces iso-probabilistes. Dans ces espaces, les variables aléatoires sont transformées de sorte à être indépendantes et à suivre une distribution connue. Les distances euclidiennes correspondent alors à des variations de probabilités et cela permet une identification des anomalies plus simple. La transformation Nataf est un exemple de transformation iso-probabiliste [Lebrun and Dutfoy, 2009].

Les transformations vers des espaces latents permettant de mieux représenter les données peuvent également être réalisées avec des réseaux de neurones particuliers. Par exemple, l'Auto-encodeur variationnel (*Variational Autoencoder*) (VAE) est un type de réseau de neurones adapté. À travers plusieurs couches neuronales, les données sont transformées en un vecteur à dimension réduite et un réseau symétrique décode ce vecteur pour reconstituer les données initiales. L'apprentissage de ce réseau vise à optimiser l'encodage-décodage. La représentation réduite des données dans un espace latent peut être utilisée pour détecter des anomalies, des groupes de données et des structures cachées. [Oliveira-Filho *et al.*, 2024] présentent par exemple l'utilisation d'un espace latent issu d'un VAE, associé à une transformation iso-probabiliste pour suivre les conditions de fonctionnement d'un moteur de propulsion de la NASA.

2.2.2 Analyse et comparaison des performances

Le support à la performance évoque également l'analyse et la compréhension des performances. Une idée consiste à étudier l'effet du contexte sur les performances.

Analyse de variance

L'ANOVA développée par Fisher [Fisher, 1925] et formalisée dans un cadre industriel par Montgomery [Montgomery, 2017], permet d'évaluer si les moyennes de plusieurs groupes diffèrent significativement. Ainsi, il est possible de déterminer statistiquement l'effet d'une variable qualitative sur une variable mesurée quantitative.

En pratique, l'ANOVA compare la variabilité présente entre les groupes par rapport à celle observée à l'intérieur des groupes, sous l'hypothèse nulle selon laquelle toutes les moyennes de groupe sont égales. Si la variabilité entre les groupes dépasse significativement celle présente au sein des groupes, l'hypothèse nulle est rejetée, indiquant qu'au moins une des moyennes n'est pas équivalente aux autres.

L'ANOVA permet d'analyser l'effet d'un ou plusieurs facteurs qualitatifs sur une variable dépendante quantitative, en testant l'égalité des moyennes entre groupes. Lorsque plusieurs variables dépendantes quantitatives sont étudiées simultanément, l'Analyse de la variance multivariée (*Multivariate Analysis of Variance*) (MANOVA) est utilisée afin de tenir compte de la corrélation entre ces variables. L'Analyse de covariance (*Analysis of Covariance*) (ANCOVA) permet d'évaluer l'effet de facteurs qualitatifs tout en ajustant les résultats en fonction de facteurs quantitatifs.

L'ANOVA repose sur l'hypothèse que la variable quantitative suit une distribution normale. Une simple transformation peut ajuster la variable à cette distribution. Il est également important de vérifier l'homoscédasticité, c'est-à-dire que les variances entre les groupes soient identiques.

Lorsque l'ANOVA révèle une différence significative entre les moyennes de plusieurs groupes, il est nécessaire de déterminer précisément quels groupes diffèrent entre eux. Des tests post-hoc sont utilisés, parmi lesquels le test de Tukey est l'un des plus couramment employés. Le test de Tukey permet de réaliser des comparaisons multiples deux à deux entre les moyennes des groupes tout en contrôlant le risque global de fausses détections de différences [Montgomery, 2017].

Ainsi, l'analyse de variance et les méthodes liées permettent d'expliquer les variations d'une variable en fonction d'autres variables explicatives. Les hypothèses de normalité et d'homoscédasticité ainsi que la nécessité d'analyser une variable à la fois sont des contraintes à prendre en compte. Cependant, l'idée de pouvoir montrer une différence de performance en fonction de facteurs est intéressante. L'analyse de variance peut être utilisée lors des phases préliminaires ou pour vérifier des résultats obtenus avec d'autres méthodes.

Distance entre des distributions

Pour analyser et comparer des performances, il est intéressant de comparer différents groupes de données et d'évaluer la différence entre leurs distributions statistiques.

La distance de Mahalanobis [Mahalanobis, 1936] est une distance généralisée prenant en compte les corrélations entre variables (2.1). En tenant compte de la variance, une plus grande importance est accordée aux variables les plus dispersées pour mesurer la similarité. Elle permet de mesurer la distance d'une observation à une distribution selon la formulation suivante.

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \tag{2.1}$$

Entre deux vecteurs aléatoires $\mathbf{x}$ et $\mathbf{y}$ provenant de la même distribution et partageant une matrice de covariance $\mathbf{\Sigma}$, elle peut être définie pour mesurer la similarité entre les deux groupes (2.2).

$$d_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})}. \tag{2.2}$$

D'autres mesures existent pour évaluer la différence entre deux distributions. La divergence de Kullback-Leibler [Kullback and Leibler, 1951] repose sur la théorie de l'information pour mesurer la dissimilarité entre deux distributions.

2.3 Maintenance et anomalies de performances

Améliorer les performances au sens global signifie aussi s'intéresser aux processus connexes et les rendre plus efficaces. La maintenance des équipements est l'un des principaux postes de dépense lors de l'exploitation d'une mine. Selon [Dhillon, 2008] 20% à 35% des coûts d'opérations sont dus à la maintenance. La mise en place de stratégies de l'Industrie 4.0 et l'exploitation des informations acquises grâce aux nouvelles technologies peuvent permettre d'augmenter la disponibilité des équipements et d'optimiser les activités de maintenance pour en améliorer l'efficacité et réduire les dépenses. Cela améliore les performances globales de l'activité et de la structure.

D'après la norme AFNOR NF X60 de 2016 [AFNOR, 2016], la maintenance est définie telle que "l'ensemble des actions techniques, administratives et de management durant le cycle de vie d'un bien destinées à le maintenir ou à le rétablir dans un état dans lequel il peut accomplir la fonction requise". Elle peut être :

- **Corrective**, "exécutée après détection d'une panne et destinée à remettre un bien dans un état dans lequel il peut accomplir une fonction requise".
- **Préventive**, "exécutée à intervalle prédéterminés ou selon des critères prescrits et destinés à réduire la probabilité de défaillance ou la dégradation de fonctionnement d'un bien".

Historiquement, la maintenance dans les mines était réalisée selon des plans systématiques. L'émergence des données et de l'Industrie 4.0 ouvre des possibilités plus efficaces pour programmer les interventions de maintenance pour les équipements. Deux types de maintenance préventive peuvent alors être mis en place :

- La maintenance conditionnelle, qui surveille le fonctionnement ou l'état de la machine pour décider du déclenchement d'interventions.
- La maintenance prévisionnelle ou prédictive, "exécutée suite à une prévision obtenue grâce à une analyse répétée ou à des caractéristiques connues et à une évaluation des

paramètres significatifs de la dégradation du bien".

[Dayo-Olupona *et al.*, 2023] présentent une revue de littérature systématique concernant les approches de maintenance prédictive dans les mines. L'enjeu principal de la maintenance prédictive consiste à anticiper les défaillances et à optimiser la programmation des opérations de maintenance. Les capteurs permettent de récolter de grandes quantités de données provenant de différentes sources (données de fonctionnement, indicateurs opérationnels). Les anomalies dans ces données peuvent représenter des signaux indiquant de potentielles défaillances et l'enjeu consiste à les détecter et à estimer la probabilité qu'elles soient effectivement annonciatrices d'un défaut à venir.

Pour le domaine minier, les approches basées sur les données, mettant en jeu des modèles statistiques ou du ML présentent les meilleurs résultats et sont les plus étudiées. Deux méthodologies principales se distinguent :

- La Maintenance conditionnelle (*Condition-Based Monitoring*) (CBM), qui consiste à surveiller de manière régulière une ou plusieurs conditions de fonctionnement d'un composant, sous-système ou système. La détection de signaux faibles est réalisées grâce à des seuils établis, déclenchant des alertes lors des dépassements. Plusieurs approches appliquées dans le domaine minier sont détaillées dans la revue systématique. Les paramètres principalement utilisés sont des paramètres physiques, spécifiques à un composant ou un sous-système tel que la température, les vibrations, le bruit, les caractéristiques de l'huile ou des signaux électriques. Les méthodes liées au CBM nécessitent un historique de données relativement limité.
- Les méthodes de ML ou d'Apprentissage profond (*Deep Learning*) (DL) exploitent de manière plus poussée les historiques opérationnels et les historiques de maintenance pour estimer la durée de vie restante d'un composant, sous-système ou système. Ces estimations sont plus informatives ou précises qu'une simple alarme. Dans la littérature propre au domaine minier, plusieurs méthodes sont utilisées telles que les Machines à Vecteur Support, les réseaux de neurones classiques ou convolutionnels, les arbres de décisions et les forêts aléatoires [Dayo-Olupona *et al.*, 2023]. Ces approches présentent de très bonnes performances pour prédire les durée de vie restantes et anticiper les défaillances mais nécessitent de très grande quantités de données.

Les approches actuellement utilisées, que ce soit pour du CBM ou lors de l'estimation de durée de vie restante, sont majoritairement axées sur un paramètre mécanique, physique ou chimique spécifique. Dans le domaine minier, ces analyses sont souvent spécifiques à un composant ou un sous-système tel que les freins, le moteur ou la transmission. Ces approches nécessitent une grande quantité de données pour être fiables et sont difficilement généralisables à l'ensemble du système ou à d'autres systèmes similaires.

Compte tenu de ces limitations et de la problématique de support à la performance, l'analyse des performances opérationnelles globales du système, avec le choix d'indicateurs clés, pourrait permettre de détecter des signaux faibles indiquant des déviations mineures susceptibles de devenir de réelles défaillances. L'objectif de la maintenance consiste à s'assurer que le système évolue dans un état de fonctionnement souhaité. Alerter lorsque cet état de performance évolue semble intéressant dans un contexte industriel.

La surveillance d'indicateurs globaux par le biais du CBM permet le développement de modèles plus simples, avec un besoin de données historiques moins important. L'obtention de données concernant la performance globale d'un système est également plus facile que l'acquisition de données physiques ou mécaniques comme les vibrations. Une telle approche serait plus généralisable et adaptée à des contextes industriels variés. Une limite prévisible serait la précision des prédictions en comparaison des méthodes de ML et DL. À terme, avec des historiques importants, les mesures de performances globales pourraient être utilisées comme données d'entrée pour les approches d'estimation de durée de vie restante.

2.4 Opportunités de recherche

À partir de la revue de littérature, il apparaît que l'amélioration des performances (au sens global) d'une activité industrielle nécessite la possibilité de comparer des itérations ou des groupes d'itérations selon différents critères prenant en compte le contexte et les conditions de réalisation. Différencier les comportements normaux d'éventuelles anomalies semble aussi important. Cela peut être réalisé de plusieurs manières différentes avec un besoin en données et en temps de calcul variable. Dans le cas où peu de données sont disponibles et où les résultats doivent pouvoir être obtenus et compris rapidement par des utilisateurs opérationnels, les techniques lourdes d'apprentissage machine ne sont pas adaptées. Dans ces cas, il est nécessaire d'utiliser des outils mathématiques et statistiques répondant à un enjeu à la fois. Un outil plus global et nécessitant peu de données, permettant directement de procéder aux analyses des performances, semble ainsi être un enjeu intéressant sur lequel travailler.La transformation des données vers un espace latent est une idée prometteuse pour permettre les analyses et faciliter l'interprétation pour l'utilisateur.

Il apparaît aussi que les pratiques de maintenance conditionnelle sont basées sur le suivi de paramètres physiques précis relatifs à certains sous-systèmes des machines. Étudier la possibilité de suivre des indicateurs de performance plus globaux pour réaliser des opérations de maintenance semble également être intéressant.

CHAPITRE 3 PROBLÉMATIQUE ET OBJECTIFS DE RECHERCHE

L'introduction (Chapitre 1) montre le besoin de l'industrie, en particulier minière, de disposer d'outils basés sur les données pour augmenter la productivité et réduire les coûts de ses activités. La revue de littérature (Chapitre 2) met en avant la méthode générique CRISP-DM permettant d'extraire des informations pertinentes à partir des données et présente diverses approches pour comparer et analyser des performances. Le lien entre anomalies de performances et maintenance est également souligné. Cependant, peu d'outils globaux, possibles à mettre en place dans des cas industriels, permettent de rassembler et de mettre en valeur les données de performances industrielles pour qu'elles puissent être utilisées pour piloter et optimiser les activités par des utilisateurs opérationnels.

En prenant en compte ces éléments, nous souhaitons adapter et mettre en œuvre la méthodologie CRISP-DM pour développer un outil qui permettra d'analyser et de comparer les performances d'une activité industrielle. Pour cela, nous proposons de mettre en place des indicateurs de performance spécifiques à l'activité étudiée. Nous projetons ensuite ces indicateurs de chaque itération dans un nouvel espace qui facilitera l'analyse visuelle et les comparaisons. La mise en place d'un tel outil pourra également permettre de lever des alertes à destination de la maintenance lors d'apparitions d'anomalies ou de dérives. La recherche se décompose ainsi en deux sous-objectifs :

- **SO1 :** Développer un cadre de comparaison et d'analyse des performances d'une activité industrielle à partir des données.
- **SO2 :** Détecter les anomalies de performances pour d'éventuelles alertes à destination de la maintenance.

L'outil se doit d'être général afin d'être applicable dans des contextes industriels différents. Il est destiné à être intégré dans un contexte opérationnel. Il devra répondre aux exigences suivantes :

- **Interprétable :** les résultats obtenus doivent être compréhensibles par les utilisateurs opérationnels pour appuyer les prises de décisions. Les modèles ou transformations de données complexes (de type "boite noire") ne sont pas souhaitables.
- **Adaptable :** l'outil doit permettre de répondre à différents usages. Il doit, par exemple, permettre de comparer les performances de l'activité selon le contexte dans lequel elle a été réalisée et permettre le suivi des performances dans le temps.
- **Visuel :** la présentation des résultats dans l'outil doit être facilement compréhensible et permettre d'obtenir des conclusions rapides pour l'utilisateur opérationnel. Le support visuel doit permettre de soutenir la prise de décision.

Dans les prochains chapitres, nous développons la méthodologie générale permettant d'arriver à l'outil(Chapitre 4) puis nous l'appliquons à un cas d'étude de l'industrie minière et observons plusieurs manières d'utiliser l'outil mis en place (Chapitre 5).

CHAPITRE 4 MÉTHODOLOGIE PROPOSÉE

Ce chapitre présente la méthodologie suivie, étape par étape, en suivant les recommandations de la méthode CRISP-DM. La Figure 4.1 résume les différentes étapes suivies. La méthode CRISP-DM permet de valoriser des données. Dans ce travail, elle est adaptée pour amener à l'obtention d'un outil utilisable par les utilisateurs opérationnels pour comprendre les performances d'une activité industrielle à partir de données.

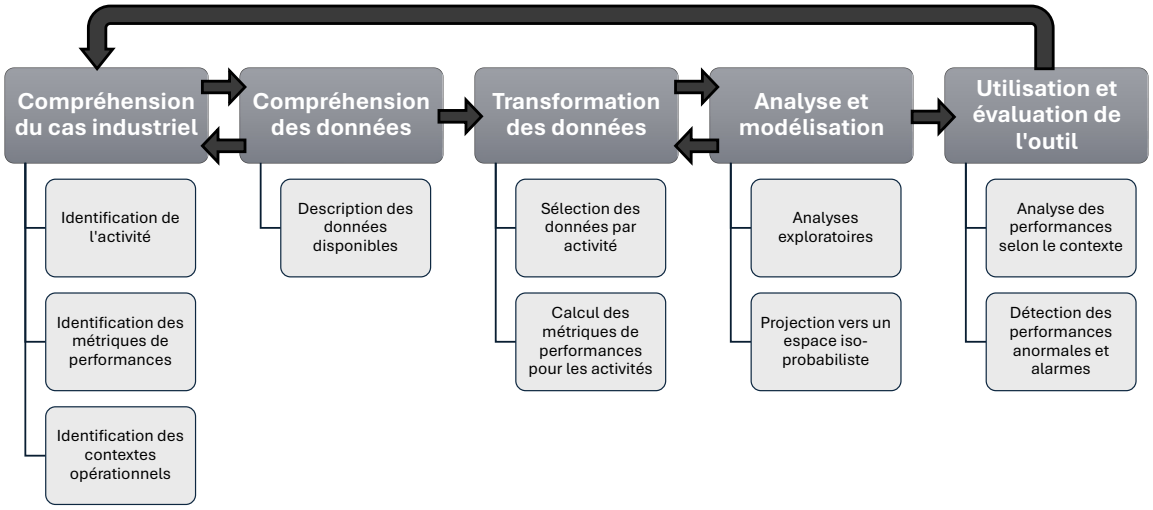

FIGURE 4.1 Vue schématique de la méthodologie proposée

4.1 Compréhension du cas industriel

La première étape de la méthodologie, comme dans CRISP-DM, correspond à la compréhension du cas d'un point de vue industriel. Avant de travailler avec les données, il convient de cadrer l'étude et ses limites. L'outil et les analyses portent sur une activité industrielle. Il est nécessaire d'identifier cette activité qui doit être répétable selon une structure commune. Le contexte de l'activité représente les paramètres en amont qui peuvent varier. Il peut comprendre le matériel, la durée ou les individus. Les performances représentent ce qui est mesuré à la sortie de l'activité et qui impacte la productivité ou les coûts de l'organisation. Deux mesures sont importantes pour définir la performance d'une activité : ce qu'elle consomme (en énergie, en matière première,..) et ce qu'elle produit (en termes de quantité par unité de

temps, par exemple). Selon le cas d'application, il est nécessaire de définir ces mesures de **consommation** et de **productivité**. Ces informations doivent être identifiées en relation avec des experts industriels.

Les mesures de performances permettent d'agréger différentes informations issues de l'activité dans un nombre restreint d'indicateurs. Ainsi, l'analyse de l'activité est simplifiée car le nombre de dimensions à comparer est réduit. Cela sera expliqué avec des indicateurs concrets dans le Chapitre 5.

À l'issue de cette phase de compréhension du cas industriel, la Figure 4.2 pourra être complétée pour résumer les informations sur l'activité industrielle.

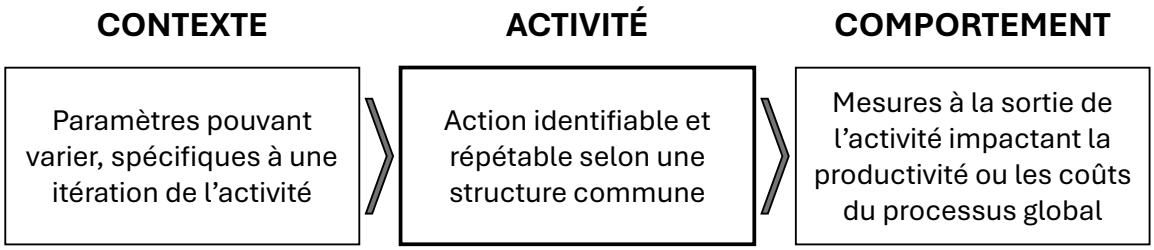

FIGURE 4.2 Activité industrielle étudiée dans le cas général

4.2 Compréhension des données

La deuxième étape de la méthodologie consiste à établir un état des lieux des données disponibles. Une fois l'activité étudiée définie, l'objectif consiste à faire l'inventaire des données. Cet inventaire intègre pour chaque variable : le nom, l'unité, la plage de variation possible, la fréquence de collecte, la catégorie (capteur, événement, mesure opérateur, etc.), ainsi que des commentaires spécifiques au cas d'étude (qualité de la mesure, source, transformations déjà appliquées,...). La structure de l'inventaire peut évoluer en fonction du cas d'étude. Un exemple de structure est présenté dans le Tableau 4.1.

TABLEAU 4.1 Exemple de structure d'inventaire des données disponibles

Nom	**Plage**	**Unité**	**Fréquence**	**Catégorie**	**Commentaires**
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

Le nombre d'itérations de l'activité étudiée disponible lors de la mise en place de la méthodologie ainsi que leur répartition dans le temps (par jour ou par semaine) sont également

importants à connaître. Cela permet d'évaluer la robustesse des résultats obtenus et les éventuels besoins en capacité d'enregistrement et de stockage de données pour enrichir la base de données et assurer la qualité et la fiabilité des résultats. Cette phase de la méthodologie est détaillée lors du Chapitre 5, elle est très dépendante du cas d'étude choisi.

4.3 Transformation des données

Cette étape de la méthodologie consiste à appliquer des transformations aux données pour préparer l'analyse et la modélisation. Elle correspond à la préparation de données dans la méthode CRISP-DM. À la fin de cette étape, la structure du modèle de données souhaitée est la suivante : une ligne par itération de l'activité avec des repères temporels, les mesures de performance et les données associées au contexte. Ce modèle facilite les comparaisons et les analyses statistiques qui sont détaillées dans la section suivante.

Les transformations à appliquer diffèrent selon le cas industriel particulier étudié, mais certaines idées sont communes. Par exemple, si les données brutes disponibles sont des données continues, les itérations des activités doivent être sélectionnées en se basant sur des informations extérieures (comme des journaux de production) ou sur le comportement de certaines variables. Une fois cette sélection réalisée, les données de contexte pour l'itération sont agrégées et les métriques de performance sont calculées. Cette phase est fortement dépendante de l'activité étudiée et des données disponibles. Elle est développée pour un cas réel dans le Chapitre 5.

4.4 Analyse et modélisation

L'étape d'analyse et de modélisation vise à extraire des informations exploitables à partir des données transformées. Elle comporte deux phases : une phase d'analyse exploratoire et une phase de modélisation.

4.4.1 Analyses exploratoires

L'analyse exploratoire consiste à appliquer des méthodes statistiques descriptives afin d'identifier les relations entre le contexte et les performances des itérations d'activité. Elle permet d'obtenir des premiers éléments d'interprétation qui peuvent être comparés aux résultats issus de la modélisation. Cette phase est réalisée pendant le développement de l'outil pour mieux comprendre les liens entre le contexte et les performances.

4.4.2 Projection vers un espace iso-probabiliste

La modélisation finale repose sur une projection des mesures de performance de chaque itération de l'activité dans un espace iso-probabiliste construit à partir d'un sous-ensemble de référence. Cet espace est conçu pour être visuellement interprétable : les distances euclidiennes mesurées représentent une proximité probabiliste entre les itérations, ce qui facilite l'identification de groupes de comportements et la détection d'activités atypiques.

La transformation appliquée aux données s'effectue en deux étapes :

1. Normalisation univariée des mesures de performance (consommation et productivité) à l'aide d'une transformation quantile vers une distribution normale. Cette étape standardise les distributions marginales vers des distributions normales, tout en conservant leur structure probabiliste : deux observations ayant la même probabilité dans la distribution d'origine auront également la même probabilité dans la distribution transformée.
2. Décorrélation multivariée qui centre les données sur la moyenne du sous-ensemble de référence, puis qui transforme la matrice de covariance des données transformées en la matrice identité I. Cela permet de rendre les deux axes comparables et de supprimer les redondances entre les dimensions.

Finalement, chaque observation est représentée comme un point dans cet espace centré sur l'activité "moyenne" du sous-ensemble de référence. La position reflète en quoi les performances s'écartent de ce comportement de référence, en intensité (distance au centre) et en nature (direction). Cette projection facilite l'analyse qualitative et quantitative des écarts de performance.

Choix du sous-ensemble pour créer l'espace de projection

Pour créer l'espace de projection, il est nécessaire de sélectionner le sous-ensemble de référence permettant la construction. Selon l'objectif d'analyse visé, ce sous-ensemble peut varier. Dans l'outil final, nous recommandons de laisser la possibilité à l'utilisateur opérationnel de choisir le sous-ensemble de création de l'espace pour permettre une plus grande liberté d'analyse et rendre l'outil adaptable.

Normalisation via transformation quantile

Les mesures de performances peuvent présenter des distributions empiriques très différentes, asymétriques, multimodales ou influencées fortement par des valeurs extrêmes. Pour les

rendre comparables et compatibles avec le calcul de distances significatives, les distributions marginales de chaque variable sont transformées vers une distribution normale centrée réduite.

Soit X une variable aléatoire avec une fonction de répartition empirique F et G la fonction de répartition d'une loi cible (ici, $Z \sim N(0,1)$). La normalisation s'écrit :

$$Z = G^{-1}(F(X)) \tag{4.1}$$

où G^{-1} est la fonction quantile de la loi cible (voir l'exemple sur la Figure 4.3). Lorsque G est la loi normale, cette fonction est la fonction *probit*. Cette transformation permet d'obtenir une variable normalisée tout en conservant la structure de rangs de la variable d'origine. La distribution obtenue est de la forme d'une distribution normale.

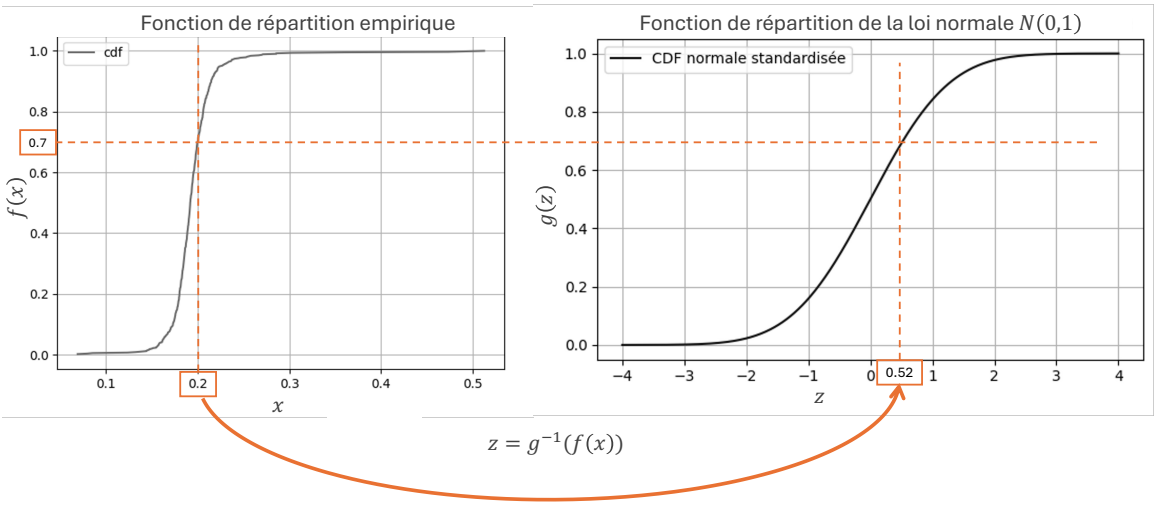

FIGURE 4.3 Transformation quantile d'une distribution empirique vers la loi normale centrée réduite

Décorrélation des données par décomposition de Cholesky

Une fois les mesures de performance normalisées de manière univariée, une décorrélation multivariée est appliquée. Cela permet de supprimer les corrélations linéaires entre les variables. Lorsque les axes sont corrélés, les points forment des ellipses orientées dans l'espace. Avec la décorrélation, les données sont centrées et les contours de densité deviennent circulaires. Les axes principaux deviennent orthogonaux et non corrélés. Il devient alors possible de tracer des cercles isoprobabilistes (1σ, 2σ, ...) dans lesquels une proportion connue d'observations est censée se trouver.

Différentes méthodes de décorrélation sont possibles (PCA, Analyse en composantes à phase nulle (*Zero-phase Component Analysis*) (ZCA)...) [Kessy *et al.*, 2018]. La décomposition de Cholesky est retenue ici pour sa simplicité d'implémentation dans un espace de faible dimension (deux variables principales).

Les étapes suivies pour obtenir l'espace iso-probabiliste sont décrites ici.

1. **Centrage des données** autour de la moyenne empirique $\boldsymbol{\mu}$ du sous-ensemble de référence :
$$\mathbf{Z}_C = \mathbf{Z} - \boldsymbol{\mu} \tag{4.2}$$
2. **Estimation de la matrice de covariance empirique** :
$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{Z}_C) \tag{4.3}$$
3. **Décomposition de Cholesky** de la matrice $\boldsymbol{\Sigma}$:
$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top \tag{4.4}$$
où $\mathbf{L}$ est une matrice triangulaire inférieure.
4. **Décorrélation des données** en appliquant l'inverse de $\mathbf{L}$ aux données centrées :
$$\mathbf{Z}_{\mathrm{sph}} = \mathbf{L}^{-1}\mathbf{Z}_C \tag{4.5}$$

Par construction, cette transformation garantit que la matrice de covariance des données projetées devient la matrice identité :

$$\mathrm{Cov}(\mathbf{Z}_{\mathrm{sph}}) = \mathbf{I} \tag{4.6}$$

Plus de détails sur la décomposition de Cholesky sont donnés en Annexe A.

Ainsi, les variables sont à la fois décorrélées et normalisées, ce qui permet une représentation géométrique cohérente des observations dans l'espace iso-probabiliste.

Mesure de distance

La distance de Mahalanobis, décrite dans le Chapitre 2, permet de mesurer la similarité entre deux séries de données en tenant compte des relations entre variables.

Dans notre contexte, après la transformation iso-probabiliste, la matrice de covariance des

données devient la matrice identité $\mathbf{I}$. La distance de Mahalanobis se simplifie et devient équivalente à la distance euclidienne classique :

$$d_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^\top \mathbf{I}^{-1} (\mathbf{x} - \mathbf{y})} = \sqrt{(\mathbf{x} - \mathbf{y})^\top (\mathbf{x} - \mathbf{y})} = \|\mathbf{x} - \mathbf{y}\|_2. \tag{4.7}$$

Cette équivalence montre l'intérêt de la transformation présentée : elle permet d'interpréter les distances probabilistes (Mahalanobis) dans l'espace transformé comme des distances euclidiennes. Cela facilite l'établissement de seuils de confiance pour évaluer si deux points diffèrent significativement ou si un point est atypique par rapport à un ensemble. En effet, l'espérance de la distance de Mahalanobis suit une loi du χ^2. Considérons un vecteur aléatoire $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ suivant une loi normale multidimensionnelle d'espérance $\boldsymbol{\mu}$ et de matrice de covariance $\boldsymbol{\Sigma}$ définie positive. Dans ce cas, le carré de la distance de Mahalanobis entre $\mathbf{X}$ et son espérance suit une loi du χ^2 à p degrés de liberté :

$$D_M^2(\mathbf{X}, \boldsymbol{\mu}) \sim \chi_p^2. \tag{4.8}$$

Ainsi, si $\chi^2_{p;1-\alpha}$ représente le quantile d'ordre $1 - \alpha$ de la loi du χ^2 à p degrés de liberté, on obtient :

$$\mathbb{P}\left[D_M^2(\mathbf{X}, \boldsymbol{\mu}) \leq \chi^2_{p;1-\alpha}\right] = 1 - \alpha. \tag{4.9}$$

Grâce à la transformation quantile, nous nous assurons que les vecteurs de données suivent une loi normale multidimensionnelle et qu'il est possible d'utiliser cette propriété. Les distances mesurées dans l'espace transformé deviennent directement interprétables.

Soit $\mathbf{Z} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ le vecteur aléatoire transformé, centré-réduit et décorrélé. Alors, la distance euclidienne au carré d'une observation à la moyenne de référence suit une loi du χ^2 à p degrés de liberté :

$$\|\mathbf{Z}\|_2^2 \sim \chi_p^2. \tag{4.10}$$

Cela permet de fixer des seuils pour détecter des points anormaux dans l'espace par rapport à l'ensemble de référence choisi. En particulier, pour deux variables ($p = 2$) (par exemple, la productivité et la consommation spécifique), un seuil d'anomalie au niveau $\alpha = 0,05$ (par exemple, selon le cas d'autres valeurs peuvent être utilisées) est donné par le quantile d'ordre $0,95$ de la loi du χ^2 à 2 degrés de liberté, soit :

$$\chi^2_{2;0,95} \approx 5,99. \tag{4.11}$$

La racine carrée de cette valeur donne le seuil de distance euclidienne correspondant :

$$\sqrt{\chi^2_{2;0,95}} \approx 2,45. \tag{4.12}$$

Ainsi, toute observation située à un rayon supérieur à une distance de $2,45$ à partir du centre $\mathbf{0}$ de l'espace iso-probabiliste serait considérée comme statistiquement anormale à un seuil de confiance de 95%.

4.5 Utilisation et évaluation

La modélisation des données de performances dans un espace iso-probabiliste, telle que décrite précédemment, vise une utilisation par les gestionnaires opérationnels. Cet espace permet de définir des seuils et des marges de normalité ; son aspect visuel offre également un grand intérêt aux décideurs. Deux utilisations classiques de l'outil sont présentées ci-dessous.

4.5.1 Compréhension des performances

Lors de la phase d'analyse préliminaire, des analyses de variances (ANOVA) sont réalisées pour déterminer les effets du contexte sur les performances. L'une des limites de ce type d'analyse est que seule une mesure de performance peut être évaluée à la fois. En utilisant la transformation iso-probabiliste à partir d'un ensemble de référence, il est possible d'étudier l'effet d'un paramètre de contexte sur plusieurs mesures de performance.

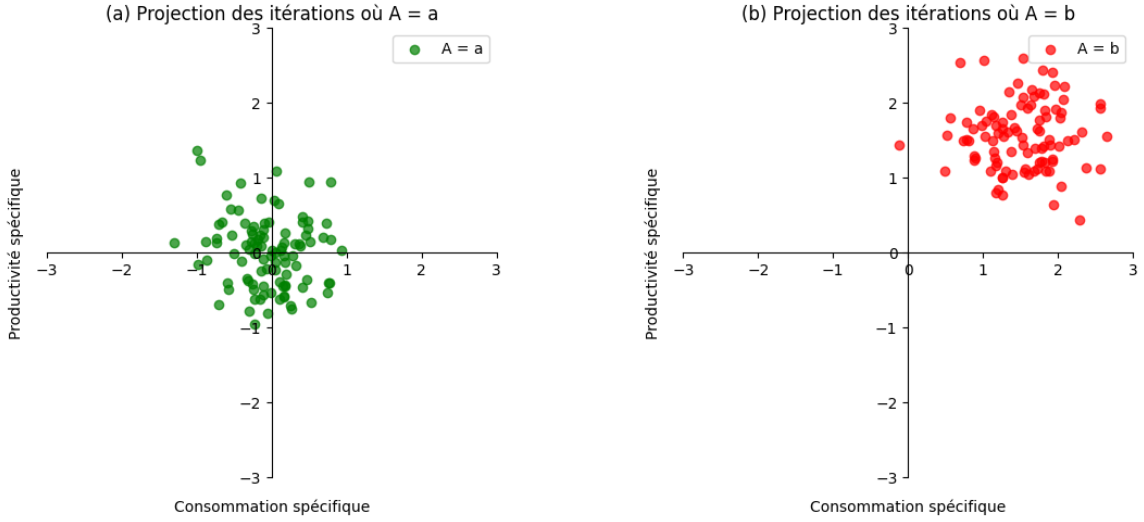

FIGURE 4.4 Projection d'itérations d'une activité dans l'espace probabiliste créé à partir de l'ensemble des points tels que $A = a$

Exemple : Le gestionnaire souhaite comprendre l'effet du paramètre de contexte A qui peut prendre les valeurs a ou b. Il choisit comme ensemble de référence les itérations de l'activité réalisée quand $A = a$ puis il projette l'ensemble des itérations dans l'espace. Il attribue une couleur différente pour chaque valeur possible de A. Les itérations réalisées avec $A = a$ forment par construction un nuage de points sphérique centré sur $\mathbf{0}$. La position et la forme des autres nuages de points permettent de déterminer l'influence du contexte sur les mesures de performance. La Figure 4.4 présente un tel exemple. L'ANOVA montre que la productivité et la consommation sont plus élevées dans le contexte $A = b$. Ceci est directement visible via l'outil et la transformation iso-probabiliste avec le nuage de points rouge qui est décalé en haut et à droite par rapport au centre $\mathbf{0}$ et au nuage de points verts.

4.5.2 Détection des performances anormales et alarmes

L'outil peut être utilisé pour suivre les performances dans le temps et pour déclencher des alarmes. La transformation iso-probabiliste permet de projeter les données vers un espace centré en $\mathbf{0}$. La position d'un point représentant une itération permet de déterminer les caractéristiques de l'itération par rapport à celles de l'ensemble de référence. La position du point est donc un indicateur intéressant ; il devient alors possible de calculer des proportions dans certaines zones.

Dans le Chapitre 2, nous montrons un intérêt pour le lien entre baisse de performance et maintenance. Nous proposons ainsi de mettre en place des alertes de maintenance à partir de la position des itérations d'activités. Tout d'abord, en fonction des mesures de performances considérées, une zone sous-optimale est définie. Elle représente la zone où toutes les performances sont moins bonnes que celles de l'activité moyenne de l'ensemble de référence. Les données historiques sont utilisées comme ensemble de référence et, pour une période donnée (un quart de travail, une durée fixe, un nombre d'itérations fixe, à définir selon le cas d'étude), les itérations sont projetées dans l'espace créé. La proportion d'itérations dans la zone sous-optimale est calculée.

Cette proportion est comparée à la moyenne des proportions des périodes précédentes et à un seuil fixe. Si les deux seuils sont dépassés pour un nombre de périodes suffisant, alors une alerte est émise. La taille de l'historique utilisé comme ensemble de référence, la nature et la taille de la période d'analyse et les seuils utilisés sont des hyperparamètres qui doivent être établis. Pour déterminer de manière optimale ces hyperparamètres, nous conseillons d'utiliser une base de données d'activités labellisées avec la durée restante avant défaillance pour vérifier si les alertes permettent effectivement d'anticiper les pannes ou maintenances.

CHAPITRE 5 CAS D'ÉTUDE

Le partenaire industriel exploite des mines dans le nord du Québec et cherche à augmenter la productivité de ses activités tout en réduisant les coûts. L'une de ces activités est l'extraction du minerai souterrain. L'extraction consiste à fragmenter et détacher la roche de son environnement géologique, puis à transporter le minerai extrait dans les chantiers souterrains jusqu'à la surface, où il sera trié et traité. Le transport est assuré par des camions miniers spécialisés capables de transporter plusieurs dizaines de tonnes (voir la Figure 5.1). Ces véhicules circulent en continu dans les galeries et la rampe d'accès pendant les quarts de travail. Cette activité a un impact direct sur la productivité et les coûts d'exploitation : des temps de cycle réduits, une consommation de carburant optimisée et une disponibilité élevée des équipements permettent d'en améliorer l'efficacité.

FIGURE 5.1 Camion de transport minier souterrain [Sandvik Mining and Rock Solutions, 2024]

Nous décidons de nous concentrer sur cette activité particulière : la remontée de roche effectuée par un certain type de camion. À partir des données récoltées lors du fonctionnement des camions, nous proposons de développer un outil de support à la performance industrielle basé sur les données et à destination du gestionnaire de flotte grâce à la méthodologie présentée au Chapitre 4. Cette application dans le cadre industriel permet de juger de la pertinence de l'approche.

5.1 Compréhension du cas industriel

L'activité industrielle étudiée consiste en un voyage de camion depuis le chantier en bas de la mine vers la surface. Le trajet commence une fois le camion chargé et se termine lorsque la benne est vide, une fois arrivé en surface. Le trajet a principalement lieu dans la rampe d'accès du site, qui permet de relier les galeries et les chantiers à la surface. La pente moyenne de la rampe est de 15%. Un trajet dure en moyenne 22 minutes, sur une distance comprise entre 1.2 et 6 kilomètres. Plusieurs types de camions, différents selon le volume de la benne, permettent de transporter du minerai. Nous nous concentrons ici sur les camions avec une benne de 63 tonnes.

Selon le chantier de départ, la distance et la durée du trajet ainsi que la nature de la roche transportée peuvent être différentes. La benne peut également être plus ou moins remplie. Ces différences entre les itérations de la même activité sont importantes à prendre en compte. Pour analyser et comparer les performances de plusieurs trajets, il est nécessaire de mesurer des indicateurs spécifiques qui agrègent toutes ces informations.

Ainsi pour l'activité qui consiste à monter une charge avec un camion, les indicateurs de performance choisis représentent la productivité spécifique et la consommation spécifique.

- La productivité spécifique mesure la quantité de travail de transport (charge transportée × distance parcourue) réalisée par unité de temps. Elles s'exprime en $t{\cdot}km{\cdot}h^{-1}$.
- La consommation spécifique mesure le volume de carburant consommé pour déplacer une unité de charge pendant une unité de temps sur une unité de distance. Elle s'exprime en $l{\cdot}km^{-1}{\cdot}t^{-1}$.

Au lieu d'analyser quatre dimensions (charge transportée, distance parcourue, temps de trajet, volume de carburant consommé) pour caractériser l'activité et comparer les itérations, les indicateurs de performance spécifiques définis permettent de travailler avec seulement deux dimensions. Cela rend l'information intelligible étant donnée la facilité d'analyse de figures à deux dimensions.

Les paramètres de contexte pour l'activité sont :

- l'âge du camion, c'est-à-dire la distance qu'il a déjà parcourue avant d'effectuer le trajet,
- le numéro du camion (un modèle étudié, mais plusieurs camions similaires dans la mine),
- le rapport principalement utilisé par le conducteur lors de la montée,
- la charge transportée dans la benne.

D'autres paramètres de contexte pourraient être ajoutés, notamment pour définir plus préci-

sément la manière de conduire ou la mission, mais nous choisissons de considérer les quatre présentés ci-dessus dans le cadre de notre étude.

Les paramètres de contexte amènent une troisième dimension aux analyses, les indicateurs de performances pourront être calculés pour l'un ou l'autre de ces paramètres. En fixant un tel cadre à l'activité étudiée, nous facilitons la suite des analyses et des comparaisons.

Ces informations permettent de compléter la Figure 5.2 qui reprend les informations de manière synthétique.

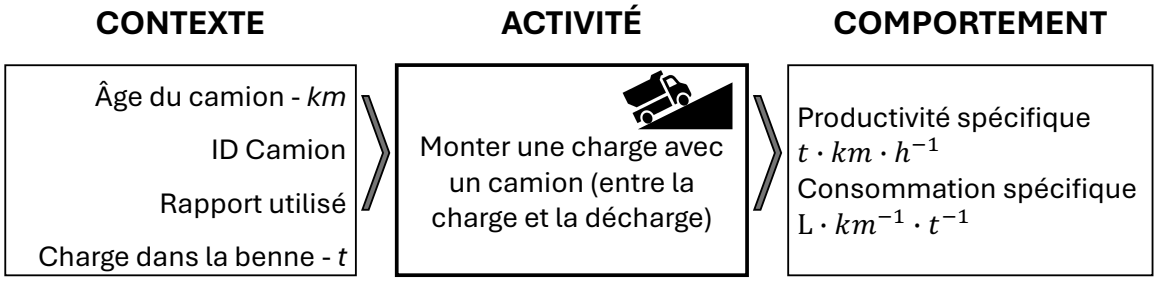

FIGURE 5.2 Activité industrielle étudiée pour le cas d'étude

5.2 Compréhension des données

Une fois l'activité étudiée définie, nous nous intéressons aux données à notre disposition. Le partenaire industriel partage des données de fonctionnement des camions. Cela comprend des mesures effectuées par des capteurs sur des sous-systèmes spécifiques, tels que le moteur ou la transmission, ainsi que des mesures plus globales sur le camion et son état, comme la charge dans la benne, le kilométrage ou l'appui sur la pédale d'accélérateur. Une liste des capteurs disponibles est présentée en Annexe B. Les données sont enregistrées toutes les 0.5 secondes.

Aucune donnée de position n'est disponible, nous ne disposons pas non plus d'informations concernant le planning ou les missions affectées aux camions, ni d'historique de défauts et de maintenance. Les données de fonctionnement sont sauvegardées pendant trois mois puis effacées par le partenaire. L'historique est donc réduit.

Trois échantillons de données concernant six camions (même modèle mais pas la même date de mise en service pour tous) sont disponibles pour le cas d'étude. Une première période entre le 1er et le 3 décembre contient 159 montées réalisées par cinq camions différents. Entre le 6 et le 21 mars, 617 montées sont répertoriées pour les cinq mêmes camions. Enfin, 173 montées réalisées entre le 24 et le 26 avril sont enregistrées. Elles ont été réalisées par deux

camions qui ont subi une panne à la suite de la période. La Figure 5.3 montre la distribution temporelle des montées pour chaque camion.

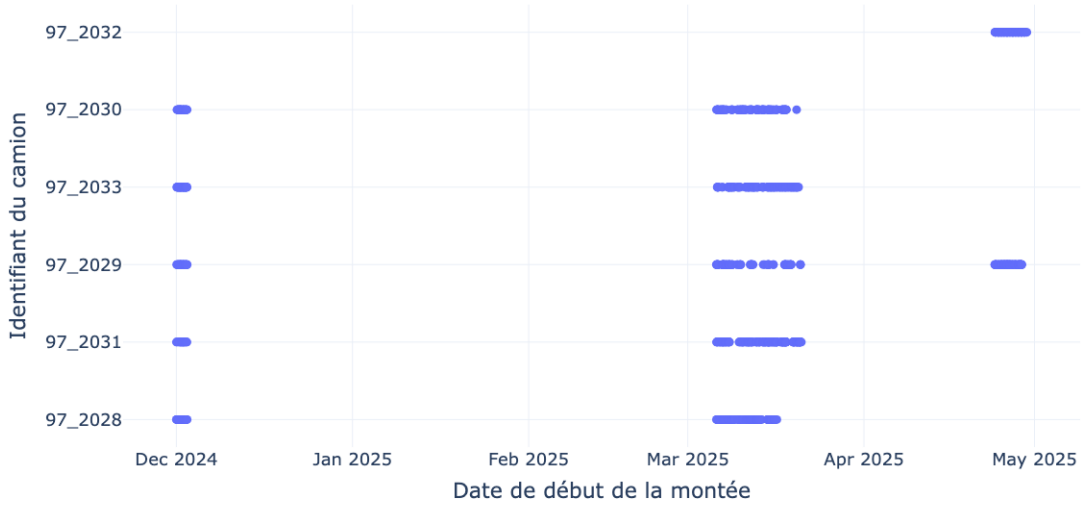

FIGURE 5.3 Distribution temporelle des montées par camion

La répartition des montées selon les paramètres de contexte est étudiée. En s'intéressant au rapport de transmission engagé par le conducteur pendant la majorité du temps de montée, il apparaît que la plupart des montées sont réalisées avec le rapport A (770, 81%). Seulement 19% des montées sont réalisées avec le rapport B. La Figure 5.4 montre la répartition. Une montée est considérée comme réalisée dans un rapport spécifique si le camion passe plus de 50% du temps de montée dans ce rapport. L'effet de l'utilisation du rapport A ou du rapport B est l'une des questions qui intéresse le partenaire pour effectuer des recommandations aux conducteurs, et l'outil développé a pour but d'y apporter une réponse.

La Figure 5.5 montre la répartition des montées selon la charge dans la benne. Pour la période de temps étudiée, la charge pour une montée est comprise entre 16 et 80 tonnes. Dans la majorité des cas, le camion est chargé avec 40 à 60 tonnes de roches.

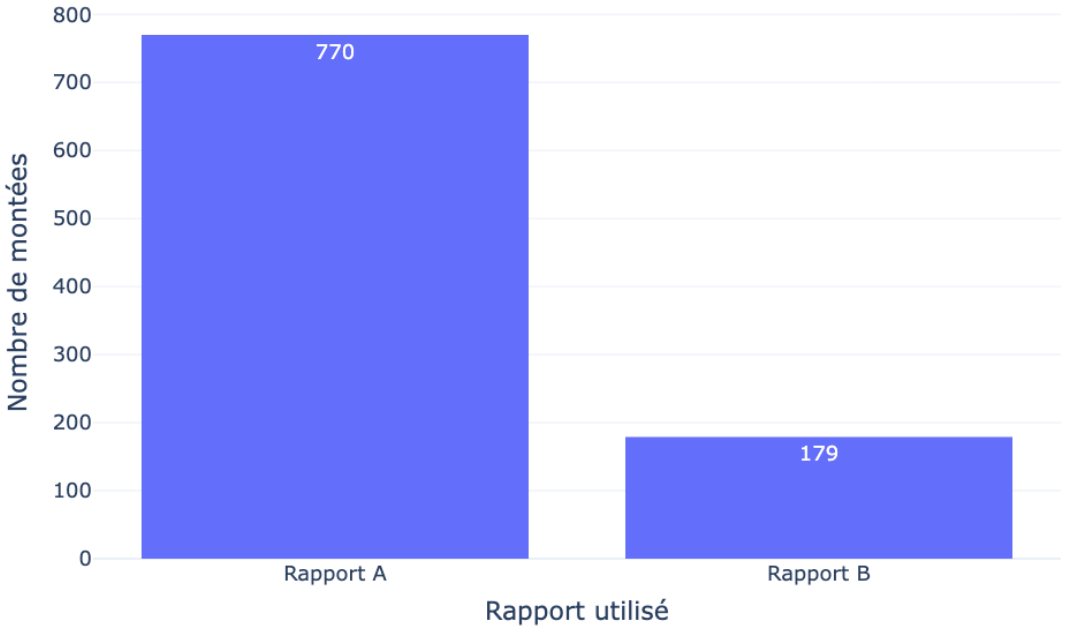

FIGURE 5.4 Répartition du nombre de montées selon le rapport de vitesse engagé (Rapport A ou Rapport B)

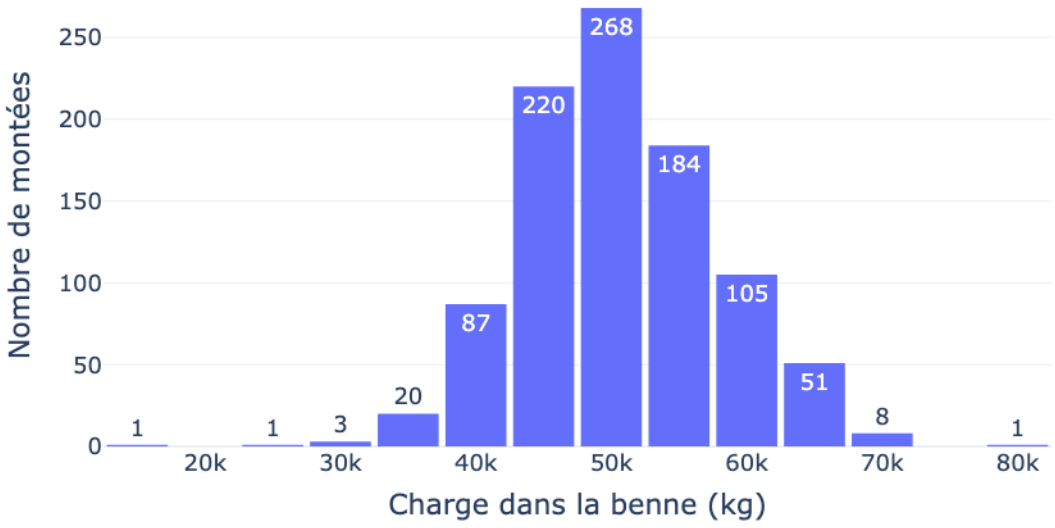

FIGURE 5.5 Répartition du nombre de montées selon la charge dans la benne

5.3 Transformation des données

Pour faciliter les comparaisons et les analyses statistiques, les données doivent être organisées de la manière suivante : une ligne par montée avec des repères temporels, les mesures de performance et les données associées au contexte.

5.3.1 Sélection des activités

Les données brutes sont continues pendant le fonctionnement du camion. Pour isoler les montées, il faut sélectionner les données correspondantes. Deux mesures permettent d'isoler les périodes de montée du reste dans les données de fonctionnement du camion : la mesure de charge dans la benne et la mesure de l'appui sur la pédale d'accélération. L'idée est la suivante : un camion qui effectue une montée est chargé et le conducteur a besoin d'appuyer plus sur l'accélérateur car le terrain est en pente ascendante. La seule mesure de la charge n'est pas suffisante car, après les discussions avec les experts industriels, il apparaît que certaines descentes sont effectuées avec du stérile dans la benne (déchets constitués des roches extraites pour accéder au minerai). Ce stérile est réutilisé pour remblayer les chantiers où le minerai a été extrait afin d'atténuer les pressions de terrain exercées sur les zones excavées. La mesure d'appui sur la pédale d'accélération vient en complément pour assurer la détection des montées.

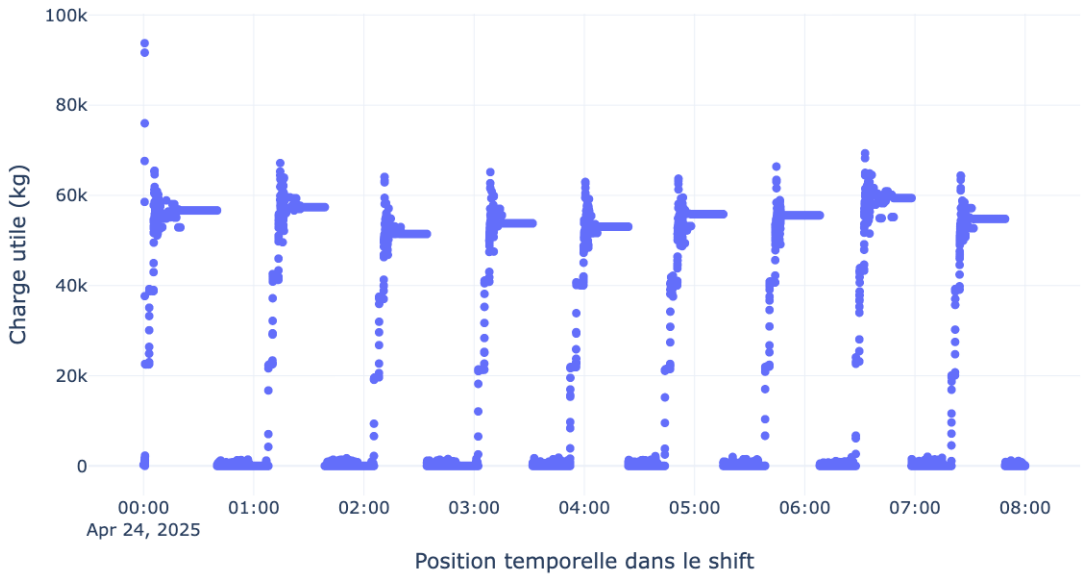

FIGURE 5.6 Évolution de la charge dans la benne au cours d'un quart de travail pour un camion en particulier

La Figure 5.6 montre l'évolution de la charge dans la benne d'un camion le long d'un quart de travail. Les périodes intéressantes pour notre étude sont les plateaux hauts, représentant les périodes de voyage en charge. La détection des montées est réalisée selon les étapes suivantes :

- Pour chaque camion, les enregistrements de charge sont triés chronologiquement.
- Pour réduire le bruit et les fluctuations instantanées dues à la mesure de charge en environnement minier, une moyenne mobile est appliquée à la mesure de charge, avec une fenêtre de 2 minutes. Cela permet de rendre plus stable l'évolution de la charge transportée.
- L'hypothèse opérationnelle est que lors d'une montée, la charge transportée est stable. La différence entre la mesure à l'instant t et l'instant $t-1$, équivalent à la dérivée première discrète, permet d'évaluer cette stabilité. Les points pour laquelle cette variation est comprise dans une bande centrée sur zéro sont considérés comme pouvant appartenir à une montée.
- Chaque bloc successif de points satisfaisant cette condition reçoit alors un identifiant unique, ce qui permet de diviser les différentes phases candidates. Ces phases représentent des périodes ou la charge dans la benne est stable.
- Un ensemble de trois règles provenant de l'expertise métier sont appliquées pour définir si une phase candidate est bien une montée ou non :

  1. Si la durée de la phase est supérieure à 12 minutes,
  2. Si l'appui moyen sur la pédale d'accélérateur est supérieur à 60%,
  3. Si la charge dans la benne est supérieure à 10 tonnes.

  Les phases respectant ces conditions sont enregistrées comme des montées, avec le premier et le dernier point représentant les instants de début et de fin.

5.3.2 Calcul des métriques de performance et des données associées au contexte

Grâce aux étapes décrites dans la section précédente, les montées et leurs repères temporels sont connus. Il faut maintenant ajouter les métriques de performance et les données associées au contexte.

L'âge du camion, c'est-à-dire le nombre de kilomètres qu'il a parcourus avant de commencer la montée, correspond à la plus petite distance au compteur enregistrée lors de la phase de montée.

Le rapport principal utilisé peut correspondre au rapport A ou au rapport B d'après les discussions avec les experts. Si le quotient calculé avec l'Équation 5.1 est supérieur à 0.5, la montée est considérée comme réalisée en rapport B ; sinon, elle est enregistrée comme étant

réalisée en rapport A.

$$\text{Rapport principal utilisé} = \frac{\text{Temps}_{\text{Rapport B}}}{\text{Temps}_{\text{Rapport B}} + \text{Temps}_{\text{Rapport A}}} \tag{5.1}$$

La charge dans la benne correspond à la charge médiane lors de la phase de montée. Pour éviter qu'une mesure ponctuelle anormale de charge impacte fortement la représentation de la charge lors de la montée, nous utilisons la médiane plutôt que la moyenne.

La productivité spécifique $P_{\text{spé}}$ exprimée en $t{\cdot}km{\cdot}h^{-1}$ est calculée grâce à la distance parcourue pendant la montée, la charge dans la benne et la durée de la montée selon l'Équation 5.2.

$$P_{spé}\ (\text{t}\cdot\text{km}\cdot\text{h}^{-1}) = \left(\frac{\text{Charge (kg)}}{1000}\right) \times \left(\frac{\text{Distance (m)}}{1000}\right) \times \left(\frac{3600}{\text{Durée (s)}}\right) \tag{5.2}$$

La consommation spécifique $C_{spé}$ exprimée en $l{\cdot}km^{-1}{\cdot}t^{-1}$ est calculée en prenant en compte la charge totale déplacée, c'est-à-dire la charge dans la benne à laquelle est ajoutée la masse à vide du camion selon le constructeur (48440 tonnes). Nous souhaitons mesurer combien le camion consomme d'essence pour se déplacer sur une distance en un certain temps. Le volume d'essence utilisé $V_{essence}$ est déterminé avec l'Équation 5.3 qui est fonction de la moyenne du débit d'essence entrant dans le moteur durant la phase ascendante.

$$V_{essence}\ (\text{l}) = \text{Débit moyen}\ (\text{l}\cdot\text{h}^{-1}) \times \frac{\text{Durée (s)}}{3600} \tag{5.3}$$

La consommation spécifique est obtenue grâce à l'Équation 5.4.

$$C_{spé}\ (\text{l}{\cdot}\text{km}^{-1}{\cdot}\text{t}^{-1}) = V_{essence}\ (\text{l}) \times \left(\frac{1000}{\text{Distance (m)}}\right) \times \left(\frac{1000}{\text{Charge déplacé (kg)}}\right) \tag{5.4}$$

Finalement, la base de données des montées obtenues après ces transformations ressemble au Tableau 5.1.

Tableau 5.1 Extrait de la base de données contenant les informations des montées

Début	**Fin**	**Âge camion** (km)	**Camion**	**Rapport**	$P_{spé}$	$C_{spé}$
05 :14 :28	05 :35 :35	1.62×10^7	97_2030	Rapport A	525.184	0.203130
07 :55 :41	08 :28 :07	1.73×10^7	97_2033	Rapport A	602.306	0.172907
01 :17 :00	01 :30 :22	3.68×10^7	97_2029	Rapport A	438.434	0.209005
13 :57 :15	14 :19 :57	2.18×10^7	97_2031	Rapport A	421.132	0.209565

5.4 Analyse et Modélisation

5.4.1 Analyses exploratoires

Grâce à la base de données des itérations de montées, il est possible de réaliser des premières analyses pour comprendre l'influence des paramètres de contexte sur les performances. Nous détaillons le processus d'ANOVA pour évaluer l'influence du rapport utilisé sur la productivité spécifique.

L'ensemble des montées est utilisé pour réaliser les analyses de variances. 770 montées sont réalisées avec le rapport A, 179 avec le rapport B (Figure 5.4). L'hypothèse de normalité pour la variable dépendante, c'est-à-dire la productivité spécifique, est validée visuellement. L'hypothèse d'homoscédasticité est vérifiée grâce au test statistique de Levene. L'hypothèse nulle testée est la suivante : les variances dans les groupes sont égales. Pour les deux groupes étudiés (montées en rapport A et montées en rapport B), le test donne une valeur-p ≈ 0.642, bien supérieure au seuil de 0.05. L'hypothèse nulle n'est pas rejetée et nous considérons donc les variances égales. L'analyse de variance peut être appliquée. L'hypothèse nulle testée considère l'égalité de la moyenne des groupes et aucune influence du facteur rapport sur la productivité. Les résultats obtenus sont les suivants : $F = 19.78$, valeur-p ≈ 0.0001. L'hypothèse nulle est rejetée, la différence de moyenne entre les groupes est significative. Un test post-hoc, le test de Tukey, permet ensuite de comparer les moyennes entre les groupes. Ce test confirme une différence significative entre les deux groupes et montre qu'en moyenne, les montées réalisées avec le rapport A présentent une productivité supérieure à celles réalisées avec le rapport B ($\Delta = 30.17$). Ces résultats apparaissent visuellement dans le diagramme de Tukey présenté en Figure 5.7.

Les mêmes analyses sont réalisées pour évaluer l'influence de l'âge du camion et de la charge dans la benne sur la productivité spécifique et la consommation spécifique des montées. Les résultats sont disponibles dans le Tableau 5.2. Pour chaque variable de contexte, les différents groupes sont détaillés et l'influence sur la mesure de performance associée est expliquée. Les diagrammes de Tukey associés sont disponibles en Annexe C.

5.4.2 Projection vers un espace iso-probabiliste

Après avoir réalisé des analyses préliminaires pour explorer et mieux comprendre le jeu de données, la création d'un espace iso-probabiliste peut être réalisée. La base de données des montées, présentée dans le Tableau 5.1, est utilisée. Pour détailler visuellement les effets de la transformation, la Figure 5.8 présente le nuage de points de l'ensemble des montées de la base avec la productivité spécifique en ordonnée et la consommation spécifique en abscisse.

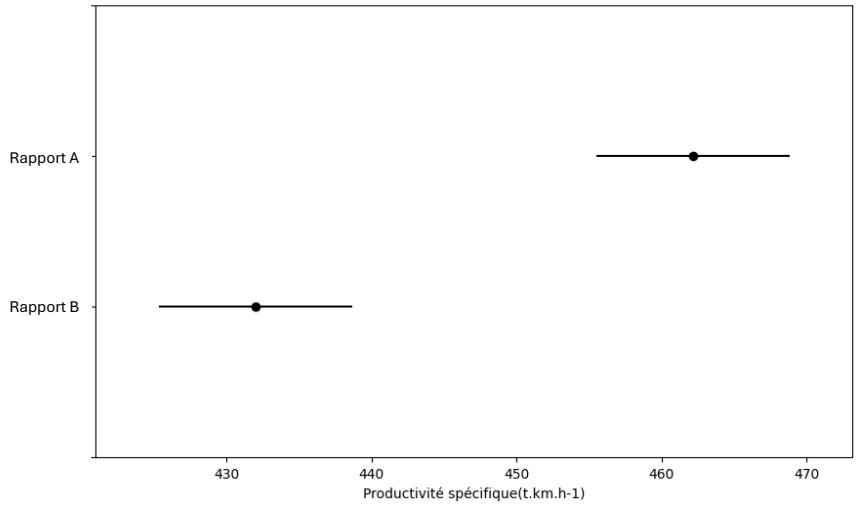

FIGURE 5.7 Diagramme de Tukey pour la productivité spécifique selon le rapport utilisé

TABLEAU 5.2 Effet des variables de contexte sur les mesures de performance

Mesure de performance	**Contexte**	**Groupe 1**	**Groupe 2**	**Groupe 3**	**Influence du contexte**
Productivité spécifique	Rapport utilisé	Rapport A	Rapport B	–	Oui – Moyenne du groupe 1 plus élevée
Productivité spécifique	Charge dans la benne	< 45 t	45–55 t	> 55 t	Oui – Moyenne du groupe 1 plus faible, moyenne du groupe 3 plus élevée
Productivité spécifique	Âge du camion	< 7000 h	> 7000 h	–	Oui – Moyenne du groupe 1 plus élevée
Consommation spécifique	Rapport utilisé	Rapport A	Rapport B	–	Oui – Moyenne du groupe 1 plus faible
Consommation spécifique	Charge dans la benne	< 45 t	45–55 t	> 55 t	Oui – Moyenne du groupe 3 plus faible ; groupes 1 et 2 semblables
Consommation spécifique	Âge du camion	< 7000 h	> 7000 h	–	Oui – Moyenne du groupe 1 plus faible

La forme du nuage est une ellipse allongée, les valeurs de consommation spécifique sont très proches.

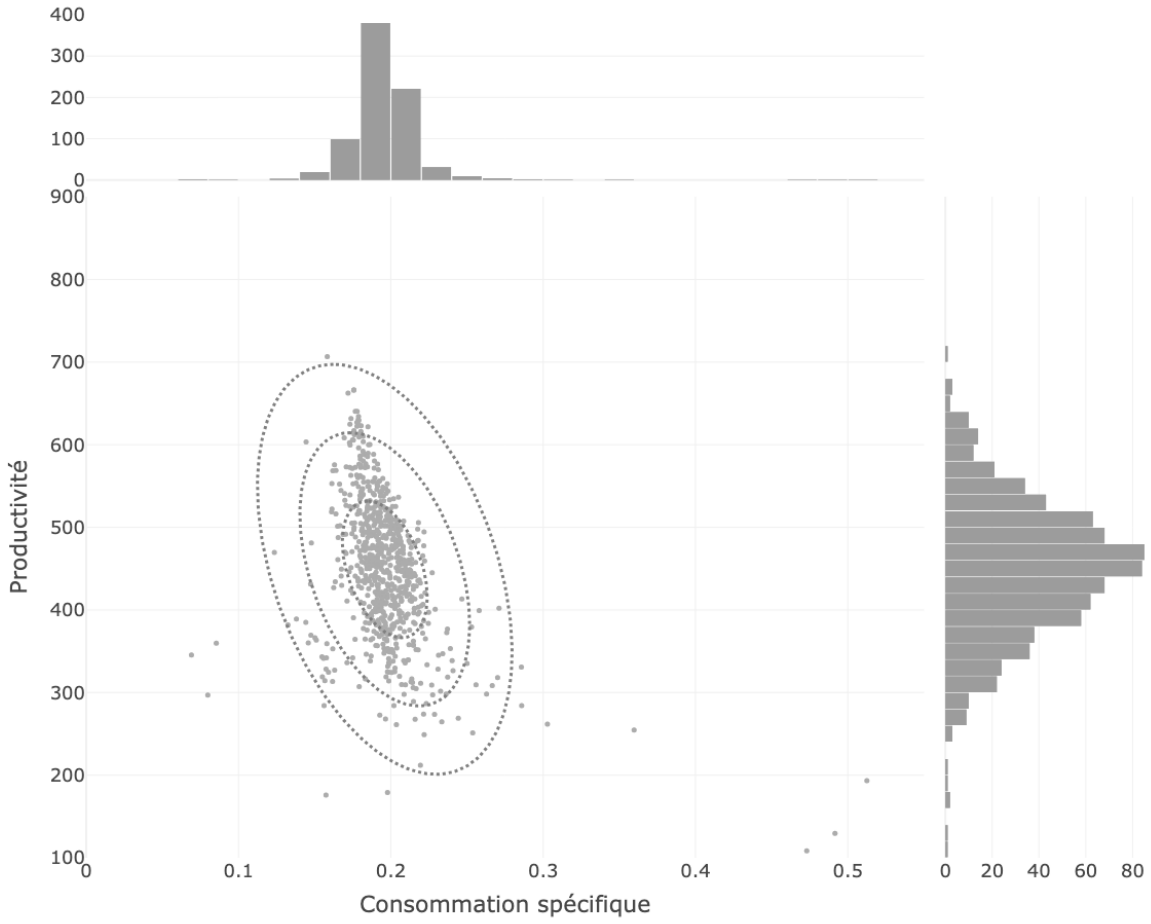

FIGURE 5.8 Projection brute des données de consommation spécifique et de productivité spécifique pour l'ensemble des activités de la base de données

Choix du sous-ensemble pour créer l'espace de projection

Dans le Chapitre 4, nous expliquons qu'un sous-ensemble de données doit être choisi pour créer l'espace de projection. Le Tableau 5.3 résume différents sous-ensembles utilisés et le but de leur utilisation. Les résultats et informations rendus disponibles sont détaillés par la suite dans la Section 5.5.

Tableau 5.3 Cas d'utilisations de différents sous-ensembles de référence

Sous-ensemble	**Données projetées**	**Utilisation**
Montées en rapport A	Toutes les montées, colorées selon le rapport utilisé	Visualiser et évaluer l'influence du rapport utilisé
Montées avec charge dans la benne entre 45 t et 55 t	Toutes les montées, colorées selon la charge dans la benne	Visualiser et évaluer l'influence de la charge dans la benne
Montées par un camion avec moins de 7000 h de fonctionnement	Toutes les montées, colorées selon l'âge du camion	Visualiser et évaluer l'influence de l'âge du camion
Toutes les montées sur une période donnée	Montées d'un seul camion	Visualiser le profil d'un camion pour le comparer par rapport à la flotte
Toutes les montées	Toutes les montées	Détecter les montées anormales et les analyser

Normalisation via transformation quantile

Après la sélection du sous-ensemble de référence, une normalisation univariée est appliquée aux données. La transformation est effectuée avec le langage Python, grâce à la bibliothèque `scikit-learn` [Pedregosa *et al.*, 2011] et à la fonction `QuantileTransformer`. La transformation est appliquée aux mesures de performance. La Figure 5.9 montre le nuage de points de l'ensemble des montées de la base après l'application de la transformation quantile. Les distributions de la productivité et de la consommation spécifique sont amenées vers des distributions normales centrées et réduites.

Les données du sous-ensemble de référence servent à étalonner la transformation et à établir les seuils pour transformer les points externes. Ainsi, il est essentiel que la taille de ce sous-ensemble soit suffisamment grande pour éviter que les queues des distributions ne soient trop compressées.

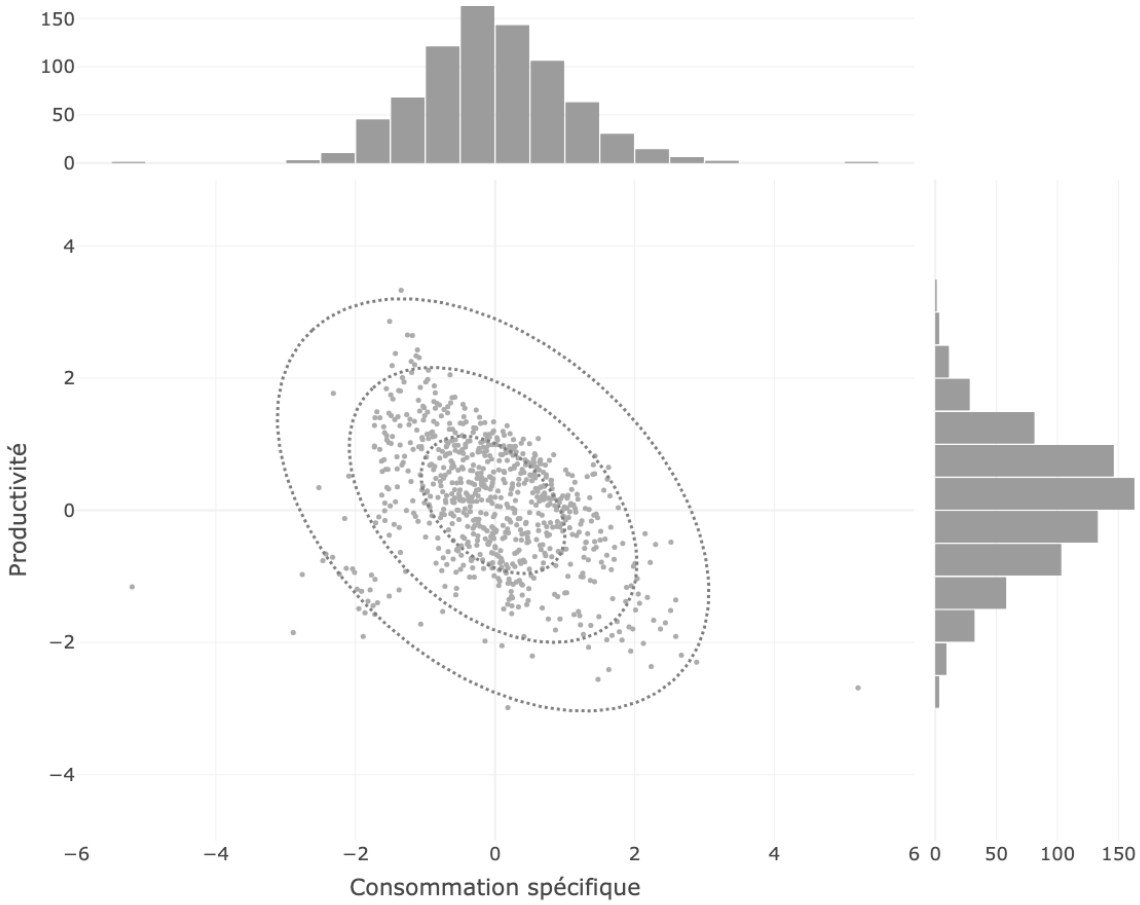

FIGURE 5.9 Projection des données de consommation et de productivité pour l'ensemble des activités de la base de données après l'application de la transformation quantile

Décorrélation des données par décomposition de Cholesky

La décorrélation des données est réalisée avec le langage Python et la bibliothèque `Numpy` [Harris *et al.*, 2020]. La matrice de Cholesky est calculée à partir du sous-ensemble de référence choisi puis la transformation est appliquée à toutes les données à analyser. La Figure 5.10 présente la projection finale des montées dans l'espace transformé, dans le cas où l'ensemble des montées est utilisé comme sous-ensemble de référence. Le nuage de points est de forme sphérique, centré sur $\mathbf{0}$. Les cercles de rayon 1, 2 et 3 correspondent à des contours iso-probabilistes, équivalents à 1, 2 et 3 écarts-types.

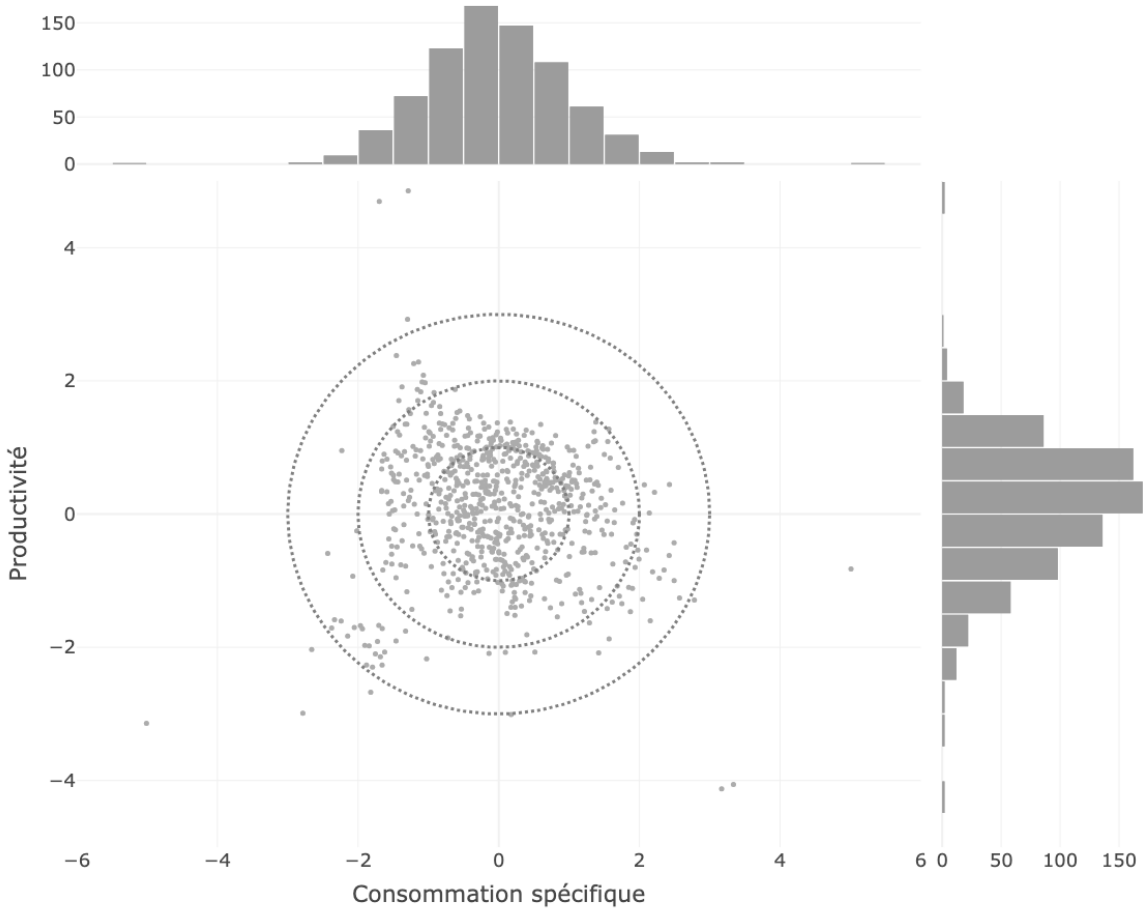

FIGURE 5.10 Projection finale des données de consommation et de productivité pour l'ensemble des activités de la base de données après l'application de la transformation quantile et de la décorrélation

5.5 Utilisation et résultats

En appliquant la méthodologie décrite au Chapitre 4, il est désormais possible de développer un outil fondé sur la transformation des données mentionnée pour supporter la performance. Cette section présente les utilisations possibles de l'outil et les informations qu'il est possible d'obtenir sur l'activité étudiée, la montée de minerai par les camions. Les résultats de l'analyse exploratoire permettent d'effectuer des comparaisons. Nous montrons ensuite qu'il est possible d'établir un suivi des performances de l'activité et de créer des alertes pour détecter des signaux faibles, potentiellement révélateurs de défaillances.

5.5.1 Compréhension des performances

Analyse de l'effet du contexte

Le premier cas d'utilisation de l'outil consiste à évaluer l'effet des paramètres de contexte pour favoriser de meilleures pratiques. C'est une alternative aux ANOVA, réalisées lors de l'analyse exploratoire. L'utilisation majoritaire du rapport A pendant la montée permet des montées plus productives en moyenne, ainsi qu'une consommation plus faible (voir Figure C.3 et Figure 5.7). En projetant l'ensemble des montées dans un espace créé à partir du sous-ensemble des montées réalisées avec le rapport A, il est possible d'arriver aux mêmes conclusions d'un seul coup d'œil. Le résultat visuel est présenté avec la Figure 5.11. Les points bleus (montées avec le rapport A) sont centrés sur $\mathbf{0}$. Les points rouges (montées avec le rapport B) sont centrés sur $(0.5, -0.1)$. La consommation est plus élevée, le rapport semble avoir un effet moyen ; la productivité est légèrement plus faible, avec un effet peu important.

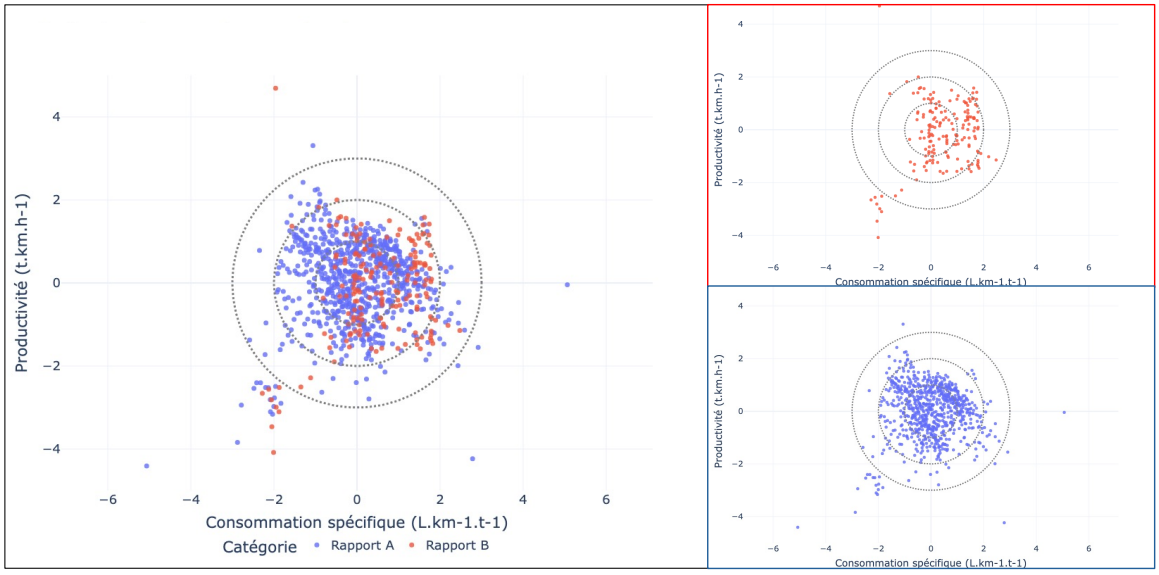

FIGURE 5.11 Projection de toutes les montées dans l'espace créé à partir des montées réalisées avec le rapport A

Le même type d'analyse est réalisé en regroupant les montées selon la charge (Figure 5.12). Ici, le groupe de référence contient les montées réalisées avec une charge dans la benne comprise entre $45t$ et $55t$ (les points rouges). Les points verts représentent les montées peu chargées (moins de $45t$) et les points bleus celles très chargées (plus de $55t$). Les centres moyens de chaque groupe sont présentés dans le Tableau 5.4. Il apparaît clairement que la charge dans la benne influe fortement sur la productivité d'une montée et, de manière plus faible, sur la consommation.

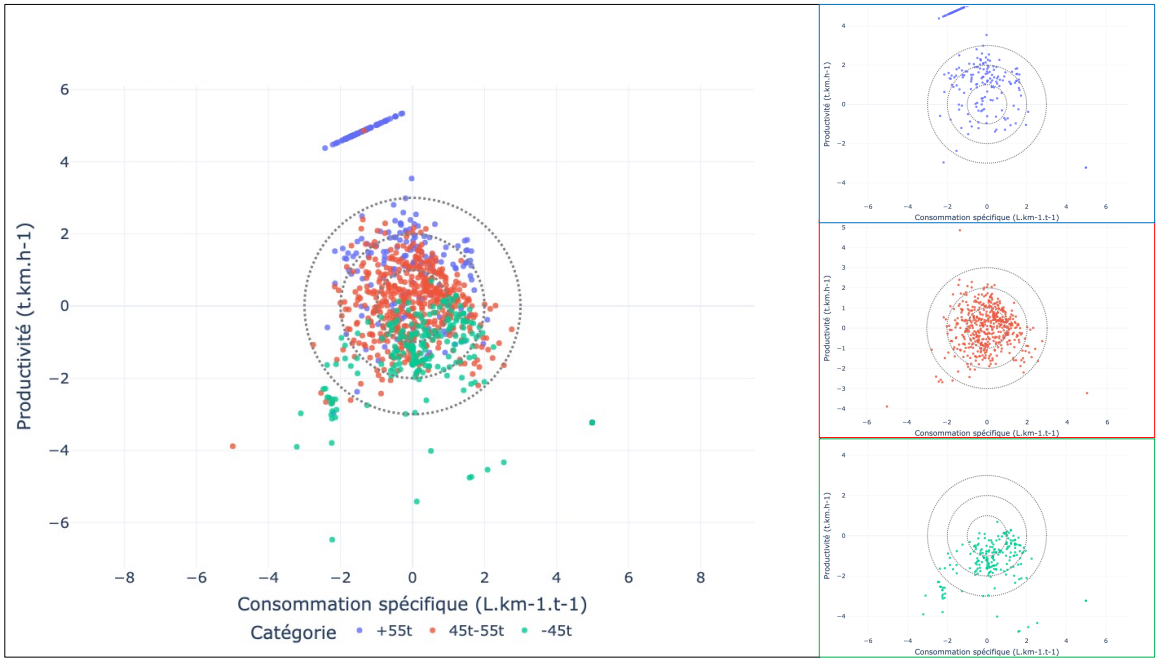

FIGURE 5.12 Projection de toutes les montées dans l'espace créé à partir des montées avec une charge dans la benne entre $45t$ et $55t$

TABLEAU 5.4 Effet de la charge dans la benne sur les mesures de performance dans l'espace transformé

Charge dans la benne	**Productivité spécifique**	**Consommation spécifique**
Plus de 55 t	2,3	−0,56
Entre 45 t et 55 t	0	0
Moins de 45 t	−1,3	0,2

L'effet de l'âge du camion au moment de la montée peut aussi être évalué avec la Figure 5.13. Les montées réalisées par des camions avec moins de $7000h$ de fonctionnement sont utilisées comme référence et sont colorées en rouge. Les points bleus représentent les montées réalisées par des camions avec plus de $7000h$ de fonctionnement. Le nuage de points bleus est assez semblable à la distribution des points rouges, la consommation étant légèrement plus élevée et la productivité plus faible. Comme avec l'ANOVA et les tests post-hoc, il est possible de conclure en disant que l'effet de l'âge est assez faible.

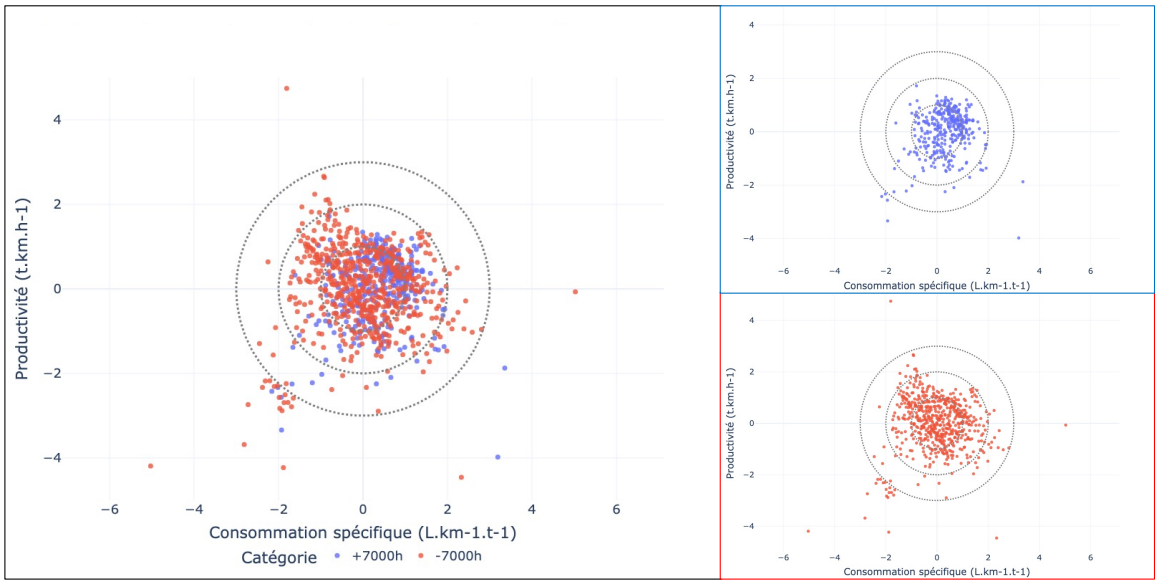

FIGURE 5.13 Projection de toutes les montées dans l'espace créé à partir des montées réalisées par des camion âgés de moins de 7000h

La valorisation des données avec l'outil de support à la performance proposé permet au gestionnaire de flotte de valider et de supporter des recommandations pour la réalisation des montées par les camions. Ainsi, il apparaît que :

- L'utilisation du rapport A lors des montées est plus productive et économe en carburant.
- Il est plus intéressant d'un point de vue production et consommation de réaliser une montée avec un camion chargé au maximum.
- Les camions plus anciens montrent une légère baisse de leur performance, qui reste faible en comparaison avec d'autres paramètres du contexte.

Profil de camion

Un autre cas d'utilisation possible consiste à visualiser les profils des camions dans l'espace créé à partir de toutes les montées. En analysant la distribution des montées pour un seul camion (un seul ID) sur une période donnée, il est possible d'évaluer les camions qui performent le mieux et le moins bien. Les données du mois de mars sont utilisées comme ensemble de référence, puis pour chaque camion, nous traçons les contours de la densité de points (représentant des montées) dans l'espace. Ainsi, chaque camion est associé à son empreinte dans l'espace transformé des performances. La Figure 5.14 présente ces profils. Plusieurs analyses sont possibles ici, par exemple, le camion au profil rouge semble présenter une consommation spécifique plus importante que le camion au profil violet (à partir des pics de densité).

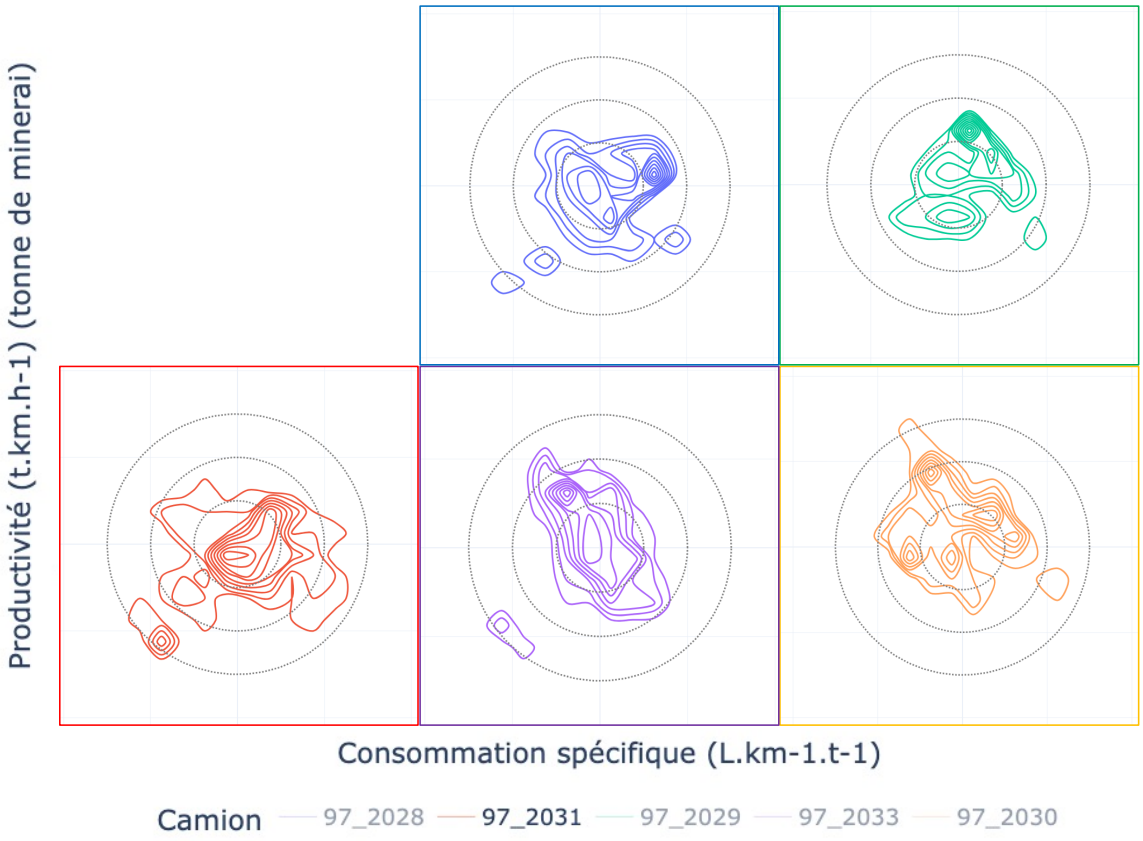

FIGURE 5.14 Projection des densités de points par camion dans l'espace créé à partir de toutes les montées de la période 'mars'

La zone optimale souhaitée pour les montées se situe dans le quart supérieur gauche de l'espace, là où la productivité spécifique est élevée et où la consommation est faible. Le camion au profil vert présente un pic de densité au-dessus du centre de l'espace. Cela signifie par exemple qu'un grand nombre de montées réalisées par ce camion ont été productives.

5.5.2 Détection des performances anormales et alarmes

Comme expliqué dans le Chapitre 4, un seuil de distance est établi pour identifier les montées anormales. Avec un seuil fixé à 95%, un cercle de rayon 2.45 aide à distinguer les montées potentiellement anormales pour une analyse approfondie. La Figure 5.15 illustre ce seuil pour les montées de mars. Certaines montées présentent des performances exceptionnellement bonnes ou mauvaises. Dans la gestion quotidienne de la flotte, ce seuil et cette projection pourraient être utilisés après chaque quart de travail pour examiner les anomalies. Un rapport pourrait ainsi être généré automatiquement et inclure la durée, la distance parcourue, la charge, et d'autres données, facilitant ainsi l'analyse et l'intervention du gestionnaire si nécessaire.

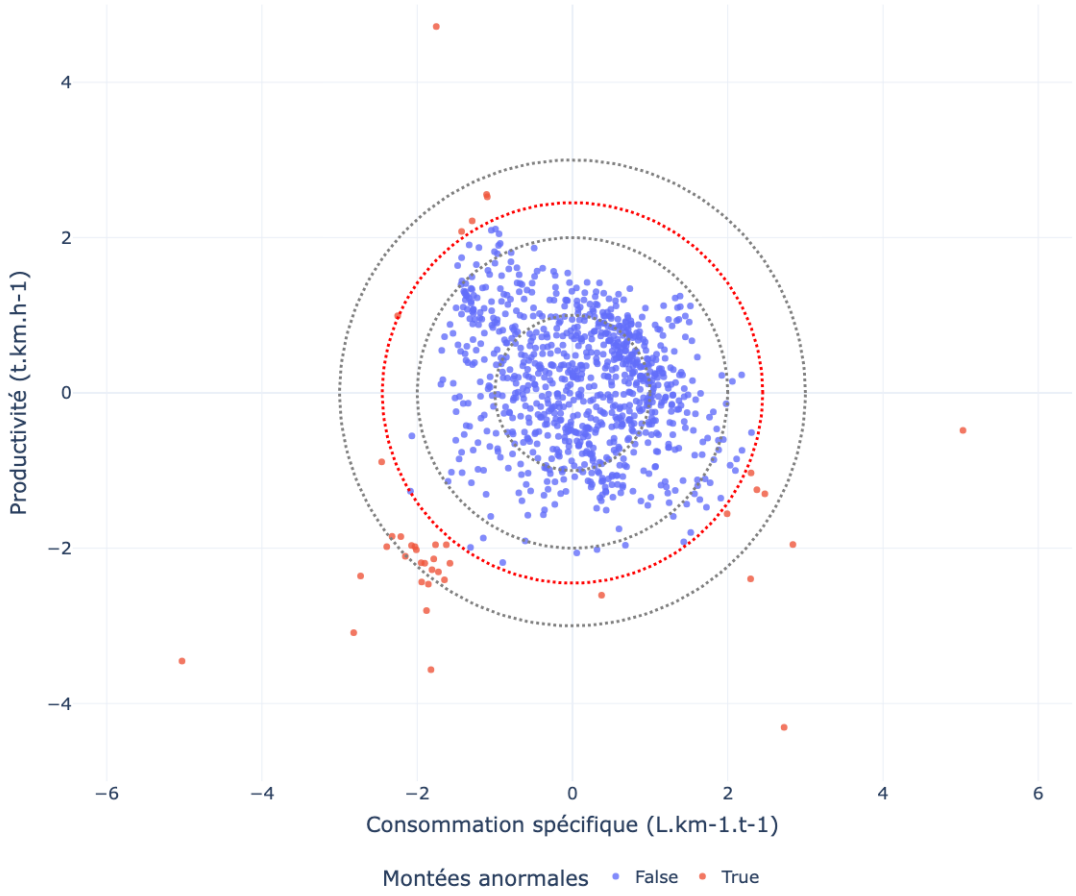

FIGURE 5.15 Projection des montées de la période du mois de mars dans l'espace avec toutes les montées, avec seuil d'anomalies à 95%

Il est également possible de mettre en place des alarmes par camions en cas de baisse de performance dans le temps. En détectant une productivité plus faible ou une consommation plus haute que la normale pendant un certain temps, il devient possible d'alerter les gestionnaires de flottes pour anticiper des potentielles défaillances.

Dans l'espace généré, une zone optimale (en vert dans la Figure 5.16) et une zone sous-optimale sont identifiées. La zone sous-optimale est située dans le quart inférieur droit, caractérisée par une consommation spécifique accrue et une productivité réduite par rapport à la moyenne. Le quart supérieur gauche correspond à la zone optimale.

Pour détecter les baisses de performances, nous disposons de données concernant deux camions qui ont subi une défaillance au mois d'avril. Pour le premier, nous disposons également de données au mois de mars. Chaque groupe de 10 montées consécutives par camions est pro-

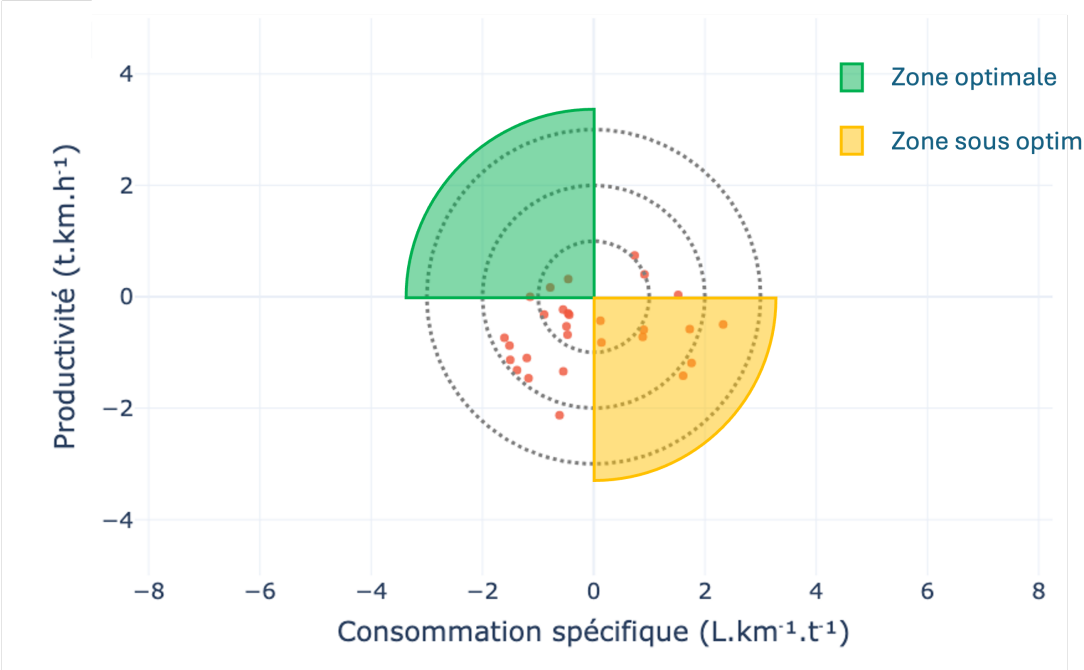

FIGURE 5.16 Zone optimale et zone sous optimale pour la projection

jeté dans l'espace créé à partir de l'historique sur les trois derniers mois pour tous les camions de la flotte. La proportion de montées situées dans la zone sous-optimale est calculée puis enregistrée pour chacun de ces groupes. Cette proportion est ensuite comparée avec deux seuils : un seuil fixe et commun à tous les camions, ici 20%, et un seuil calculé à partir de la moyenne des 5 dernières périodes de 10 montées du camion étudié (ou moins si non disponibles). Si les deux seuils sont dépassés pour trois groupes de 10 montées consécutifs, alors une alerte est émise. Elle indique que les performances du camion observé sont anormales.

Pour le camion 2029, qui subit une défaillance après la dernière montée le 28 avril 2025, une alerte aurait été émise le 26 avril 2025 à 05 :13, après que le seuil fixe et le seuil propre au camion aient été dépassés trois fois consécutivement. Le Tableau 5.5 résume les proportions pour chaque groupe de 10 montées et indique le dépassement des seuils. Plusieurs fois, les seuils sont dépassés une ou deux fois sans déclencher d'alerte. Pour le gestionnaire de flotte, un tableau de bord rassemblant les cinq derniers groupes de 10 montées pourrait être présenté, avec une pastille jaune au centre du graphe pour les groupes dépassant le seuil sans déclencher l'alerte et une pastille rouge pour ceux déclenchant l'alerte. Par exemple, à partir de 5h16 le 26 avril 2025, le tableau de bord du gestionnaire aurait pu ressembler à la Figure 5.17.

TABLEAU 5.5 Résumé des proportions de montées en zone sous-optimale et des seuils pour les groupes de 10 montées du camion 2029 entre mars et avril

Début	Fin	Prop. sous-opt.	Moyenne 5 périodes préc.	Seuil fixe	Seuil camion
28/04/2025 12 :30	28/04/2025 18 :28	0%	32%	–	–
27/04/2025 20 :27	28/04/2025 11 :48	0%	44%	–	–
27/04/2025 06 :50	27/04/2025 19 :05	30%	46%	Dépassé	–
26/04/2025 16 :33	27/04/2025 05 :26	50%	36%	Dépassé	Dépassé
26/04/2025 05 :49	26/04/2025 15 :29	50%	32%	Dépassé	Dépassé
25/04/2025 18 :41	26/04/2025 05 :13	30%	28%	Dépassé	Dépassé
25/04/2025 06 :31	25/04/2025 18 :01	60%	20%	Dépassé	Dépassé
24/04/2025 15 :19	25/04/2025 05 :11	40%	18%	Dépassé	Dépassé
24/04/2025 01 :18	24/04/2025 14 :39	0%	26%	–	–
19/03/2025 00 :23	24/04/2025 00 :10	30%	22%	Dépassé	Dépassé
18/03/2025 01 :12	18/03/2025 17 :03	10%	24%	–	–
14/03/2025 12 :01	18/03/2025 00 :34	20%	25%	–	–
11/03/2025 23 :36	14/03/2025 06 :27	30%	23%	Dépassé	Dépassé
08/03/2025 17 :18	10/03/2025 07 :41	40%	15%	Dépassé	Dépassé
07/03/2025 19 :11	08/03/2025 09 :12	10%	20%	–	–
07/03/2025 04 :08	07/03/2025 18 :10	20%	–	–	–

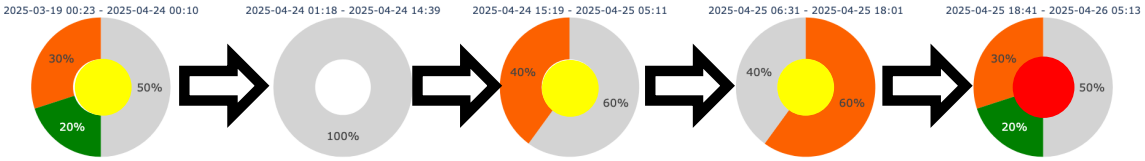

FIGURE 5.17 Aperçu du tableau de bord potentiel pour le camion 2029 le 26 avril, à destination du gestionnaire de flotte

Pour ce camion 2029, une alerte aurait pu être émise avant la défaillance. Pour l'autre camion qui subit une défaillance, nous disposons de moins de données et nous remarquons qu'aucune alerte n'aurait été émise avant la défaillance. Les proportions et le dépassement des seuils pour ce camion sont renseignés dans le Tableau 5.6. Ici, le camion ne dépasse jamais le seuil fixe, ce qui semble indiquer qu'il est plus efficace que le reste de la flotte et ses performances ne se dégradent pas assez pour déclencher une alerte.

Pour les quatre camions restants et en utilisant les mêmes seuils, nous remarquons que l'alerte aurait été levée dans deux cas mais nous ne pouvons pas dire si elles correspondent réellement à des défaillances. Nous ne disposons pas d'autres données labellisées, avec des informations concernant de réelles défaillances qui sont arrivées. Pour poursuivre les tests et ajuster les seuils, le nombre de montées prises en compte, le nombre de périodes avec seuils dépassés avant de déclencher l'alerte, plus de données sont nécessaires.

Tableau 5.6 Résumé des proportions de montées en zone sous-optimale et des seuils pour les groupes de 10 montées du camion 2032 en avril

Début	**Fin**	**Prop. sous-opt.**	**Moyenne 5 périodes avant**	**Seuil fixe**	**Seuil camion**
28/04/2025 19 :00	29/04/2025 15 :59	20%	6%	–	Dépassé
28/04/2025 03 :45	28/04/2025 18 :00	0%	8%	–	–
27/04/2025 08 :14	28/04/2025 02 :29	20%	4%	–	Dépassé
26/04/2025 18 :42	27/04/2025 07 :18	0%	4%	–	–
26/04/2025 03 :13	26/04/2025 16 :53	10%	3%	–	Dépassé
25/04/2025 12 :08	26/04/2025 02 :05	0%	3%	–	–
24/04/2025 18 :55	25/04/2025 08 :34	10%	0%	–	Dépassé
24/04/2025 02 :51	24/04/2025 17 :55	0%	0%	–	–
24/04/2025 00 :58	24/04/2025 01 :55	0%	–	–	–

CHAPITRE 6 CONCLUSION

Ce mémoire présente une méthode pour développer un outil d'amélioration des performances industrielles basé sur les données. Cet outil est destiné à être utilisé dans un contexte opérationnel. Des données de fonctionnement de camions miniers ont été mises à notre disposition par un partenaire industriel qui souhaitait comprendre et trouver des pistes d'amélioration pour augmenter la productivité de ses activités et réduire les coûts.

L'application des différentes étapes de la méthodologie basée sur CRISP-DM [Wirth and Hipp, 2000] (la compréhension du cas industriel, la compréhension des données, la transformation des données, puis la modélisation et la projection dans un espace iso-probabiliste) a permis le développement d'un outil adaptable sous la forme d'une visualisation des itérations de l'activité étudiée selon leurs performances et leurs paramètres de contexte. Cet outil visuel donne lieu à une interprétation rapide pour un gestionnaire opérationnel qui peut effectuer des recommandations ou prendre des décisions soutenues par les données. Il est également possible de mettre en place des alertes lorsque les performances de l'activité baissent afin d'anticiper de potentielles défaillances.

6.1 Avantages et limites

La définition d'indicateurs spécifiques à l'activité étudiée permet de réduire le nombre de variables à analyser et comparer. En agrégeant différentes informations pour obtenir deux indicateurs expliquant la consommation et la productivité de l'activité, il devient plus simple d'appliquer des transformations, de visualiser des résultats et d'effectuer des comparaisons car la dimensionnalité de l'étude est amenée au cas 2D. Ce travail d'ingénierie des attributs (*feature engineering*) est l'une des bases de la méthode développée.

La suite de la méthode proposée est basée sur une transformation de données simple et non paramétrique, nécessitant seulement le choix d'un ensemble de référence. Le résultat final, sous la forme d'une cible avec le point central représentant l'activité moyenne de la référence, est interprétable facilement, selon le principe des cartes de contrôle. Plus un point représentant une itération est éloigné du centre, plus il est anormal, avec des seuils statistiques établis. La position par rapport au centre indique si les performances sont anormalement bonnes ou mauvaises. En choisissant judicieusement l'espace de référence selon un paramètre de contexte ou bien une période de temps, il est possible d'adapter l'utilisation de l'outil au besoin.

La quantité de données à disposition peut être limitante, notamment lors du choix de l'espace de référence pour éviter une compression trop importante des données lors de la phase de normalisation. Cependant, comparativement aux méthodes de prévision et d'apprentissage automatique, la transformation mise en jeu peut fonctionner avec des jeux de données réduits. Trois mois d'historique, dans le cas du partenaire industriel, suffisent à mettre en place une première version de l'outil. L'ajout de plus de données, de plus de contexte et d'informations sur les missions réalisées ou sur les défaillances subies permettrait de meilleurs résultats et des analyses plus poussées.

L'anticipation de défaillances à partir des baisses de performances, qui correspondait au deuxième sous-objectif de recherche, semble prometteuse mais ne peut pas entièrement être validée lors de cette étude. Elle nécessite également plus d'hyperparamètres, qui doivent être sélectionnés selon le cas industriel étudié et validés avec des données historiques.

6.2 Recommandations pour le partenaire industriel

L'application de la méthodologie aux données du partenaire industriel permet de proposer plusieurs recommandations.

- Les analyses réalisées avec les données fournies montrent que l'utilisation du rapport A permet de meilleurs performance (productivité et consommation). Utiliser le camion avec la benne à pleine charge est également plus profitable. Nous recommandons donc que les camions effectuent les montées avec la benne remplie et que le rapport A soit principalement utilisé. Ces recommandations ne prennent pas en compte l'usure ou les pratiques recommandées par le constructeur du camion, ni les contraintes de planification et des chantiers dans la mine.
- Le stockage des données de fonctionnement sur une courte période (trois mois) ne permet pas des analyses historiques poussées. Garder les informations issues de ces données sur une période plus longue permettrait la création d'un historique riche, utile pour de futures analyses. Pour ne pas surcharger les capacités de stockage, les données pourraient être échantillonnées à une fréquence moins élevée que celle d'acquisition (passer d'un enregistrement toutes les $0.5s$ à un enregistrement toutes les $10s$ par exemple). Sauvegarder seulement les informations concernant une activité précise, comme la base de données de montées utilisée ici pourrait également être une solution.
- Lors de cette étude, seules les données de fonctionnement des camions sont utilisées. Il serait intéressant d'élargir les analyses en enrichissant le jeu de données avec des informations concernant les positions des camions, les missions effectuées et les informations concernant la maintenance. L'agrégation de données provenant de plusieurs

sources différentes est l'un des enjeux de l'analytique industriel et de la transition numérique.

6.3 Futures recherches

Dans le futur, la méthodologie proposée pourrait être appliquée dans d'autres domaines. Les activités de machines d'usinage, où la consommation d'énergie et la productivité (quantité de matière usinée) peuvent être mesurées et où la répétition d'une même activité a lieu, pourraient être analysées. Les voyages de camion de transport ou de véhicules effectuant des trajets récurrents et assez similaires pourraient également convenir à l'application de la méthodologie.

La transformation de données présentée met en jeu une sphérisation des données grâce à la matrice de Cholesky. Cette technique a été choisie pour sa simplicité computationnelle, mais d'autres méthodes de sphérisation pourraient être envisagées. De plus, la méthodologie et le cas d'étude se concentrent sur deux mesures de performance. La prise en compte d'indicateurs supplémentaires peut être envisagée. Les enjeux d'interprétabilité et d'affichage pour les acteurs opérationnels seraient alors plus importants. Lors de la réalisation des analyses de variances, d'autres tests statistiques pour valider les hypothèses de normalité pourraient être utilisés à la place d'une validation visuelle, le test de Shapiro–Wilk par exemple.

La détection des baisses de performances et la mise en place d'alarmes nécessitent plus de données labellisées pour être évaluées et comparées avec des méthodes classiques de maintenance basée sur les données. Utiliser les baisses de performances et les performances globales comme entrée dans des modèles pour estimer la durée de vie restante d'équipement pourrait être intéressant.

Enfin, la transformation de données présentée avec une normalisation univariée, puis la sphérisation des données, pourrait être utilisée comme prétraitement avant l'application d'algorithmes de prédiction ou de classification.

Finalement, dans le contexte de l'Industrie 4.0 et des données massives, nous proposons de développer un outil de surveillance de performance adaptable, interprétable et à destination d'utilisateurs opérationnels. De nombreuses entreprises industrielles collectent des données à propos de leurs activités et pourraient bénéficier de la méthodologie proposée comme première étape pour valoriser ces informations, avant de mettre en place des méthodes d'apprentissage automatique et d'intelligence artificielle plus complexes, demandant plus de moyens et de compétences.

RÉFÉRENCES

[AFNOR, 2016] AFNOR (2016). NF X60-000 - maintenance industrielle – fonction maintenance. Norme française.

[Ahmad *et al.*, 2022] Ahmad, N., Hamid, A., and Ahmed, V. (2022). Data science : Hype and reality. *Computer*, 55(2) :95–101.

[Chapman *et al.*, 2000] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). *CRISP-DM 1.0 : Step-by-Step Data Mining Guide*. CRISP-DM Consortium / SPSS. Disponible en ligne : `https://mineracaodedados.wordpress.com/wp-content/uploads/2012/12/crisp-dm-1-0.pdf`.

[Danjou *et al.*, 2017] Danjou, C., Pellerin, R., and Rivest, L. (2017). Le passage au numérique : Industrie 4.0 : des pistes pour aborder l'ère du numérique et de la connectivité. Num Pages : 26 Place : Québec Publisher : Centre francophone d'informatisation des organisations (CEFRIO).

[Dayo-Olupona *et al.*, 2023] Dayo-Olupona, O., Genc, B., Celik, T., and Bada, S. (2023). Adoptable approaches to predictive maintenance in mining industry : An overview. *Resources Policy*, 86 :104291.

[Deming, 1982] Deming, W. E. (1982). *Quality, Productivity and Competitive Position*. The MIT Press, Cambridge, MA.

[Dhillon, 2008] Dhillon, B. (2008). Mining equipment reliability, maintainability, and safety. *Mining Equipment Reliability, Maintainability, and Safety, by B.S. Dhillon. Berlin : Springer, 2008. ISBN : 978-1-84800-287-6.*

[Drath and Horch, 2014] Drath, R. and Horch, A. (2014). Industrie 4.0 : Hit or Hype ? [Industry Forum]. *IEEE Industrial Electronics Magazine*.

[Fayyad *et al.*, 1996] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3) :37–37.

[Fisher, 1925] Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

[Harris *et al.*, 2020] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with numpy. *Nature*, 585(7825) :357–362.

[International Organization for Standardization, 2015] International Organization for Standardization (2015). ISO 9001 :2015 – quality management systems – requirements. Accessed : 2025-07-11.

[Kessy *et al.*, 2018] Kessy, A., Lewin, A., and Strimmer, K. (2018). Optimal whitening and decorrelation. *The American Statistician*, 72(4) :309–314.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86.

[Lebrun and Dutfoy, 2009] Lebrun, R. and Dutfoy, A. (2009). An innovating analysis of the nataf transformation from the copula viewpoint. *Probabilistic Engineering Mechanics - PROBABILISTIC ENG MECH*, 24 :312–320.

[Lööw *et al.*, 2019] Lööw, J., Abrahamsson, L., and Johansson, J. (2019). Mining 4.0—the Impact of New Technology from a Work Place Perspective. *Mining, Metallurgy & Exploration*, 36(4) :701–707.

[MacGregor, 1994] MacGregor, J. F. (1994). Statistical process control of multivariate processes. *IFAC Proceedings Volumes*, 27(2) :427–437.

[Mahalanobis, 1936] Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2 :49–55.

[Mannila, 1996] Mannila, H. (1996). Data mining : machine learning, statistics, and databases. In *Proceedings of 8th International Conference on Scientific and Statistical Data Base Management*, pages 2–9.

[Martin *et al.*, 1998] Martin, E. B., Morris, A. J., and Kiparissides, C. (1998). Multivariate statistical process control and process performance monitoring. *IFAC Proceedings Volumes*, 31(11) :347–356.

[Montgomery, 2017] Montgomery, D. C. (2017). *Design and Analysis of Experiments*. John Wiley & Sons, 9th edition.

[Mukherjee and Marozzi, 2022] Mukherjee, A. and Marozzi, M. (2022). Nonparametric phase-ii control charts for monitoring high-dimensional processes with unknown parameters. *Journal of Quality Technology*, 54(1) :44–64.

[Oliveira-Filho *et al.*, 2024] Oliveira-Filho, A., Zemouri, R., Pelletier, F., and Tahan, A. (2024). System Condition Monitoring Based on a Standardized Latent Space and the Nataf Transform. *IEEE Access*, 12 :32637–32659.

[Pedregosa *et al.*, 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn : Machine learning in python. *Journal of Machine Learning Research*.

[Porter and Heppelmann, 2014] Porter, M. E. and Heppelmann, J. E. (2014). How Smart, Connected Products Are Transforming Competition. *Harvard Business Review*.

[Provost and Fawcett, 2013] Provost, F. and Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1) :51–59.

[PWC, 2024] PWC (2024). Mine 2024 : 21st edition. Technical report, PWC.

[Rüßman *et al.*, 2015] Rüßman, M., Lorenz, M., Gerbert, P., Waldner, M., Engel, P., Harnisch, M., and Justus, J. (2015). Industry 4.0 : The Future of Productivity and Growth in Manufacturing Industries.

[Sandvik Mining and Rock Solutions, 2024] Sandvik Mining and Rock Solutions (2024). Sandvik th663i underground truck. Consulté le 25 juin 2025.

[SAS Institute Inc., 2003] SAS Institute Inc. (2003). *Data Mining Using SAS® Enterprise Miner™ : A Case Study Approach*. SAS Institute Inc., Cary, NC, USA, 2nd edition. Consulté le 26 juin 2025.

[Shimaoka *et al.*, 2024] Shimaoka, A. M., Ferreira, R. C., and Goldman, A. (2024). The evolution of CRISP-DM for data science : Methods, processes and frameworks. *SBC Reviews on Computer Science*, 4(1) :28–43.

[Wetherill and Brown, 1991] Wetherill, W. B. and Brown, D. (1991). *Statistical Process Control for the Process Industries*. Chapman and Hall, London, 3rd edition.

[Wirth and Hipp, 2000] Wirth, R. and Hipp, J. (2000). Crisp-dm : Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*.

[Xue and Qiu, 2021] Xue, L. and Qiu, P. (2021). A nonparametric cusum chart for monitoring multivariate serially correlated processes. *Journal of Quality Technology*, 53(4) :396–409.

ANNEXE A DÉCOMPOSITION DE CHOLESKY DE LA MATRICE DE COVARIANCE

Dans la procédure de sphérisation des données, une étape clé consiste à transformer la matrice de covariance Σ des données centrées en une matrice diagonale par une transformation linéaire. Pour cela, on utilise la décomposition de Cholesky, qui permet d'écrire toute matrice symétrique définie positive $\Sigma \in \mathbb{R}^{n\times n}$ sous la forme :

$$\Sigma = LL^\top$$

où :

- L est une matrice triangulaire inférieure (tous les éléments au-dessus de la diagonale sont nuls),
- $L^\top$ est sa transposée.

Conditions d'existence La décomposition de Cholesky existe si et seulement si :

- Σ est symétrique, i.e. $\Sigma = \Sigma^\top$,
- Σ est définie positive, c'est-à-dire que :

$$\forall x \in \mathbb{R}^n \setminus \{0\}, \quad x^\top \Sigma x > 0$$

Pour une matrice de covariance empirique (calculée sur des données réelles), ces conditions sont généralement vérifiées, à condition que les variables ne soient pas parfaitement colinéaires. En revanche, si la matrice est mal conditionnée (nombre d'observations trop faible, redondance linéaire entre variables), la décomposition peut échouer numériquement.

Calcul de L La décomposition de Cholesky est réalisée à l'aide d'algorithmes numériques efficaces, disponibles dans la plupart des bibliothèques scientifiques (par exemple `numpy.linalg.cholesky()` en Python).

À titre d'exemple, pour une matrice $\Sigma \in \mathbb{R}^{2\times 2}$ telle que :

$$\Sigma = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \quad \text{on cherche} \quad L = \begin{pmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{pmatrix} \quad \text{tel que} \quad \Sigma = LL^\top$$

En développant $LL^\top$, on obtient :

$$LL^\top = \begin{pmatrix} l_{11}^2 & l_{11}l_{21} \\ l_{11}l_{21} & l_{21}^2 + l_{22}^2 \end{pmatrix}$$

Ce qui permet d'identifier l_{11}, l_{21}, l_{22} en fonction de a, b, c.

ANNEXE B LISTE DES CAPTEURS DISPONIBLES

Tableau B.1 Liste des capteurs disponibles

Capteur	**Description**	**Plage**	**Unité**
AccelerationPedal	Pression sur la pédale d'accélération	0–100%	%
BatteryVoltage	Tension aux bornes de la batterie	–	Volts
EngineActualPercentLoad	Charge du moteur en % de la charge maximale	0–100%	%
EngineFuelRate	Consommation en essence du moteur, débit avant injection	–	L/h
EngineSpeed	Vitesse de rotation en sortie du moteur	–	RPM
FuelLevel	Niveau d'essence dans le réservoir	0–100%	%
InstantStandardPayload	Charge instantanée transportée	–	kg
MachineSpeed	Vitesse du camion	–	km/h
TotalDistance	Distance totale couverte par le camion depuis sa mise en fonction	–	km
TotalVehicleRuntime	Temps de fonctionnement total du camion depuis sa mise en fonction	–	s
TransmissionCurrentGear	Rapport de transmission actuellement engagé	A–B	–

ANNEXE C DIAGRAMMES DE TUKEY, RÉSULTAT DES ANOVA

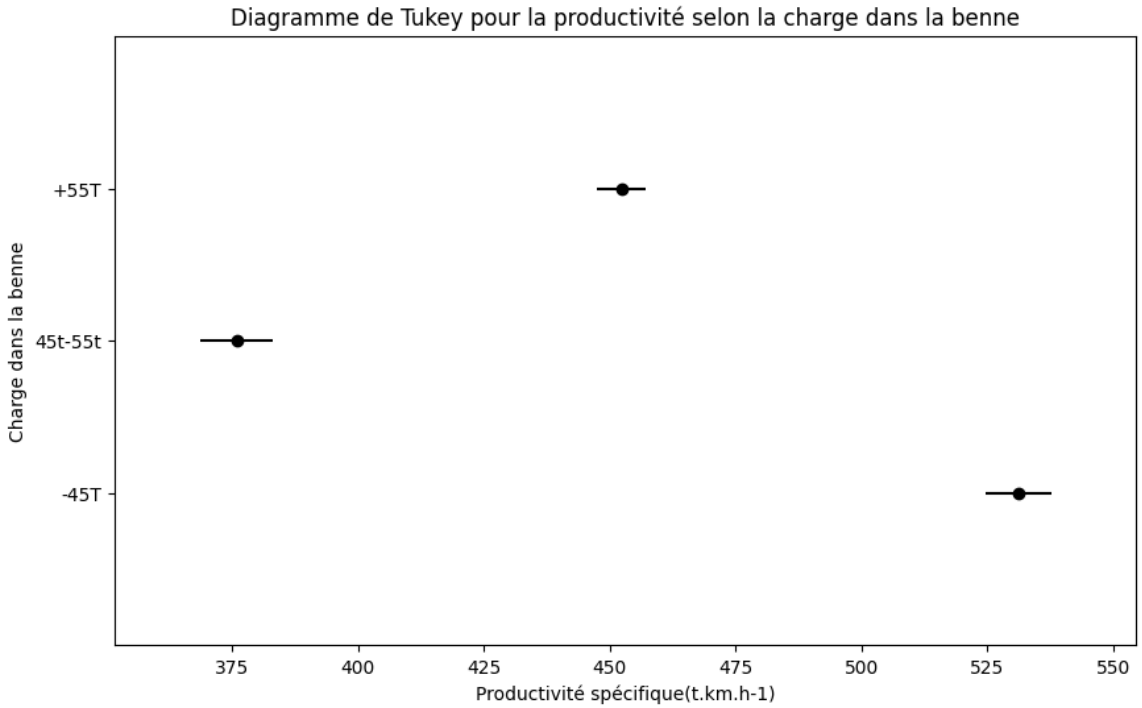

FIGURE C.1 Diagramme de Tukey pour la productivité spécifique selon la charge dans la benne

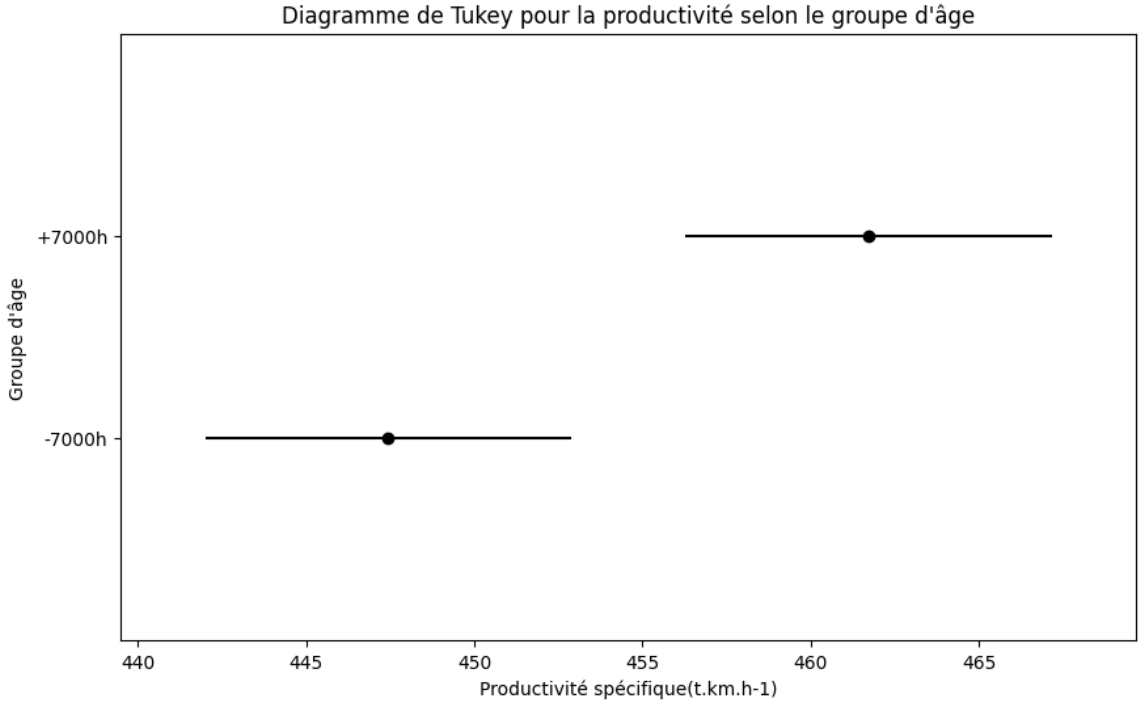

FIGURE C.2 Diagramme de Tukey pour la productivité spécifique selon le groupe d'âge

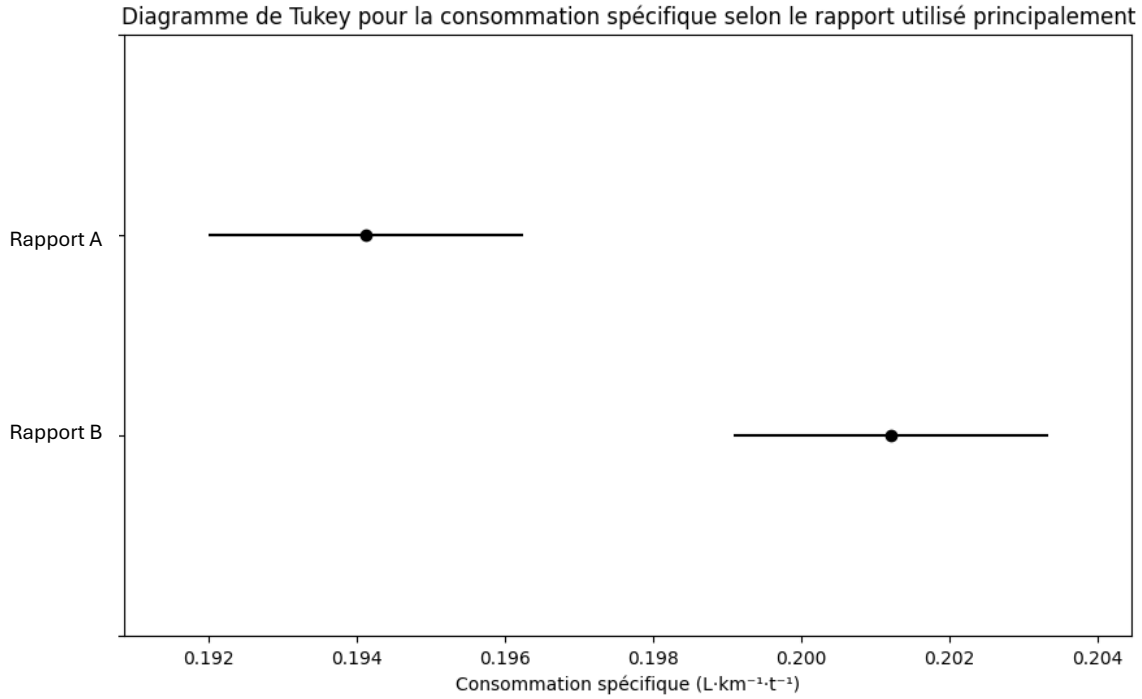

FIGURE C.3 Diagramme de Tukey pour la consommation spécifique selon le rapport utilisé

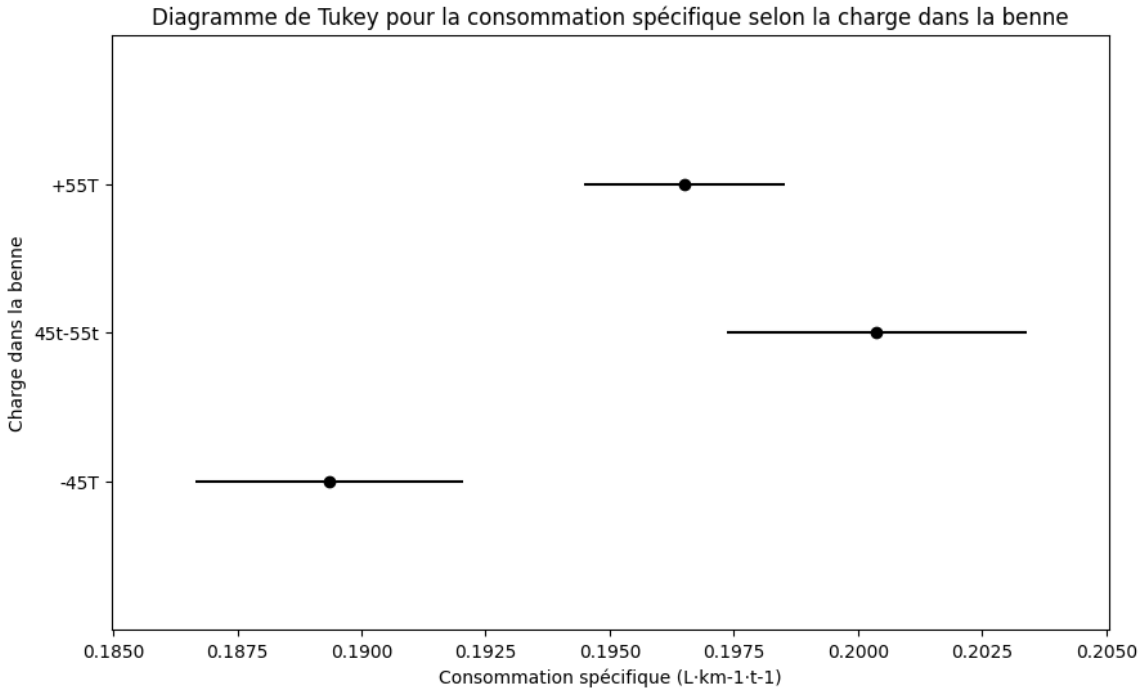

FIGURE C.4 Diagramme de Tukey pour la consommation spécifique selon la charge dans la benne

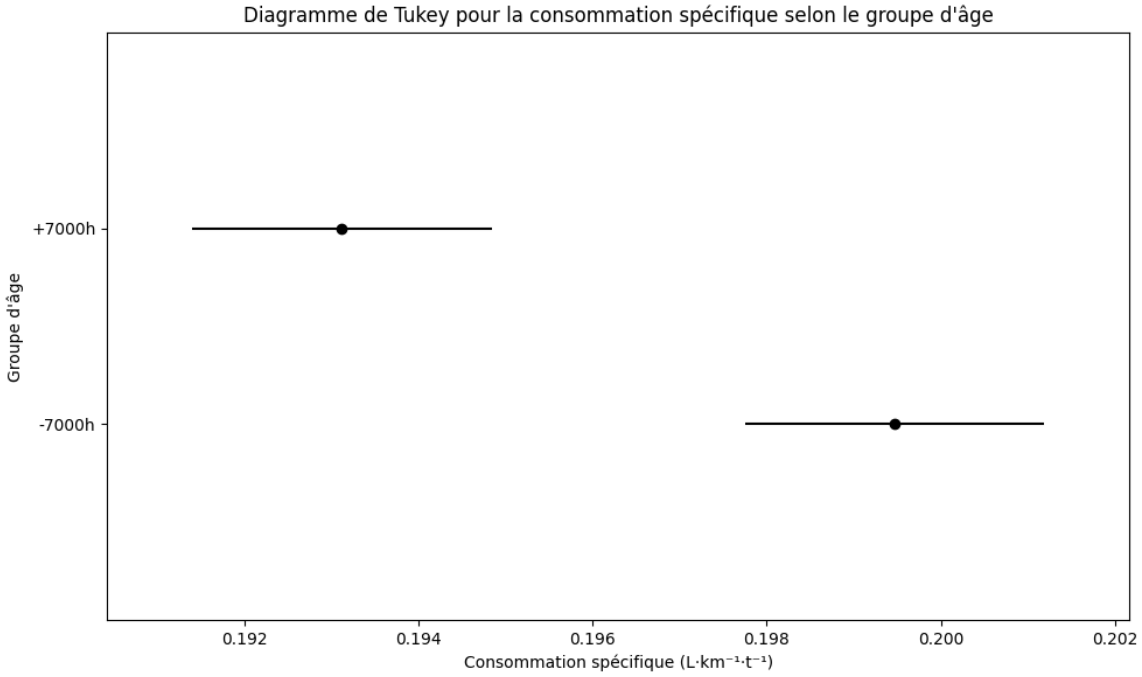

FIGURE C.5 Diagramme de Tukey pour la consommation spécifique selon le groupe d'âge