Titre: Title:	Méthode de valorisation de données de production pour l'évaluation des pertes de cadence
Auteur: Author:	Mathilde Guendon
Date:	2022
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Guendon, M. (2022). Méthode de valorisation de données de production pour l'évaluation des pertes de cadence [Master's thesis, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/10556/

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Méthode de valorisation de données de production pour l'évaluation des pertes de cadence

MATHILDE GUENDON

Département de mathématiques et de génie industriel

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

Génie industriel

Septembre 2022

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Méthode de valorisation de données de production pour l'évaluation des pertes de cadence

présenté par **Mathilde GUENDON**

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

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

Martin TRÉPANIER, président
Robert PELLERIN, membre et directeur de recherche
Camélia DADOUCHI, membre et codirectrice de recherche
Bruno AGARD, membre et codirecteur de recherche
Alexandre LEBLANC-RICHARD, membre

REMERCIEMENTS

Tout d'abord, je tiens à remercier Robert PELLERIN, directeur de recherche, pour son encadrement et ses conseils avisés tout au long de ma maîtrise. Je souhaite remercier Camélia DADOUCHI et Bruno AGARD, codirectrice et codirecteur de recherche pour leur accueil au sein du Laboratoire en Intelligence des Données de Polytechnique Montréal, leur bienveillance et leurs aides précieuses durant cette année de recherche.

Je voudrais exprimer ma reconnaissance à Alexandre LEBLANC-RICHARD pour son aide et son implication dans ce projet, ainsi qu'à Samuel LUPIEN pour sa confiance et à Zacharie ST-PIERRE accompagné de toute l'équipe du service informatique qui ont donnés de leur temps et ont grandement facilité cette collaboration.

Enfin, je souhaite remercier mes nouveaux amis de Montréal, mes amis de France, ma famille et tout particulièrement mes parents, Adrien et mes chers amis du laboratoire, pour leur soutien moral inestimable et leur encouragement.

RÉSUMÉ

La réduction des pertes de cadence en production représente un enjeu majeur pour les entreprises manufacturières. Toujours à la recherche d'optimisation de leur rendement, elles désirent identifier les facteurs qui peuvent lui nuire. Avec le développement de l'Industrie 4.0, la production s'est réinventée, notamment en devenant de plus en plus automatisée. Cela a donc rendu possible la collecte de données toujours plus sophistiquées et dans des quantités toujours plus importantes. La valorisation de données se présente alors comme un processus essentiel pour traiter ces données qui regorgent de connaissances inexploitées. La littérature scientifique témoigne du développement de plusieurs méthodes de valorisation de données ces dernières années, utilisant elles aussi des outils de plus en plus élaborés. Cependant très peu de méthodes proposent une évaluation des pertes de cadence allant jusqu'à identifier les comportements dans un contexte de production où l'automatisation est prédominante.

En collaboration avec une usine manufacturière de l'industrie automobile du grand Montréal, ce projet a été l'occasion de mieux appréhender l'évaluation des pertes de cadence de machines d'assemblage semi-automatiques. En adaptant la démarche CRISP-DM (*Cross-Industry Standard Process for Data Mining*) aux besoins de cette usine, il a été possible de développer une méthodologie de valorisation de données industrielles pour l'évaluation des pertes de cadence dans un contexte de production semi-automatique. Cette méthode, centrée sur les processus de production, s'appuie dans un premier temps sur des outils d'analyse exploratoire permettant de visualiser et d'évaluer l'impact de différentes caractéristiques de production sur la productivité. À la lumière de ces caractéristiques, la méthode se poursuit par de la découverte de processus à partir de données événementielles. La découverte de processus permet alors d'identifier les transitions critiques entre les différentes étapes d'un processus de production au moyen de graphes dirigés. À partir de ces graphes et des données explicitées par l'analyse exploratoire, il sera possible pour les experts du domaine de cibler des pratiques à privilégier pour réduire les pertes de cadence.

Le choix d'utiliser des outils favorisant la visualisation des données permet de garder les experts du domaine dans la boucle d'analyse, considération parfois laissée de côté dans la littérature. Ces outils représentent un support majeur dans la conduite de ce travail et éclairent les résultats obtenus par leurs interprétations. C'est notamment le cas des graphes dirigés qui, d'une part, offrent la possibilité de distinguer des pratiques relatives à des niveaux de productivité différents et, d'autre

part, permettent d’approcher les pertes de cadence en production par une étude des comportements. La valorisation des données du partenaire industriel démontre l’applicabilité de la méthode à un cas concret de production complexe. L’application se concentre sur 3 mois de production d’une machine d’assemblage, mais pourrait être étendue. Le partenaire industriel a pu alors approfondir l’évaluation des pertes de cadence dans sa situation de production, bien que plus de précisions auraient pu être apportées concernant la complexité des processus étudiés. D'une investigation des pertes de cadence passant par l’évaluation du nombre de produits assemblés en moyenne par jour, le partenaire peut désormais identifier des transitions critiques du processus d’assemblage selon différentes conditions de production.

ABSTRACT

The reduction of lost production rates is a major issue for manufacturing companies. Always looking to optimize their performance, they seek to identify the factors that can affect it. With the development of Industry 4.0, production has been reinvented, notably by becoming more and more automated. This has made it possible to collect even more sophisticated data in even greater quantities. Data valorization is therefore an essential tool to process this data, which hides unexploited knowledge. The scientific literature shows the development of several methods of data valorization in recent years, using increasingly sophisticated tools. However, very few methods propose an evaluation of the losses of rate leading to identifying behaviors in a context of production where automation is predominant.

In collaboration with a manufacturing plant in the automotive industry in the greater Montreal area, this project was an opportunity to understand better the evaluation of cycle losses of semi-automatic assembly machines. By adapting the CRISP-DM (*Cross-Industry Standard Process for Data Mining*) approach to the needs of this plant, it was possible to develop a methodology to value industrial data to evaluate reduced speeds in a semi-automatic production context. This method, centered on production processes, is based initially on exploratory analysis tools allowing to visualize and evaluate the impact of various production characteristics on productivity. Considering these characteristics, the method continues with process discovery based on event data. Process discovery identifies the critical transitions between the different steps of a production process using DFGs (*Directly Follows Graphs)*. From these graphs and the data highlighted by the exploratory analysis, it will be possible for domain experts to target practices to be favored in order to reduce cycle time losses.

The choice to use tools promoting data visualization allows us to keep the domain experts in the analysis loop, a consideration that is sometimes overlooked in the literature. These tools represent major support in the conduct of this work and enlighten the results obtained by their interpretations. This is notably the case of DFGs which, on the one hand, offer the possibility to distinguish practices linked to different levels of productivity and, on the other hand, they allow an approach of the losses of production rate studying behaviors. The use of the industrial partner's data demonstrates the applicability of the method to a concrete case of complex production. The application focuses on 3 months of production of a semi-automatic assembly machine but could be

extended. The industrial partner was then able to deepen the evaluation of cycle losses in its production case, although more details could have been provided concerning the complexity of the studied processes and machines. From an investigation of cycle losses based on the evaluation of the number of products assembled on average per day, the partner can now identify problematic transitions of the assembly process according to various production conditions.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

AI	Artificial Intelligence
CRISP-DM	Cross-Industry Standard for Data Mining
DFG	Directly-Follows Graph
IC	Integrated Circuit
KNN	K-Nearest Neighbors
MCD	Minimum Covariant Determinant
PM4Py	Process Mining For Python
RF	Random Forest
SSD	Single Shot Detector
SVM	Support Vector Machines
5M	Méthode, Milieu, Matériel, Matière, Main d'œuvre

CHAPITRE 1 INTRODUCTION

Dans un contexte de production où la main-d'œuvre est déjà manquante, la pandémie a considérablement modifié les manières de travailler. Les entreprises manufacturières doivent alors faire preuve de résilience afin de pouvoir se réinventer et rester compétitives. L'automatisation de certains processus a permis d'assurer une certaine cadence de production en minimisant les pertes et les erreurs. Bien que les usines de production soient de plus en plus automatisées, des pertes de cadence subsistent.

Les équipements sont de plus en plus équipés de capteurs qui donnent accès à des quantités massives de données sur leur fonctionnement. La mise en place de ces capteurs permet de suivre l'état des équipements de production, qui sont aussi des mines d'or d'informations concernant les processus de production (Souza et al., 2021). Aujourd'hui, les techniques d'analyse de ces données se sont beaucoup développées et sont de bons outils d'aide à la décision pour les entreprises (Bhogal et Garg, 2020). Le transfert de connaissances issues de la valorisation des données est, depuis plusieurs années, incontournable pour les entreprises manufacturières désireuses d'améliorer leur productivité (Çiflikli et Kahya-Özyirmidokuz, 2008).

Cependant, automatiser ne veut pas dire faire disparaître la totalité des pertes de cadence. Les apparitions de pannes non anticipées sont à l'origine de pertes de production conséquentes (Bhogal et Garb (2020), Trunzer et al. (2017)). Par ailleurs, il reste impossible d'automatiser les opérations de résolution de problème dans leur intégralité. Le développement de l'Industrie 4.0 a offert l'accès à des données toujours plus sophistiquées (Liang et al. 2021), de plus en plus stockées (Cerquitelli et al., 2020), mais encore peu exploitées. Des outils existent et permettraient d'étudier ces données, car il y a une réelle nécessité d'en extraire la connaissance sous-jacente (Liang et al., 2021) et l'interprétabilité des données est un réel manque à gagner (Alfeo et al., 2020). Cependant, cet effort de valorisation reste encore à être développé pour l'étude de problématiques industrielles. Bien que l'Industrie 4.0 promeuve plus d'automatisation, elle doit rester cohérente avec l'environnement de travail existant (Cerquitelli et al., 2020). Les experts du domaine concerné restent encore peu concertés tandis que l'analyse de données pourrait faciliter leur insertion dans le processus, notamment par la visualisation de données (Campos et al., 2017).

Dans le contexte d'une production semi-automatique, où les processus peuvent parfois s'avérer complexes, et avec une main d'œuvre changeante, il est difficile de correctement cibler les pertes

de cadence et donc d'agir en conséquence. Or, diverses données sont à disposition et l'industrie manufacturière montre de l'intérêt quant à leur exploitation.

En collaboration avec un partenaire manufacturier de l'industrie automobile, ce travail propose une méthodologie de valorisation des données de production dans un contexte de production semi-automatique. L'objectif est de *réduire les pertes de cadence dans un contexte de production semi-automatique.* Cette proposition se démarque par le développement d'une méthodologie qui permettra d'identifier des bonnes pratiques. Les bonnes pratiques sont définies comme les comportements permettant d'éviter les pertes de cadence, à adopter face aux machines de production et inspirées des pratiques existantes. Cette démarche de recherche cible donc des pratiques relatives à différents niveaux de productivité sur des lignes de production semi-automatiques. La performance est alors mesurée à partir du nombre de produits réalisés, mais ne prend pas en considération la qualité de ces produits.

Ce mémoire est constitué de 5 chapitres. Dans un premier temps, une revue de littérature présentera l'avancée actuelle de la recherche concernant la valorisation de données de production quant à la gestion des pertes de cadence. Le chapitre 3 sera l'occasion de présenter les objectifs de ce travail de recherche ainsi que le développement de la méthodologie de valorisation de données de production pour l'évaluation des pertes de cadence dans un contexte de production semi-automatique. Au cours du chapitre 5, cette méthode sera appliquée sur un cas réel avec un partenaire industriel souhaitant réduire les pertes de cadence qu'il rencontre à la suite de l'automatisation d'une partie de ses équipements. Ensuite, il sera question de mettre en perspectives les résultats obtenus avant de conclure dans le dernier chapitre.

CHAPITRE 2 REVUE DE LITTÉRATURE

Dans ce chapitre, il est question de faire un état des lieux des travaux réalisés jusqu'à présent en lien avec cette étude. Avant de procéder à cette revue de littérature, nous définissons dans un premier temps les concepts de base en matière de valorisation de données et de pertes de cadence en production. Puis, nous présentons la stratégie de recherche mise en place pour sélectionner des articles pertinents. Par la suite, ces articles sont analysés afin de mettre en lumière les principales contributions du domaine et d'en souligner leurs limitations en regard du problème spécifique ici étudié.

2.1 Définitions

Afin de garder un œil sur l'évolution de leur efficacité et de pouvoir rester compétitives, les usines modernes de production intègrent à leurs outils de gestion des systèmes de mesure de performance, rendus concrets par l'utilisation d'indicateurs clés de performance (Kang et al., 2015). Ces indicateurs permettent d'évaluer la performance, mais aussi de mettre en lumière les différentes pertes (Lindberg et al., 2015). Aujourd'hui, l'utilisation des indicateurs de performance est normée, mais il existe encore des lacunes concernant leur capacité à mesurer la performance dans un contexte de production continue (Lindberg et al., 2015). Lorsqu'on s'intéresse à la production, les indicateurs les plus fréquemment utilisés sont ceux regroupés sous les concepts de taux de rendement global (« Overall Equipment Effectiveness» en anglais), qui évalue la productivité. Ce dernier découle de la théorie des six grandes pertes de Nakajima dues :

- aux pannes machines et aux changements de séries;
- aux micro-arrêts et aux cadences ralenties; et
- aux défauts de qualité et aux défauts causés lors de la mise en marche des machines.

Elles permettent respectivement de rendre compte de la disponibilité de la production, de son efficacité de la production et de sa qualité (Soltanali et al., 2021).

Dans ce chapitre de revue de littérature, les pertes de cadence feront référence aux micro-arrêts, aux baisses de cadence ainsi qu'aux arrêts non-planifiés ne nécessitant pas de maintenance.

Pour sa part, et d'après l'Office Québécois de la Langue Française (2019), la valorisation de données correspond au « *processus de collecte, de traitement et d'analyse de données, permettant*

l'utilisation optimale de celles-ci dans la poursuite d'un objectif donné ». Dans le cadre de l'exploitation de données industrielles, l'enjeu principal est d'être capable d'en extraire le maximum de connaissances. En effet, selon Dogan et Briant (2021), « *La majorité des problèmes industriels sont riches en données, mais pauvres en savoir* ». En relevant tout type de variables, dans des quantités toujours plus importantes et à des vitesses grandissantes, beaucoup d'entreprises s'inscrivent dans l'ère du *Big Data* (IBM, s.d.). Ces données massives permettent l'étude des processus, de la planification, de la qualité et autres sous-parties de la production. La valorisation de données représente alors un atout majeur pour apporter un soutien aux différentes prises de décision et rester compétitif (Cerquitelli et al. (2021), Dogan et Birant (2021)). Les données peuvent être valorisées par l'utilisation de méthodes d'analyses (*Data Analytics*) permettant de les décrire de manière intelligible. Des méthodes de fouilles de données (*Data Mining)* peuvent aussi être utilisées afin de découvrir des connaissances sous-jacentes au processus étudié, comme l'apprentissage machine (*Machine Learning*), très en vogue lorsque l'objectif est de faire de la prédiction (Dogan et Birant, 2021). Toutes ces technologies sont impliquées, de près ou de loin, dans le développement de l'intelligence artificielle (*Artificial Intelligence, AI*) qui, dans le domaine de la production, vise à améliorer ses performances par l'apprentissage machine à des cas industriels (Lee et al., 2018).

2.2 Démarche de recherche bibliographique

La revue de littérature est réalisée à partir de la base de données scientifiques appelée 'Scopus' (ELSEVIER, s.d.), ou encore 'Google Scholar' (Google, s.d.) offrant une recherche moins rigoureuse, mais permettant d'accéder à une plus grande quantité de documents. Cela a permis de distinguer des documents scientifiques détaillant des méthodes de valorisation de données au service de l'étude des pertes de cadence en production.

Dans un premier temps, une revue de littérature permet d'axer les recherches sur trois notions : la donnée, les pertes de cadence et la production. L'analyse des mots-clés des différents articles obtenus a permis d'étoffer la stratégie de recherche. Les mots-clés utilisés sont recensés dans le tableau ci-dessous.

Tableau 2.1 Mots-clés utilisés pour la revue systématique

Pertes de cadence	Production	Donnée
time loss OR *downtime* OR *interruption* OR *breakdown*	*production* OR *manufacturing*	*big data* OR *data analytics* OR *data mining* OR *AI* OR *machine learning*

À partir de ces résultats, une revue de littérature systématique est réalisée. Dans un premier temps, une restriction aux mots-clés dans le titre et le résumé est choisie, ainsi que des publications ultérieures à 2010. Avec cette première approche, 248 documents sont obtenus. Par la suite, seuls les articles de journaux scientifiques, les articles de conférence ainsi que les chapitres de livres sont conservés. Les domaines présentant des enjeux éloignés de ceux des pertes de cadence en production sont jugés non pertinents et sont donc retirés. Il en est de même pour les documents se concentrant sur la maintenance prédictive, afin de se concentrer sur une approche différente des pertes de cadence en production. Les articles rédigés dans une autre langue que l'anglais sont retirés. Après l'application de ces différents filtres, il reste 107 documents. À partir des articles restants, une lecture des résumés et des introductions permet la suppression de 74 documents, jugés non pertinents par leur domaine d'application ou leurs objectifs. Il reste alors 33 articles, puis 2 articles requérant des accès qui n'ont pas pu être obtenus ont aussi été écartés. Après une lecture plus approfondie, il reste finalement 16 articles. Les articles retenus ciblent alors différents types de pertes de cadence en production et les solutions qui ont été apportées grâce au traitement de données.

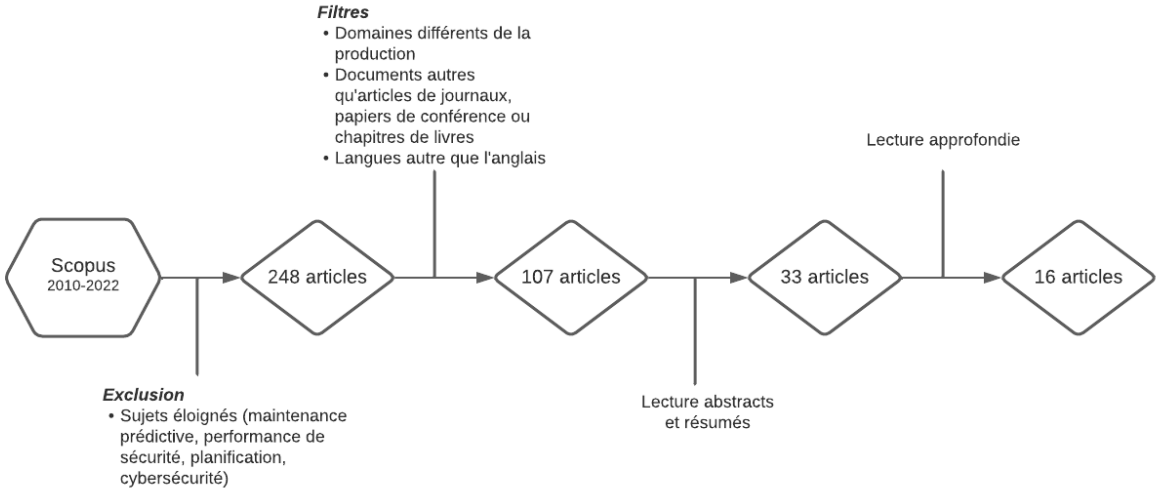


Figure 2.1 Démarche de recherche

Tableau 2.2 Articles retenus

	Auteurs	**Articles**
1	Ahmed et al. (2015)	System availability enhancement using computational intelligence-based decision tree predictive model
2	Alfeo et al. (2020)	Using an autoencoder in the design of an anomaly detector for smart manufacturing
3	Bhogal et Garg (2020)	Anomaly Detection and Fault Prediction of Breakdown to Repair Process Using Mining Techniques
4	Bouché et Zanni-Merk (2011)	Improving the performance of production lines with an expert system using a stochastic approach
5	Cerquitelli et al. (2021)	Manufacturing as a Data-Driven Practice: Methodologies, Technologies, and Tools

Tableau 2.2 Articles retenus (suite et fin)

6	Chen et al. (2019)	A Data Mining Approach for Optimizing Manufacturing Parameters of Wire Bonding Process in IC Packaging Industry and Empirical Study
7	Çiflikli et Kahya-Özyirmidokuz (2010)	Implementing a data mining solution for enhancing carpet manufacturing productivity
8	Dagnigo (2019)	Data Mining Methods to Analyze Alarm Logs in IoT Process Control Systems
9	Denkena et al. (2020)	Continuous modelling of machine tool failure durations for improved production scheduling
10	Elangovan et al. (2015)	Condition monitoring of a valve in a reciprocating compressor using machine learning approach
11	Haasbroek et al. (2018)	Fault Diagnosis for an Industrial High Pressure Leaching Process with a Monitoring Dashboard
12	Hrcka et al. (2017)	Using text mining methods for analysis of production data in automotive industry
13	Klaeger et al. (2020)	Applying SSD to Real World Food Production Environments
14	Liang et al. (2021)	Industrial time series determinative anomaly detection based on constraint hypergraph
15	Pabolu et al. (2022)	A Dynamic System to Predict an Assembly Line Worker's Comfortable Work-Duration Time by Using the Machine Learning Technique
16	Trunzer et al. (2018)	Failure mode classification for control valves for supporting data-driven fault detection

Afin de mieux appréhender les informations contenues dans ces différents documents scientifiques, ces derniers ont été classés selon le type de données qu'ils traitaient ainsi que les méthodes utilisées pour les valoriser.

2.3 Étude des pertes de cadence par la valorisation de données

Dans un premier temps, les documents scientifiques sélectionnés sont présentés selon les objectifs à atteindre par la valorisation de données de production dans un but d'évaluation des pertes de cadence.

2.3.1 Détection d'anomalies

Dans le contexte de la détection d'anomalies, plusieurs chercheurs ont traité différents types de données dans différents contextes. Diverses méthodologies de détection d'anomalies exploitent des séries temporelles. C'est le cas par exemple de Liang et al. (2021) qui proposent une méthode basée sur des hypergraphes de contraintes pour faire détecter des anomalies. Cette méthode est appliquée à trois contextes industriels différents, à savoir pour un système de contrôle de température dans une centrale électrique, une usine chimique ainsi qu'un système hydraulique complexe. Elle permet de prendre en compte les différents types de contraintes, comme celles imposées par les connaissances du domaine d'application, et cela permet aussi de conserver un temps de calcul correct. À partir de données relevées par des capteurs de contrôle opérationnel et de surveillance de processus dans une usine de pétrochimique de styrène, Souza et al. (2021) proposent une méthodologie de détection d'anomalies comme suit : un réseau de neurones convolutif avec une architecture d'auto-encodeur traitent des séries temporelles pour identifier les défauts. L'objectif est de fournir des pistes d'interventions pour cibler les anomalies entraînant des pertes de cadence et les éviter. De la même façon, Alfeo et al. (2020) exploitent des séries temporelles dans différents contextes industriels. Leur méthodologie consiste cette fois-ci en l'association d'un réseau de neurones profond ayant l'architecture d'un auto-encodeur, permettant d'assigner un score d'anomalies à une occurrence, puis d'un discriminateur capable de distinguer une anomalie d'une occurrence normale. Quant à eux, Bhogal et Garg (2020) développent leur méthodologie à partir de données événementielles d'une usine manufacturière. Ils utilisent des outils de *Process Mining* afin de faire de la détection d'anomalies lors des pannes machine et du processus de réparation. Leur objectif est de réduire ces pannes et d'améliorer le processus de réparation des machines en

observant les différents comportements adoptés. Bouché et Zanni-Merk (2010) s'intéressent aussi aux comportements qui peuvent causer les différentes pertes de cadence sur ligne de production. Ils accordent une importance particulière aux connaissances des experts du domaine. L'analyse de données temporelles permet dans un premier temps d'étudier la distribution des durées de pannes puis de faire de la classification.

Dans une optique de remonter à l'origine des interruptions sur une ligne de conditionnement alimentaire, Klaeger et al. (2019) réalisent de la détection d'anomalies par la méthode du *Minimum Covariant Determinant* (MCD). L'objectif est d'assister la production en détectant les interruptions qui ne peuvent pas être relevées manuellement. Cela permet par la suite d'entraîner un algorithme classification grâce à la méthode *Random Forest (RF)*. Dans la même volonté de classification, Chen et al. (2019) analysent des données relatives au conditionnement de circuits intégrés avec pour objectif de réduire les pannes dans le processus de soudure de fils. Ils traitent alors des données temporelles et mettent en œuvre les algorithmes RF et *Extreme Gradient Boosting* pour classer les occurrences défectueuses. D'autres travaux mettent en pratique des méthodologies s'appuyant sur la production d'arbre. Ahmed et al. (2015) s'appuient sur un modèle d'arbre de décision pour détecter des anomalies au niveau de machines, à partir de données relatives aux vibrations de celles-ci. Elangovan et al. (2015) étudient les pertes de cadence dues à l'apparition d'anomalies lors du fonctionnement de compresseurs à pistons. Ils se concentrent notamment sur les valves, détectent les anomalies qui se produisent et les différents états des valves sont classés, avec une meilleure précision pour la méthode RF. Haasbroek et al. (2018) présentent un cheminement menant à l'identification d'anomalies dans l'affinage industriel de métaux de base. Ils utilisent ici une méthode statistique. L'analyse des composantes principales (*Principal Component Analysis)* et l'utilisation de méthodes de contribution permettent de distinguer les variables qui sont à l'origine des défauts, notamment l'étranglement et les fuites des valves d'alimentation du système de lixiviation à haute pression. Dans la production de tapis, l'utilisation de données qualitatives ainsi que de données temporelles permet de recenser la structure des différents arrêts machine, allant des causes jusqu'à la durée de l'arrêt (Çiflikli & Kahya-Özyirmidokuz, 2010). La détection d'anomalies est utilisée ici pour ne conserver que les pannes machines qui font du sens afin de pouvoir les étudier.

Généralement, les méthodologies présentant des outils de détections d'anomalies approfondissent leur travail en proposant différentes façons de comprendre ces anomalies afin de pouvoir les traiter

par la suite. D'autres travaux abordent aussi le problème des pertes de cadence en production sous un angle différent de la détection d'anomalies.

2.3.2 Extraction de modèles, analyse de causes racines et autres approches des pertes de cadence par la valorisation de données

Certains travaux vont au-delà de la détection d'anomalie en identifiant des modèles d'anomalies (Liang et al., 2021). Dans le développement de leur méthodologie de *Process Mining*, Bhogal et Garg (2020) analysent les performances de production et les différents flux afin, eux aussi, d'identifier des motifs de pannes et de générer de la connaissance pour réduire les pertes de cadence engendrer par les pannes et le processus de réparation. Après avoir fait de la détection d'anomalies, Bouche et Zanni-Merk (2010) poursuivent leur méthodologie en construisant aussi des modèles de pannes, mais en suivant une démarche différente. Grâce à la combinaison des modèles de Poisson et des chaînes de Markov, ils créent un arbre permettant la visualisation des relations séquentielles les plus probables menant à une panne en particulier. L'objectif est d'éviter une panne majeure le plus tôt possible en anticipant une succession des événements d'une séquence révélée par cet arbre.

Aussi, certains travaux traitent le sujet des pertes de cadence en production en faisant seulement de l'extraction de modèles. En effet, Dagnino (2019) réalise de la fouille de motifs séquentiels et analyse des alarmes horodatées au sein de systèmes de contrôles de processus. Grâce à l'extraction de motifs d'alarmes au sein de séquences, il vise la prédiction de l'apparition d'une panne sur une machine afin de pouvoir les anticiper. Afin de détecter les modes de défaillance de valves de contrôles régulées en pression, Trunzer et al. (2017) développent une méthodologie qui place au centre la formalisation des connaissances du domaine d'application. Elle consiste à transformer une classification manuelle des différents modes de pannes en classification automatique à partir de facteurs qui permettent de les distinguer, sélectionnés par les experts du domaine.

Par ailleurs, d'autres travaux mettent en lumière l'analyse des causes racines des anomalies détectées. Selon Klaeger et al. (2017), les interruptions fréquentes sur ligne de production sont majoritairement dues à une mauvaise compréhension de la source du problème. Ainsi, à la suite du développement d'un auto-encodeur pour détecter des anomalies, Souza et al. (2021) analysent les causes racines grâce à l'outil de classification RF, qui permet de remonter jusqu'aux origines des anomalies détectées. Afin de donner suite à l'analyse des principales composantes, Haasbroek et

al. (2018) développent des graphes comprenant des connaissances des ingénieurs du domaine et permettant aussi d'identifier les variables qui seraient à l'origine d'une erreur. Quant à eux, Çiflikli & Kahya-Özyirmidokuz, (2010) développent dans leur méthodologie une étude des causes racines grâce à l'utilisation de l'algorithme d'arbres de décision C4.5. Il permet d'expliquer le cheminement vers une panne par des règles logiques.

D'autres recherches présentent des méthodologies différentes pour appréhender les pertes de cadences en production. Dans le cadre de son projet, Pabolu et al. (2022) collectent un ensemble de données relatives au processus dans son intégralité : conditions autour du poste de travail, données relatives aux opérateurs, complexité et intensité de la tâche. Ils développent alors une méthode de prédiction du temps d'opération le plus confortable pour un opérateur sur ligne afin d'optimiser ses conditions de travail et donc de minimiser les pertes de cadence dues aux rotations de personnel et changement de série. Divers modèles basés sur l'apprentissage supervisé (*KNN, SVM, Logistic Regression, Linear or Polynomial Discriminant Analysis* et des méthodes basées sur des arbres) sont comparés en fonction de la précision de la prédiction. Dans la même idée, la fouille de données textuelles est la méthode choisie par Hrcka et al. (2017) afin de recueillir de l'information sur ce qu'il se passe réellement sur une ligne de production, notamment au moment d'une panne, à partir de formulaires remplis directement par les opérateurs. Ils recueillent des données textuelles auprès des intervenants présents sur la ligne de production, concernant la structure des micro-arrêts. À la suite de la classification d'anomalies, Chen et al. (2019) s'intéressent aussi au temps d'opération. Ils utilisent une méthode statistique de régression *(Multivariate Adaptative Regression Splin*) afin de modéliser le temps de soudure puis entraîne un algorithme génétique de choisir le meilleur ensemble de paramètre menant à une occurrence non défectueuse et un temps de soudure optimal. Quant à eux, Denkena et al. (2020) s'appuient sur des séries temporelles et se concentrent sur la distribution des durées de pannes des machines-outils. Avant de réaliser de la détection d'anomalies, Çiflikli & Kahya-Özyirmidokuz, (2010) étudient la pertinence des différents attributs impliqués dans leur travail, beaucoup d'entre eux provenant de données qualitatives. Par ailleurs, ils se concentrent sur la relation entre les attributs conservés avec une analyse de corrélation et de régression.

Ainsi, la valorisation de données autour pertes de cadence sur ligne de production se concentrent majoritairement sur la détection et la classification d'anomalies. Les méthodologies peuvent s'arrêter là ou être approfondies afin d'extraire de la connaissance des anomalies relevées.

2.4 Revue critique

Les résultats présentés ci-dessus mettent en lumière de nombreuses méthodes de détection d'anomalies pour traiter le problème de réduction des pertes de cadence, mais très peu d'analyses causales sont réalisées. La détection d'anomalies est une méthode non supervisée. Elle permet de traiter des données non labellisées et permet d'avoir accès à des connaissances inhérentes aux données initiales. Çiflikli & Kahya-Özyirmidokuz (2010) soulèvent la possibilité d'identifier à la fois des comportements normaux et des points aberrants qui ne correspondent à aucun modèle préalablement déterminé. Par ailleurs, les résultats soulignent le manque d'applications de ces méthodes à des cas concrets avec une quantité importante de données (Haasbroek et al., 2018). Il est rappelé ici que l'objectif du mémoire est de réduire les pertes de cadence persistantes en définissant, d'une part, les facteurs influant la production et, d'autre part, en développant une démarche d'analyse des étapes critiques du processus de production. Cette proposition se démarque par le développement d'une méthodologie qui permettra d'identifier des bonnes pratiques. Les bonnes pratiques étant définies comme les comportements permettant d'éviter les pertes de cadence, à adopter face aux machines de production et inspirées des pratiques existantes. Au regard de cette problématique, les documents présentés ci-dessus sont critiqués.

Alfeo et al. (2020) évoquent trois types de façon de faire de la détection d'anomalies en production, dépendamment de l'emphase mise sur les connaissances du domaine. Les méthodes dites statistiques sont détachées des connaissances métiers et n'évaluent les anomalies que par des différences statistiques avec les événements normaux. Les méthodes basées essentiellement sur les connaissances du domaine permettent de définir des modèles d'anomalies. Enfin, les approches dites « phénoménologiques » interprètent les statistiques au regard des connaissances métiers. Selon Alfeo et al. (2020), pas assez d'importance n'est accordée à la connaissance métier alors que cela permet d'avoir de meilleurs résultats. Dans cette revue, quelques travaux prennent en considération la connaissance métier. Tandis que Liang et al. (2021) développent même une méthodologie essentiellement basée sur ces connaissances, Haasbroek et al. (2018) les utilisent, mais ne se procurent pas d'informations auprès des intervenants les plus proches des processus étudiés. Trunzer et al. (2017) montrent aussi de l'intérêt quant aux connaissances du domaine et les formalisent afin de savoir quels sont les éléments qui font d'un événement, une anomalie. Les anomalies sont alors détectées à partir de caractéristiques sélectionnées grâce aux connaissances

métiers puis classées dans des modes prédéfinis. Cette méthode est développée dans le but d'avoir une classification fiable et en accord avec la réalité de la production pour faire, par la suite, de la prédiction de fautes. Cependant, cette approche peut laisser passer des anomalies elles, car ne correspondent pas aux caractéristiques sélectionnées par les experts. Il en est de même pour Denkena et al. (2019), qui présentent une méthode capable de gérer quelques classes de durées de panne qui doivent être très différenciables. De plus, Denkena et al. (2019) ne s'intéressent qu'à un type d'erreurs et ne prennent pas en compte du processus général. Leur méthode est donc questionnable pour les anomalies peu fréquentes ayant des caractéristiques d'identification peut être différentes. Ainsi, l'apport de connaissance des experts doit pouvoir guider le travail et permettre l'interprétation des résultats à la fin des analyses. En cours d'analyses, il doit permettre de clarifier certains points, mais il y a un risque de biaiser l'analyse prématurément.

D'après Cerquitelli et al. (2020), la valorisation de données de production prend tout son sens dès lors que des actions sont mises en place pour donner suite à l'analyse de données. Par ailleurs, Liang et al. (2020) évoquent la difficulté à quantifier l'impact de chaque anomalie détectée. Cela peut poser problème dans la priorisation des actions à mener. Cependant, dans plusieurs travaux de cette revue, les retours concernant les actions qui pourraient être développées ou corrigées ne sont pas toujours clairs. Haasbroek et al. (2018) proposent par le biais de méthodes de contribution de remonter aux variables à l'origine des anomalies. Cependant, cela ne permet pas de faire un retour sur les pratiques existantes. Liang et al. (2020) ne prennent pas en compte les pratiques menées sur les machines étudiées et préfèrent se concentrer sur les connaissances d'experts hors-ligne de production. Chen et al. (2019) montrent une volonté de réduire les pertes de cadence en optimisant le choix des paramètres de l'outil utilisé. Cependant, il n'est jamais question de pratiques et ils développent une méthodologie sans garantie que cela n'impactera pas la qualité. Par ailleurs, leur travail est réalisé en collaboration avec un partenaire industriel, mais ils disposent de peu de données, qui plus est, ne sont pas récentes. Klaeger et al. (2019) présentent une méthode de détection d'anomalies permettant d'assister la classification manuelle existante. Lorsque la cadence de production est importante, les anomalies sont classées automatiquement et les occurrences qui auraient été mal classées sont corrigées. Cependant, cela ne permet pas d'améliorer les pratiques, juste de les substituer ou de corriger les fautes a posteriori. De plus, ils soulignent que cette la méthode n'est pas applicable à toutes les bases de données notamment dans le cas de processus discrets avec peu d'anomalies.

D'autres travaux ne considèrent pas directement les pertes de cadence. En effet, Alfeo et al. (2020) étudient les anomalies de détection dans un but de détecter la détérioration d'un équipement, donc relatives à la maintenance, mais pas pour réduire les pertes de cadence non liées à la machine. Ils démontrent cependant un intérêt pour la compréhension de la panne. Souza et al. (2021) sélectionnent les caractéristiques les plus pertinentes pour faire de la détection d'anomalies afin d'appuyer, dans le futur, une méthode de prédiction de pannes de la machine pour de la maintenance prédictive. Cette étude comprend alors tous les types de pannes, même celles dépendantes de la machine. Il est à souligner que les travaux de détection d'erreurs représentent des analyses descriptives et sont souvent des bases pertinentes pour faire de la prédiction (Cerquitelli et al., 2021). Pabolu et al. (2022) développent d'un modèle théorique dans un premier temps avant de passer à la pratique, ils disposent de peu de données pour commencer l'étude. Qui plus est, cette étude reste très dépendante des capteurs et demande une grande implication de la part des opérateurs. Elle ne permet pas d'évaluer directement les pertes de cadence, mais elle offre une approche de l'impact de l'environnement de production sur la productivité. Cependant, ce travail ne propose pas de notions d'amélioration des pratiques, mais plutôt d'investiguer les conditions de travail qui optimisent les comportements de production. Quant à Dagnino (2019), il propose lui aussi une approche différente des pertes de cadence. Il étudie des alarmes qui annoncent une erreur dans un but de les gérer d'une meilleure façon. Cette démarche pour aborder les pertes de cadence reste pertinente, mais ne permet pas réellement de faire un état de la productivité et des pratiques qui peuvent mener à une panne, puisque le but est d'intervenir avant l'occurrence de la panne à la suite d'une alerte. Dans leur cas, Bhogal et Garg (2020) s'intéressent à ce qu'il se passe à partir de la panne, notamment au processus de réparation, mais ce qui a mené à la panne n'est pas étudié.

Certains travaux proposent une approche simplifiée de l'étude des pertes de cadence en production. Klaeger et al. (2019) soulignent par exemple que le simple fait de considérer un arrêt pour une anomalie ne permet pas de prendre en compte tous les scénarios possibles, puisque dans certains cas, un arrêt peut être prévu. Bhogal et Garg (2020) exploitent des données collectées manuellement, donc en faible quantité. Ainsi, ils n'ont pas de problèmes de choix de variables, de paramètres ou bien de filtrage. Par ailleurs, ils appuient leur analyse sur des graphes dirigés, ne considérant alors que la durée des activités ainsi que les temps entre les différentes activités comme indicateurs de performance. Cette approche de la performance se retrouve dans le travail de Bouché

et Zanni-Merk (2020). La seule mesure de la performance est celle du nombre de bonnes pièces produites en un certain temps. Avec l'obtention de règles représentant des successions d'événements menant à une faute, leur méthode permettra de faire de la prévention pour éviter d'atteindre cette faute grâce à une alarme. Cependant, il n'est pas précisé quelles sont les comportements liés à une faute afin de corriger les pratiques qui y mènent. Hrcka et al. (2017) présentent une démarche d'identification des conditions dans lesquels se produit une panne par traitement de données textuelles. Cette démarche est incomplète dans un objectif de réduire les pertes de cadence. Elle offre une première investigation des pannes, mais ne fournit pas beaucoup d'informations pertinentes. Une condition est identifiée, mais aucune proposition n'est faite pour l'améliorer ou la corriger.

2.5 Conclusion

D'après cet état de l'art, l'analyse des pertes de cadence par des outils de valorisation de données semble être une problématique étudiée. En effet, il est majoritairement question de détections d'anomalies rendues possibles, soit par des méthodes semi ou non supervisées, soit par des techniques de classification. Il est aussi question d'extraction de modèles d'anomalies ou d'analyse causale des origines de ces anomalies.

Cependant, un manque est à combler dans l'étude des processus et des pratiques à mettre en place ou à améliorer pour mener à bien le processus étudié. Par ailleurs, certains travaux sont développés à partir d'une faible quantité de données, et d'autres utilisent, à des degrés différents, les connaissances des experts du domaine auquel s'applique leur projet. Un autre point à combler, majeur pour des questions d'interprétabilité, est la visualisation des données. Elle permet notamment d'insérer les experts du domaine dans le cycle d'analyse et à terme serait vouée à offrir un guidage personnalisé de la production et à être plus ciblées (Cerquitelli et al., 2020). Ce sont ces deux manquements que ce travail tentera de combler.

CHAPITRE 3 MÉTHODOLOGIE

3.1 Objectif de recherche

L'objectif principal de cette étude est de *réduire les pertes de cadence dans un contexte de production semi-automatique*. Pour l'atteindre, ce travail propose une méthodologie de valorisation de données de production afin d'évaluer les pertes de cadence dans ce contexte de production. Les sous-objectifs de cette méthodologie sont les suivants :

- **identifier et préparer les données pertinentes de production** : après une phase de prise en main des données, elles doivent être traitées afin de refléter la réalité du contexte industriel et être exploitables pour les analyses qui suivent;
- **identifier les facteurs influant sur la production**: le but est d'analyser les données de production, et d'en identifier les facteurs qui peuvent influer sur la production et de quantifier leur impact sur la cadence de production; et
- **déterminer les étapes critiques du processus étudié** : il s'agit d'approfondir la phase d'analyses des données. Cette étape se concentre sur les événements intervenant au cours du processus de production et sur les réactions qu'ils engendrent.

3.2 Méthodologie générale de recherche

Le choix méthodologique de cette première partie s'inspire la méthode *Cross-Industry Standard Process for Data Mining* (CRISP-DM) et sera adaptée à la problématique de cette étude. En effet, selon Schröer et al. (2020), ce modèle de processus est une référence lorsqu'il s'agit d'appliquer des méthodes d'extraction de données à des cas industriels. La méthode CRISP-DM se caractérise par des échanges entre une phase de compréhension du cas industriel et une phase de compréhension des données. Par la suite vient une étape de préparation des données puis de modélisation. Enfin, il est question d'évaluer le modèle et de le déployer. Dans le cadre de ce travail, les phases de compréhension du cas industriel, des données qui le décrivent ainsi que de préparation sont conservées afin d'obtenir les données exploitables les plus pertinentes pour la suite des analyses. Une étape d'investigation des conditions de production permettra d'identifier les facteurs susceptibles d'influer sur la cadence de production, suivie d'une analyse du processus de production étudié afin de souligner les étapes critiques vis-à-vis du processus en cadence habituelle.

3.3 Démarche proposée d'identification des bonnes pratiques

La démarche proposée se différenciera de la méthode CRISP-DM afin de faire de l'exploration de données, tel que soutenu par IBM (2021). La phase de modélisation sera remplacée par de l'analyse exploratoire, de la visualisation de données ainsi que la découverte de processus. La figure 3.1 présente ainsi la démarche proposée pour identifier les bonnes pratiques. Les bonnes pratiques sont définies comme les comportements permettant d'éviter les pertes de cadence, à adopter face aux machines de production et inspirées des pratiques existantes.

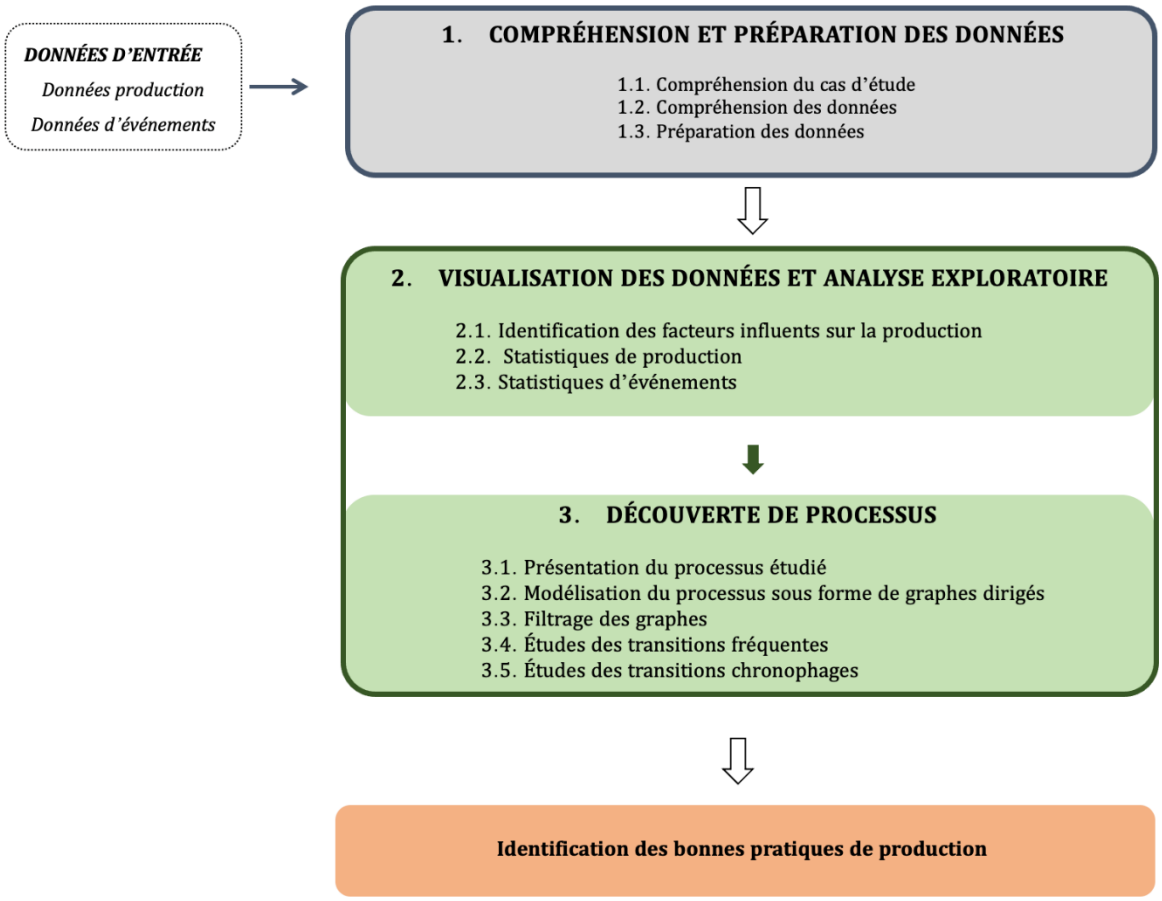


Figure 3.1 Démarche proposée pour identifier les bonnes pratiques

Après la récupération des données d'entrée, la démarche proposée se divise en trois parties. Dans un premier temps, il s'agit de s**'assurer de la compréhension et de la préparation des données**. Pour cela, il faut mettre en parallèle la réalité du cas d'études avec les données récupérées (étape

1.1 et 1.2). Une fois les données prises en main, elles doivent être préparées afin d'être exploitées dans les meilleures conditions (étape 1.3). Cette première partie peut être sujette à différents retours et itérations afin de faire correspondre au mieux les données à la réalité de la production. Dans un second temps, il s'agit de faire de **la visualisation et de l'analyse exploratoire de données** dans le but d'avoir une première mesure de la performance vis-à-vis des différents facteurs qui impactent la cadence de production (étapes 2.1 à 2.3). Cette seconde partie est aussi l'occasion de créer des profils de production, recensant différentes conditions de production relative à un niveau de performance. Par la suite, ces profils permettent de comparer des niveaux de performance présentant des caractéristiques de production similaires. **La découverte de processus** sera l'occasion d'approfondir les résultats obtenus précédemment et d'étudier des comportements de production (étapes 3.1 à 3.5) grâce à l'outil visuel qu'est le graphe dirigé. Ces graphes permettent de se concentrer sur les étapes du processus qui reviennent fréquemment ou qui prennent du temps dans le processus de production étudiée. Cette démarche mènera à l'**identification des bonnes pratiques de production** dans le contexte étudié. En effet, elle permettra aux experts du domaine d'analyser des graphes dirigés correspondant à des profils de production similaires, mais affichant des écarts de productivité. Ces graphes mettent alors en évidence différents comportements et les experts du domaine pourront s'assurer de la validité des pratiques sur le terrain. Les prochaines sous-sections présentent en détail chacune des étapes énoncées ci-dessus.

3.3.1 Compréhension et préparation des données

Dans un premier temps, il est question d'appréhender les données à disposition, que le contexte industriel qu'elles décrivent ainsi que de préparer les données

3.3.1.1 Compréhension du cas d'étude

Cette première étape de compréhension permet de poser les bases de l'étude. Il s'agit d'expliciter les attentes du projet et de comprendre la situation initiale dans laquelle se trouve le partenaire industriel. Il est important à ce stade de se familiariser avec les mécanismes en place, de comprendre les priorités du partenaire et les difficultés anticipées. Cela permettra d'identifier les objectifs tenant compte des contraintes liées à la réduction des pertes de cadence. La possibilité de pouvoir expérimenter la réalité du cas d'études ou encore d'échanger avec des experts du domaine permet de fixer une direction commune, avec des informations précises et pertinentes.

3.3.1.2 Compréhension des données

Collecte des données initiales

Un premier point à prendre en compte concerne la manière dont sont collectées les données. Certaines données peuvent être collectées automatiquement par la machine, d'autres peuvent être recueillies manuellement. Il s'agit de savoir précisément quelles connaissances sont accessibles. Une fois que le cadre des données disponibles est déterminé, il est alors possible d'avoir une idée de la quantité de données à disposition, de leurs types, de la taille des différentes tables et des informations qu'elles transmettent. Afin de s'attaquer aux pertes de cadences en production, certaines données doivent minimalement être collectées. Nous considérons les données ci-dessous nécessaires à l'évaluation des performances de production :

- des données de production : elles doivent permettre de suivre les entités réalisées au cours du processus étudié. Tout type d'informations supplémentaires permettant de caractériser ces entités peuvent faire partie des tables initiales, comme les paramètres utiles à la production; et
- des données d'événements : **un événement** correspond à une étape du processus de production étudié. Il peut s'agir d'une activité essentielle du processus ou bien d'une anomalie. Les données doivent permettre de suivre les différents événements qui ont lieu au cours du processus étudié.

Description des données

Il est également nécessaire de comprendre la structure des données et les relations entre les tables ainsi que le format dans lequel elles sont stockées. Une fois ces données principales collectées, elles sont décrites afin d'avoir une compréhension plus profonde des informations qu'elles détiennent. Il s'agit de connaître la taille des tables, de connaître le type des données collectées et le nombre de valeurs que chacun des attributs peut prendre. Cette phase de description met en lumière les informations recueillies, mais pointe aussi les informations manquantes. Chacune des tables est décrite en profondeur pour faire un état des attributs à disposition et mesurer leur capacité à décrire la production et les événements. À ce stade, il est déjà possible de se rendre compte des informations qui ne sont pas suffisamment claires pour complètement comprendre le cas d'étude ainsi que de savoir si les connaissances métiers nécessaires sont toutes présentes sous forme de

données enregistrées dans les différentes tables à disposition. Il est donc possible de faire prendre du recul et de revenir à la compréhension du cas d'étude avant de poursuivre.

Exploration des données

Lorsque les outils et les ressources à disposition sont compris, il est possible de passer à l'exploration de leur contenu. Cette phase permet d'extraire une première couche de connaissances en regardant notamment les relations entre les différents attributs. Dans un contexte de production, il est courant d'utiliser la méthode des 5M, inspirée du diagramme d'Ishikawa, qui propose des catégories de sources d'erreurs, à savoir la Matière, le Matériel, les Méthodes, la Main-d'œuvre et le Milieu (American Society for Quality, s.d.). À partir de ces catégories, il est possible d'explorer les données par des méthodes de visualisation de statistiques. Dans le cadre de l'étude des pertes de cadence en production, il s'agit d'étudier la production en fonction des 5M, par exemple en regardant le nombre de produits réalisés ou les temps d'opérations au moyen de graphes de distribution (boîtes à moustaches, graphiques à barres et autres).

Vérification de la qualité des données

La vérification de la qualité des données permet de s'assurer de traiter les problèmes d'intégrité des données. La présence de doublons et la proportion des données manquantes doivent être évaluées afin de prendre des décisions adéquates quant à leur traitement. Il faut également étudier les points aberrants par des méthodes telles que l'affichage de boîtes à moustache ou d'autres outils de détection univariée des points aberrants.

3.3.1.3 Préparation des données

Sélection des données

Dès lors que le cadre est posé, à la lumière des étapes précédentes, les données non pertinentes pour la suite de l'étude sont mises de côté. Les données conservées sont alors présentées. Dans le déroulement de cette méthodologie, les données sélectionnées sont :

- pour les données de production : afin de suivre la production, elles doivent, *a minima*, identifier les produits de façon unique, les types de produits, l'environnement de production (machine, ressources extérieures) et la date de production. Le temps d'opération relatif au processus étudié doit être accessible directement ou à partir des dates. D'autres

informations supplémentaires peuvent venir affiner chacun des produits afin de réaliser d'autres analyses pour les besoins de l'étude;

- pour les données d'événements : afin de suivre les différents événements qui ont lieu au cours du processus, elles doivent, *a minima*, suivre une nomenclature permettant d'identifier chaque événement, ainsi qu'un horodatage de début et de fin. D'autres informations supplémentaires peuvent venir affiner chacun des événements afin de réaliser d'autres analyses pour les besoins de l'étude;
- toutes données supplémentaires notamment des définitions peuvent être sélectionnées si elles permettent une meilleure compréhension.

Ces données doivent pouvoir être entrecroisées afin d'associer un événement à un produit. Si ce n'est pas déjà le cas lors de la collection, elles seront mises en parallèle par la suite.

Construction et intégration des données

La sélection des données ne permet pas toujours d'avoir toutes les informations nécessaires à portée de main afin de réaliser les différentes analyses. Des informations peuvent être présentes dans les différentes tables de données, mais de façon implicite. Il s'agit alors d'enrichir des données par la création de nouveaux attributs (*Feature Engineering*). Pensons ici à l'exemple des durées d'événements. Si un horodatage de début et de fin sont collectés dans les données initiales, il est possible de créer une nouvelle variable *Durée* à partir de l'horodatage de fin moins l'horodatage de début. L'apport des connaissances des experts du domaine peut aussi mener à la construction de données non présentes initialement dans les données collectées. Enfin, selon la disponibilité des données, il est parfois nécessaire de joindre des tables pour compléter des données existantes.

Nettoyage des données

Les données de mauvaise qualité sont traitées. Comme évoqué précédemment, les doublons sont retirés. Si certains attributs présentent une faible proportion de données manquantes, les enregistrements qui s'y réfèrent sont retirés ou non traités lorsqu'elles sont négligeables. Dans le cas où les données manquantes sont présentes en proportion importante, il s'agit s'évaluer si l'attribut auquel elles se réfèrent est pertinent pour l'étude. S'il est, il est alors judicieux d'essayer de les combler à partir des connaissances métiers.

Formatage des données

Les données ayant été nettoyées, il faut s'assurer ensuite que les données du même type sont au bon format. Il peut alors s'agir de discrétiser des variables, de catégoriser des éléments, binariser (encodeur « One Hot ») ou encore de standardiser des formats. Cela permet d'éviter les problèmes d'exploitation pour la suite.

3.3.2 Visualisation et analyse exploratoire des données

À ce stade, les données doivent représenter, de la façon la plus fidèle possible, la réalité du contexte dans lequel elles s'inscrivent, ainsi qu'être prêtes à être exploitées. L'objectif de cette deuxième étape est alors, à partir de ces données, de faire un état de la productivité dans le contexte de production étudiée. La visualisation des données est alors un élément fondamental dans la valorisation de données de production, car elle fournit un support permettant d'avoir un regard critique sur les données et d'en extraire de la connaissance (Cerquitelli et al., 2020). La productivité fera ici référence à la quantité de produits réalisés pour des ressources données par unité de temps.

3.3.2.1 Identification des facteurs influents sur la production

Afin d'évaluer la performance, il est tout d'abord question de déterminer un outil adéquat, à savoir une métrique d'évaluation de la performance, afin de pouvoir quantifier ce qui l'influence. Dans un contexte de production, le nombre de produits réalisés par unité de temps est une métrique souvent utilisée. Les données collectées concernant la production ainsi que la connaissance métier doivent permettre d'avoir une première idée de ce qui influence la cadence de production. Il est aussi possible de réaliser des entretiens avec les experts du domaine afin de distinguer d'autres facteurs influents spécifiques. L'état de l'art est aussi une source fiable d'idées d'exploration des facteurs qui pourraient influer sur la cadence. Différentes ressources peuvent intervenir sur une ligne de production : par exemple, certaines matières peuvent être plus ou moins facile à travailler, la taille des produits peut influencer le temps d'opération, l'environnement de travail que ce soit le type de machine et de ressources consacrées au processus étudiés. Une première utilisation de visualisation de données peut être utilisée afin d'évaluer la métrique sélectionnée en fonction de différents facteurs.

3.3.2.2 Statistiques de production

Dans cette phase, il s'agit d'utiliser l'analyse exploratoire de données afin de faire parler les données sélectionnées. Dans un premier temps, ce sont les données de production qui sont traitées. Dans une optique d'identifier des bonnes pratiques, cette étape doit mener à la création d'une base de comparaison de profils de production, présentant des niveaux de productivité différents. En effet, afin de juger les différentes pratiques mises en œuvre, il faut pouvoir comparer ce qui est comparable. Ainsi, l'objectif est de sortir des statistiques, sous forme de graphes afin de rendre plus facile l'interprétabilité. Par exemple, il peut être question de :

- diagrammes circulaires pour présenter des proportions; et
- diagramme à barres ou boîtes à moustaches pour afficher des distributions.

Il s'agit donc de créer des profils de production ayant des caractéristiques similaires, à la lumière des facteurs influents déterminés précédemment, menant par ailleurs à des niveaux de productivité différents.

3.3.2.3 Statistiques d'événements

Dès lors que des profils de production ayant des caractéristiques similaires sont identifiés, il est temps de se concentrer sur les différents éléments qui peuvent interférer dans leurs processus de production. Ainsi, chacun des profils est affiné par l'affichage de statistiques et de graphes permettant de comprendre les événements rencontrés. Ici, ce sont les données d'événements qui sont exploitées. De la même façon que pour les statistiques de production, l'analyse exploratoire de données permet cette fois-ci d'avoir accès à différentes informations à savoir :

- proportion des événements rencontrés durant le processus;
- statistiques de pertes de cadence :
    - Parmi les événements rencontrés, quelle proportion mène à un ralentissement ?
    - Parmi les événements rencontrés, quelle proportion mène à un arrêt?
    - Parmi les événements rencontrés, quelle proportion fait référence à des événements de maintenance?

Finalement, il est possible pour chacun des profils de connaître les tendances de production ainsi que les événements qu'il rencontre lors de la production. Cependant, afin d'identifier des bonnes pratiques, les statistiques à elles seules ne suffisent pas. En effet, à ce stade, il possible de différencier des niveaux de productivité à la lumière des différentes caractéristiques de production. Cependant, cela ne permet pas d'identifier les décisions prises lors des différentes étapes du processus, qui ont mené à ces niveaux de productivité. La suite de ce travail sera alors l'occasion de remonter d'un cran dans le processus afin de mieux comprendre les pratiques à l'origine de ces profils de production.

3.3.3 Découverte de processus

La découverte de processus est une sous-partie de l'exploration de processus. Selon Van der Aalst (2016), l'exploration de processus vient compléter la fouille de données et l'apprentissage machine en y ajoutant une dimension plus axée sur les processus. L'exploration de processus est un moyen de valoriser des données événementielles et de comprendre des comportements (Van der Aalst, 2016). L'exploration de processus permet notamment de faire de l'acquisition de connaissances en créant des modèles à partir d'un journal d'événements enregistrés (Ahmed & al., 2019). Il existe divers outils pour réaliser et visualiser de la découverte de processus. Dans ce travail, les graphes dirigés (DFG) sont utilisés. Ils sont un bon outil pour représenter visuellement les relations directes pondérées entre des événements, facilitant ainsi l'interprétabilité des résultats (Dupuis et al.,2022).

3.3.3.1 Présentation du processus étudié

Dans cette phase de présentation, il s'agit essentiellement de décrire le processus à partir des données d'événements et à la lumière des connaissances métiers. Une description plus approfondie des codes événements ainsi que de leurs catégories respectives est réalisée afin de comprendre comment leur apparition se traduit en pratique sur la machine. C'est aussi l'occasion de comprendre ce que cela implique pour l'intervenant au contact de la machine et qu'elles sont les scénarios qui se présentent à lui lorsqu'il fait face à l'apparition d'un événement. En effet, la plupart du temps en production, les actions non automatisées sont soumises à des procédures à suivre. Cette première démarche permet de savoir à quoi s'attendre ainsi que de se concentrer sur les informations importantes et révélatrices dans l'évaluation des pertes de cadence.

3.3.3.2 Modélisation du processus sous forme de graphes dirigés

Dès lors qu'une meilleure connaissance de la réalité du processus est acquise, il s'agit de le modéliser. Ce sont les données d'événements qui sont utilisées ici. La découverte de processus se fait à partir d'un journal d'événements. Ce dernier décrit ligne par ligne les différentes étapes par lesquels passe un cas. Dans le cadre de ce travail, chaque ligne décrira :

- l'identifiant unique de chaque produit;
- l'« événement » ou état par lequel le produit passe c.-à-d. le code d'événement; et
- l'horodatage de l'événement.

Il est aussi possible d'ajouter des informations concernant les ressources intervenant à chaque événement rencontré, mais ce cas-là n'est pas traité dans cette méthodologie. En effet, la question des ressources est traitée par filtration des données en entrée de l'algorithme de découverte de processus.

Le tableau suivant donne un exemple du type de table de données pour cette étude.

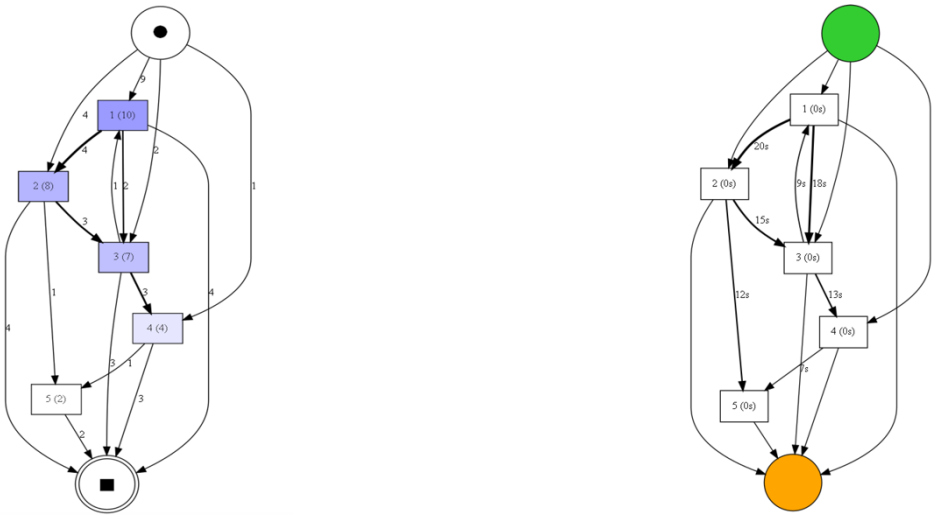
Tableau 3.1 Exemple de tables de données événementielles pour la découverte de processus

Produit	***Horodatage***	***Événement***
A	01/01/2022 - 8 :00	1
A	01/01/2022 - 8 :01	2
A	01/01/2022 - 8 :02	3
B	01/01/2022 - 8 :05	1
C	01/01/2022 - 8 :25	1
C	01/01/2022 - 8 :26	2

Lors de la production du produit A, l'événement 1 a lieu le jour 1 à 8:00, suivi de l'événement 2 le même jour à 8:01 puis l'événement 3 à 8:02. Pour le produit B, seul l'événement 1 a lieu le jour 1

à 8:05. Enfin, pour le produit C, l'événement a lieu à 8:25 puis l'événement 2 à 8:26, le même jour. Ainsi, plusieurs lignes peuvent faire référence à un produit unique, mais chacune des lignes se réfèrent à une seule activité qui se produit à un instant précis.

Dès lors que la table est formatée de la façon présentée ci-dessus, il est possible de tracer des graphes dirigés.



a) Pondération fréquentielle

b) Pondération temporelle

Figure 3.2 Exemple de graphes dirigés pondérés

Les graphes dirigés sont des outils de découverte de processus permettant de représenter l'intégralité des étapes suivis par différents produits. Les points noir et vert représentent l'entrée du produit dans le processus. Le point orange et le carré indiquent la sortie du processus. Ces graphes offrent la possibilité de connaître le nombre d'occurrences de chacun des événements rencontrés dans la table de données et le nombre d'occurrences de chacune des transitions entre événements (Figure 3-2-a) ou encore la durée moyenne de chacune des transitions (Figure 3-2-b). La durée indique en réalité le temps passé dans l'état initial de la transition avant d'aller à l'état final. Dans la figure 3-2-a, l'intensité de la couleur est proportionnelle au nombre d'occurrences :

plus le nombre d'occurrences d'un événement dans la table de données est élevé, plus la case correspondante dans le graphe sera foncée.

L'idée poursuivie est d'exploiter les DFG afin de juger quelles sont les transitions critiques dans l'évaluation des pertes de cadence, à savoir les transitions les plus fréquentes ainsi que les transitions les plus chronophages. Dans l'exemple donné ci-dessus, l'échantillon reste faible. Pour des échantillons plus importants et dans un tel contexte de production, il est possible que les graphes obtenus soient des graphes dit « spaghetti » c.-à-d. avec un nombre important de transitions, empêchant leur bonne lecture et donc de se concentrer sur les transitions les plus critiques. Afin de pallier ce phénomène, il est possible de filtrer les informations en fonction de leur pertinence.

3.3.3.3 Filtrage des graphes

Dans un contexte de production semi-automatique et pour des produits complexes, il est fréquent d'avoir des graphes dont la lecture est délicate. Pour avoir une meilleure approche, il est alors courant de filtrer les données. Cependant, il est important de rester vigilant dans cette phase de filtrage, car cela peut mener à des informations erronées (Van Der Aalst, 2019). En effet, déterminer le choix du filtre est complexe. Afin de se concentrer sur les transitions critiques, le principe de Pareto est proposé dans cette méthodologie pour filtrer l'affichage des transitions critiques. Ce principe affirme que 20% des causes sont responsables de 80% des conséquences. Ainsi sont conservées :

- les transitions dont les occurrences cumulées représentent 80% de la totalité des occurrences des transitions; et
- les transitions dont les durées cumulées représentent 80% de la durée totale des transitions.

Les graphes sont alors interprétables pour la suite de l'analyse.

3.3.3.4 Études des transitions fréquentes

À ce stade, les graphes dirigés filtrés des différents profils de production déterminés au 3.3.2 sont accessibles. L'objectif est alors de comparer les différents graphes afin de déceler des différences au niveau des comportements et réactions face aux événements rencontrés lors de la production. Concernant l'étude des transitions fréquentes, deux études sont menées. La première consiste à

évaluer le nombre d'occurrences et la seconde consiste à regarder pour ces mêmes transitions, la durée moyenne. La mise en parallèle des graphes de différents profils permet alors d'identifier des comportements en :

- comparant les occurrences des événements et des transitions : sont-elles similaires ou présentent-elles des écarts importants ?
- comparant les durées moyennes de transitions : sont-elles similaires ou présentent-elles des écarts importants ? Sont-elles significatives au regard des temps d'opération?

L'analyse des graphes dirigés est un premier pas vers l'identification de bonnes pratiques. En effet, ils permettent de mettre en avant des étapes du processus qui nécessiteraient d'être observé de plus près sur le terrain. Chacune de ces informations est à mettre en perspective avec la réalité du processus étudié. En effet, certains événements sont nécessaires au bon fonctionnement du processus. Ainsi, regarder leur nombre d'occurrences n'est pas forcément révélateur. Cependant, le fait que ces événements soient nécessaires ne veut pas forcément dire qu'ils sont appréhendés de la bonne façon. Ainsi, étudier la durée des transitions impliquant ces événements nécessaires peut s'avérer révélateur. Cela soulève aussi la nécessité d'évaluer les transitions peu fréquentes, mais occupant une longue durée lors du processus.

3.3.3.5 Études des transitions chronophages

De la même façon, les graphes dirigés temporels filtrés sont étudiés dans cette partie. L'objectif est alors de compléter l'étude précédente en s'assurant que des transitions moins fréquentes, mais dont la durée moyenne reste élevée ne sont pas laissées de côté. La mise en parallèle de ces nouveaux graphes pour les différents profils permet alors :

- d'identifier d'éventuels nouveaux événements moins fréquents vis-à-vis des graphes dirigés d'occurrence, pour chacun des profils; et
- de comparer les durées moyennes de chacun des profils pour des événements identiques: sont-elles similaires ou présentent-elles des écarts importants ? Sont-elles significatives au regard des temps d'opération?

L'analyse de types de ces graphes dirigés temporels permet de cibler un autre type de pertes de cadence et d'approfondir la compréhension des comportements adoptés face à la machine.

L'interprétation des DFG n'est finalement pas triviale. Elle présente des limites quant aux informations qui peuvent en être extraites. En effet, les DFG permettent de connaître la place qu'occupe un événement ou une transition entre deux événements. Cependant, hors contexte, des résultats peuvent être laissés de côté ou même mal interprétés. L'apport des connaissances métiers est alors primordial à cette étape de l'analyse. Les experts peuvent alors traduire les différentes transitions, identifier des anomalies ou encore distinguer des transitions révélatrices de bons comportements. Dans l'évaluation des pertes de cadence, il est maintenant possible d'avoir une approche plus approfondie de la productivité. Cependant, cette méthodologie ne permet pas l'évaluation du respect des normes et des pratiques menant à la réalisation de produits dont la qualité est valide.

3.3.4 Conclusion

Ce chapitre a permis de préciser nos objectifs de recherche et de proposer une démarche de valorisation de données de production afin d'évaluer les pertes de cadence dans un contexte de production semi-automatique. La valorisation est rendue possible par l'utilisation d'outils d'analyse exploratoire de données et d'exploration de processus. En suivant cette méthode, il est possible d'aller de la collecte de données d'une ligne de production industrielle jusqu'à la visualisation d'étapes critiques d'un processus de cette ligne. Cette méthode propose alors un outil permettant de distinguer des comportements qui pourraient être à l'origine de pertes de cadence. L'utilisation de cet outil par les experts du domaine et l'interprétation qu'ils en dégagent permet d'identifier des comportements et des bonnes pratiques relatives à la productivité. La mise en œuvre de cette méthodologie pour un cas d'application sera présentée au chapitre suivant.

CHAPITRE 4 CAS D'ÉTUDE

Au cours de ce chapitre, il sera question d'appliquer la méthodologie développée au chapitre précédent au cas d'étude du partenaire industriel de ce projet.

4.1 Mise en contexte

Le partenaire industriel est une usine de la division nord-américaine d'une multinationale manufacturière, située au Québec. Cette usine produit notamment des produits pour véhicules particuliers, véhicules utilitaires sport, mais aussi pour camionnettes. En termes de production, elle a suivi une évolution de ses équipements afin de parvenir à des processus de plus en plus automatisés.

4.2 Description du cas d'étude

Cette étude se concentre sur une partie précise de la ligne de production, à savoir l'assemblage. À l'heure actuelle, l'usine a remplacé une majorité de ses machines d'assemblage par des machines semi-automatiques afin d'augmenter sa productivité. Certaines machines non automatiques sont encore présentes sur le plancher de production afin de permettre d'atteindre les quotas fixés par l'entreprise. En effet, les machines semi-automatiques sont à l'origine d'une baisse de productivité notamment due à des pertes de cadence comme des micro-arrêts ou des ralentissements. Ce chapitre explicite l'application de la méthode développée au chapitre précédent au cas particulier du partenaire industriel de ce projet. L'objectif est alors d'identifier les différentes pratiques menant à une meilleure productivité sur les machines d'assemblage semi-automatiques.

4.3 Méthodologie de mesure de la performance

4.3.1 Compréhension du cas d'étude

Dans un premier temps, diverses rencontres avec le partenaire industriel ont permis de clarifier l'objectif de ce travail ainsi que de mieux comprendre le contexte dans lequel il allait s'inscrire. Afin d'atteindre les quotas fixés par la maison mère, le partenaire souhaite analyser le comportement de ses machines d'assemblage. En effet, bien que semi-automatiques, elles présentent différentes pertes de cadence à savoir des micro-arrêts et des ralentissements. Différents

capteurs permettent de relever plusieurs types d'informations relatives à la production. Le partenaire souhaite les exploiter afin de mieux comprendre l'origine et l'impact des pertes de cadence.

4.3.2 Compréhension des données

Collecte des données initiales

Initialement, les données sont collectées automatiquement à partir de capteurs présents sur les machines ou mis en place ultérieurement par des ingénieurs de l'usine, à des fins d'analyse. Les données initiales sont extraites depuis un système de gestion de base de données mis à disposition sur le serveur de l'entreprise. Dans un tel contexte, la majorité des données sont collectées automatiquement. Elles sont exportées sous forme de fichiers CSV afin de pouvoir être exploitées sur l'environnement de développement *Spyder* pour le langage de programmation Python.

Les données extraites sont des données relatives au fonctionnement des machines d'assemblage. Différentes tables permettent de classer les informations du processus :

- un journal de production : cette table recense toutes les informations concernant les produits réalisés sur la ligne à savoir un identifiant unique du produit, différents compteurs permettant de dénombrer les produits, la machine concernée, les ressources engagées dans la production, l'horodatage de réalisation du produit, les paramètres utilisés et le temps d'opération. À ces informations principales s'ajoutent différentes informations plus techniques concernant les opérations appliquées sur ce produit;
- un journal d'événements : cette table fait l'inventaire des différents événements qui se produisent sur une machine donnée. Ils se caractérisent par un code permettant l'identification de l'événement, un horodatage de début et un horodatage de fin ainsi que le code de la machine concernée. Les interactions de ces machines d'assemblage avec leur environnement se font à l'aide de ces codes. Ils peuvent révéler une succession d'événements normaux lors de l'assemblage des produits, comme un approvisionnement de matières. Cependant, ils sont aussi utilisés dans le cas d'apparition d'anomalies lors du processus. Les codes peuvent être automatiques ou manuels et sont donc significatifs dans l'étude des pratiques d'assemblage; et

- des tables de définitions : certaines tables permettent d'apporter des précisions sur des données présentes dans les autres tables, par exemple les catégories des différents codes d'événements.

Description des données

Une première prise en main des tables avec l'accompagnement des experts du domaine permet une meilleure compréhension des valeurs relevées ainsi que des variables qui permettraient de mesurer la performance de production.

D'après le journal de production, 15 machines permettent l'assemblage de différents produits. Ces machines interagissent avec 175 intervenants et assemblent 118 types de produits. Comme évoqué précédemment, le temps d'opération est enregistré pour chacun des produits. La connaissance métier indique que selon le type de produit et ses caractéristiques, le temps d'assemblage varie.

D'après la table de définitions des codes d'événements, il existe 150 codes d'événements, répartis dans 14 catégories. Chacun des codes possède une contrainte d'exclusion basée sur une durée maximale. En revanche, cette table ne permet pas de savoir si chacun des codes est à l'origine d'un ralentissement ou d'un arrêt. Au-delà de la dénomination utilisée pour chacun des codes, elle ne permet pas non plus de savoir si le code est un code automatique ou s'il doit être inséré manuellement. De plus, elle ne transmet pas de renseignement sur l'implication des intervenants dans l'apparition d'un code donné.

D'après le journal d'événements, chacun des codes a une date de début, une date de fin et est associé à une machine par un identifiant. Chacune des occurrences de code est repérée par un identifiant unique. Cependant, cette table ne permet pas de savoir à quelles catégories appartiennent chacun des codes enregistrés.

Afin de pouvoir s'intéresser aux différentes pertes de cadence affectant la production, il est nécessaire de pouvoir croiser ces données. En effet, le journal de production ne permet pas de décrire l'apparition d'événements. De la même façon, le journal d'événements ne permet pas d'associer un code événement à l'assemblage d'un produit.

Exploration des données

Lors de cette première phase d'exploration, différentes méthodes de visualisation de statistiques sont utilisées. La métrique d'évaluation de la production utilisée par le partenaire industriel est le nombre de produits assemblés par jour. Inspiré de la méthode des 5M, cette phase est l'occasion d'évaluer cette métrique au regard des facteurs que sont les machines, les types de produits, la main-d'œuvre ou encore les jours de la semaine.

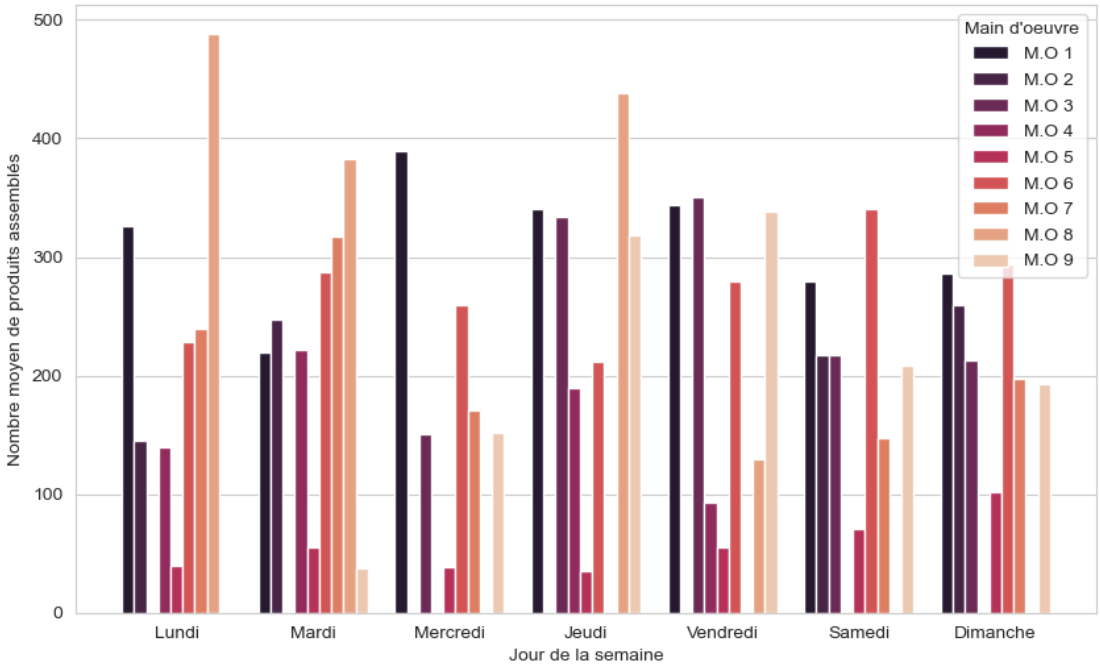


Figure 4.1 Distribution du nombre moyen de produits assemblés par jour en fonction des jours de la semaine, pour différents intervenants d'une machine, sur une période de trois mois

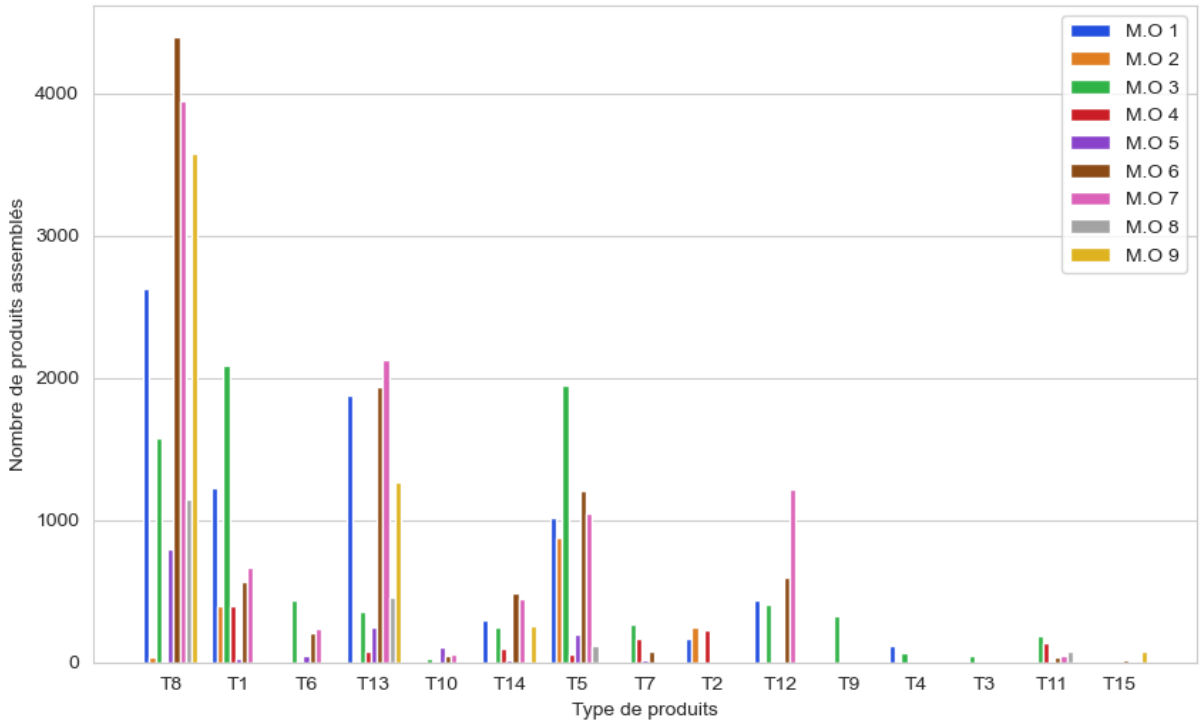


Figure 4.2 Distribution du nombre moyen de produits assemblés en fonction des types de produits, pour différents intervenants d'une machine, sur une période de trois mois

Sur les figures 4.1 et 4.2, les différentes couleurs représentent la main-d'œuvre. Ces graphes ne permettent pas d'identifier de modèles pertinents. L'analyse du nombre de produits assemblés par unité de temps devra être approfondie par la suite. En effet, le choix d'analyse du nombre de produits assemblés par jour n'est peut-être pas le plus pertinent, la taille de l'échantillon de données pourrait être augmentée et d'autres caractéristiques pourraient être étudiées. L'analyse du nombre de produits assemblés en fonction du type de produits indique que le type T8 est le plus assemblé sur la période étudiée (Figure 4.2). Cependant, cette observation seule ne permet pas d'avoir d'informations claires quant à la productivité. De plus, à ce stade de l'étude, il n'est pas possible de mettre en parallèle les produits et les événements de production puisqu'aucun attribut ne permet de relier les deux tables en question.

Par ailleurs, l'observation d'une machine en fonctionnement a permis d'approfondir la compréhension des données. En effet, les machines semi-automatiques d'assemblage sont décomposées en 3 sections, 2 sections assemblant deux sous-parties du produit et une section

centrale assurant l'assemblage complet du produit. Ces trois sections fonctionnent en parallèle, rendant l'enregistrement des données plus complexe que pour un processus linéaire. La figure 4.3 décrit l'apparition d'un événement à un instant donné de l'assemblage. Dans cette figure, un événement peut survenir lors de l'assemblage de la sous-partie 2 du produit N ou lors de l'assemblage de la sous-partie 1 du produit N+1. Ainsi, un code événement est enregistré, mais l'information concernant la section pour laquelle il apparaît n'est pas renseignée dans le journal d'événements.

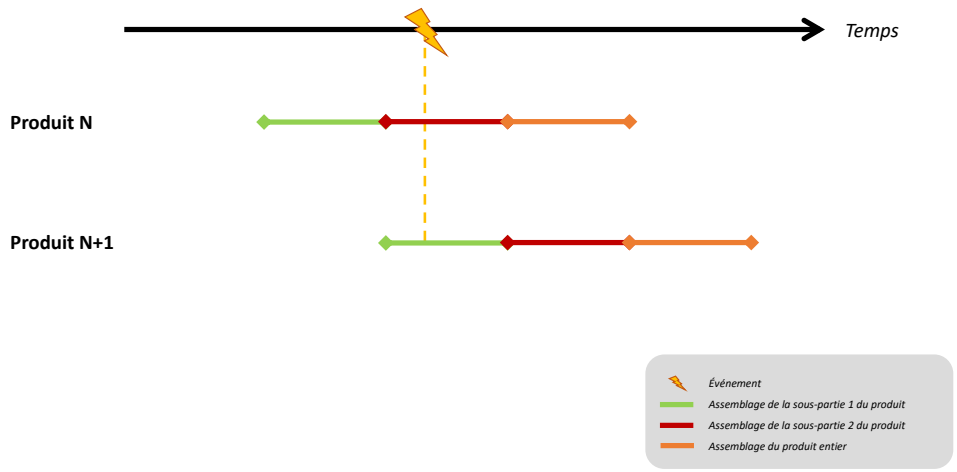


Figure 4.3 Explication du conflit temporel lors de la considération d'un événement

Cette phase montre que l'évaluation des pertes de cadence n'est pas triviale et ne se limite pas à regarder l'évolution du nombre de produits en fonction de différentes caractéristiques de production. Par ailleurs, des éléments de la réalité du processus de production sont à prendre en compte pour préparer les données adéquatement par la suite.

Vérification de la qualité des données

Après le premier traitement des données, il est possible de se rendre compte du mauvais formatage de certaines d'entre elles ainsi que de l'existence de doublons. Certaines données sont manquantes dans les tables de production et d'événements, mais représentent pour chacun des attributs concernés moins de 0,5% des données de l'échantillon.

4.3.3 Préparation des données

Sélection des données

L'analyse présentée portera sur une seule machine d'assemblage sur une durée de trois mois. L'échantillon de données est constitué d'un journal de production et d'un journal d'événement. Comme vu précédemment, différents types d'informations sont présents dans les tables de données, cependant ils ne sont pas tous pertinents pour cette étude. Afin de pouvoir évaluer les différents niveaux de productivité, les attributs suivants sont sélectionnés (Tableaux 4.1 et 4.2).

Tableau 4.1 Données de production sélectionnées

Journal de production	*Types de données*
Identifiant unique du produit	Chaîne de caractères
Types de produit	Chaîne de caractères
Horodatage	Date time
Machine	Chaîne de caractères
Temps d'assemblage	Nombre flottant

Tableau 4.2 Données événementielles sélectionnées

Journal d'événements	*Types de données*
Code d'erreur	Chaîne de caractères
Catégorie d'erreur	Chaîne de caractères
Horodatage	Date time
Machine	Chaîne de caractères

Construction et intégration des données

Pour les besoins de l'analyse, les tables existantes ont été enrichies par les variables suivantes :

- des données temporelles présentes de façon implicite, notamment des durées d'opération ou des durées d'événements, sont calculées à partir des horodatages;
- des informations supplémentaires concernant les événements sont ajoutées. En effet, la table des événements ne permet pas de connaître la catégorie de chacun des codes recensés. Ainsi, les tables des événements et de définition des codes d'événements sont jointes afin d'avoir accès à toutes les informations d'un événement lorsqu'il apparaît sur la machine d'assemblage; et
- des données concernant le fonctionnement contrôlé, ou non, de la machine lors de l'apparition d'un événement au cours du processus sont intégrées aux données existantes sous forme d'un attribut binaire. Il permet de caractériser un code d'événement.

Comme évoqué précédemment, il est nécessaire de joindre les tables de production et d'événements. L'objectif est de pouvoir associer un événement ou une succession d'événements à un produit assemblé. Or, dans le cas particulier du partenaire industriel, ces deux tables ne partagent aucun autre attribut permettant de réaliser une jointure autre que le temps. La figure 4.4 décrit comment les différents horodatages sont réalisés.

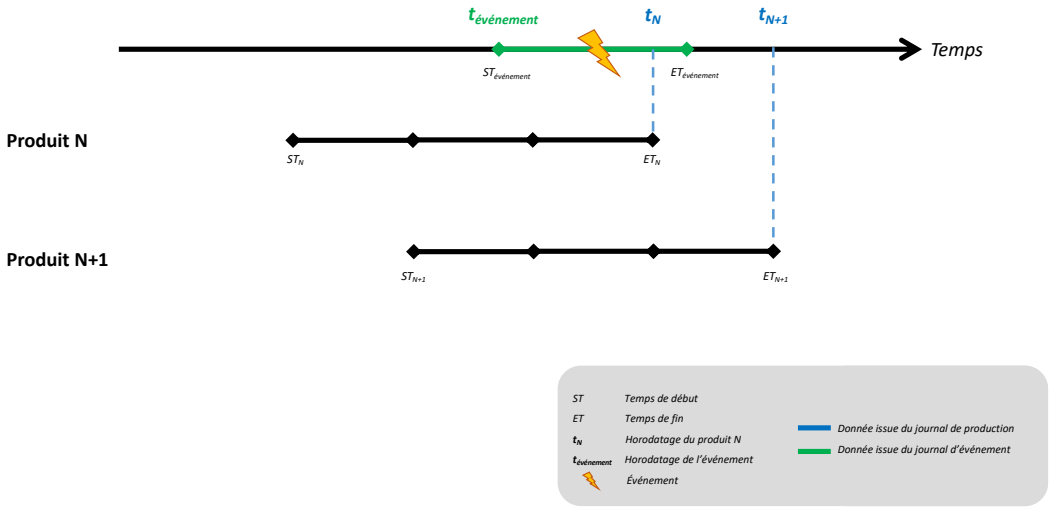


Figure 4.4 Explication du choix d'horodatage d'un événement

Ainsi, les deux tables sont jointes bout à bout à partir de l'horodatage et la table finale est triée par ordre chronologique. Comme détaillé dans la figure 4.4, le choix est fait de considérer la date de début de l'événement comme son horodatage, afin de rester cohérent avec la chronologie du processus d'assemblage. Finalement, cela permet d'avoir dans une même table l'évolution de la production, incluant l'apparition d'événements, anomalies ou non.

Cependant, l'intégration de nouvelles données amène des incohérences au sein des tables, qu'il faut nettoyer avant de pouvoir les exploiter pour analyses.

Nettoyage des données

Dans un premier temps, le nettoyage concerne les données initiales. Il s'agit principalement de supprimer des données aberrantes, présentes en très faible quantité. La deuxième partie du nettoyage concerne les données intégrées *a posteriori* :

- la jointure des tables des événements et de définition entraîne l'apparition de doublons qui sont alors retirés; et
- la jointure des tables de production et d'événements fait apparaître plusieurs données manquantes. Comme énoncé dans la partie précédente, elles ne partagent aucun attribut autre que l'horodatage. Ainsi, les valeurs des attributs d'une table sont manquantes pour l'autre table et inversement. L'objectif de cette jointure est d'obtenir les informations requises en vue de réaliser de l'exploration de processus, c.-à-d. regarder les différents états par lesquels passent les différents produits. Les événements sont alors associés aux produits auxquels ils se réfèrent.

  Pour se faire, il est nécessaire de poser certaines hypothèses. D'après la connaissance métier, l'horodatage d'un produit N est relevé quand celui-ci quitte la machine. Lorsqu'un événement ou une succession d'événements apparaissent, le choix est fait de les associer au prochain produit horodaté (Figure 4.4). Cela suppose alors que les événements sont toujours associés aux deux mêmes sous-parties de l'assemblage. Dans le cas de cette étude, cette hypothèse ne pose pas de problème, car il ne s'agit pas d'étudier la machine en elle-même, mais d'évaluer quels comportements externes peuvent être à l'origine de pertes de cadence.

Ainsi, les valeurs des attributs de la table de production (énumérés lors de la sélection des données) sont copiées à partir du produit N, premier produit horodaté succédant l'événement ou la succession d'événements. Désormais, dans la table finale, une ligne relative à un code événement recense les informations du produit auquel il se réfère, à savoir son identifiant unique, son type et les ressources engagées dans l'assemblage. Seules ses données temporelles initiales sont conservées (horodatage et durée de l'événement).

Il est aussi question de retirer de l'échantillon d'étude les produits réalisés par les ressources qui ne sont pas régulières sur la machine étudiée. Cette étape est réalisée après la jointure réalisée ci-dessus afin de ne pas obtenir des données erronées. En effet, d'après l'hypothèse émise ci-dessus, si un produit horodaté juste après un événement ou une succession d'événement est retiré de l'échantillon, cet événement serait alors associé au plus proche produit réalisé par une ressource régulière. Ainsi, tous les événements sont associés aux produits réalisés par toutes les ressources, puis les événements et les produits relatifs à des ressources non régulières sont retirés.

Formatage des données

Comme évoqué lors de la vérification de la qualité des données, il s'agit ici de gérer les doublons initialement présents, les données manquantes ainsi que les données sous un format contraignant pour l'exploitation. En ce qui concerne les doublons, ils sont retirés. Les données manquantes étant négligeables (voir « *Vérification des données* ») ou faisant référence à des attributs non sélectionnés précédemment, elles ne sont pas traitées. Les données temporelles sont toutes mises au même format *datetime*, les différents identifiants numériques sont tous convertis en nombres entiers ou flottants et les codes en chaînes de caractères.

4.3.4 Visualisation et analyse exploratoire des données

Afin de pouvoir suivre un objectif d'amélioration de la productivité, il convient dans un premier temps de faire un état des lieux des performances actuelles des lignes. L'objectif est alors d'identifier les différents niveaux de productivité. La mesure des performances de production fait référence à l'évaluation de la productivité à un instant donné, pour une ressource donnée, dans des conditions données. Afin de dépeindre ces différents niveaux, les journaux de production et d'erreurs préparés sont exploités au moyen de divers outils d'analyse exploratoire. La volonté de mettre l'accent sur la visualisation des données est rendu possible par l'affichage de nombreux

graphes permettant de réaliser des analyses statistiques. Pour cela, différentes bibliothèques *Python* ont été utilisées comme *Matplotlib* ou encore *Seaborn.*

4.3.4.1 Identification des facteurs de production

Le principal indicateur de performance de l'entreprise est le nombre de produits réalisés par jour. Afin de débuter cette étude, il est important de déterminer quels sont les différents facteurs de production qui influent sur cet indicateur. En vue d'identifier les bonnes pratiques et ces dernières étant propres à chacune des ressources au contact de la machine, la première analyse porte sur l'évolution du nombre moyen de produits par unité de temps en fonction des différentes ressources qui interagissent avec celle-ci. La figure 4.5 montre la répartition du nombre moyen de produits assemblés par jour en fonction des différentes ressources opérant sur la machine.

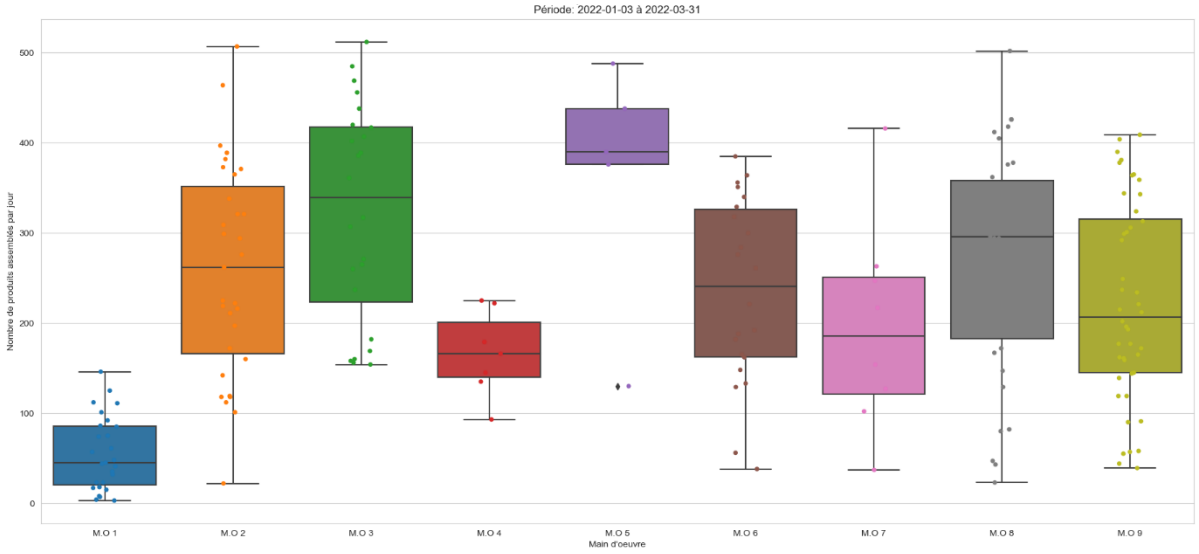


Figure 4.5 Distribution du nombre de produits réalisés par jour en fonction de la main-d'œuvre

Le nombre moyen de produits assemblés par jour n'est pas consistant en fonction de la main d'œuvre, la figure 4.5 affichant d'importants écarts. La seule évaluation de cette métrique n'est alors pas suffisante pour juger l'impact sur la cadence de production. À ce stade, les raisons menant à de tels résultats ne sont pas connues. Par ailleurs, en vue d'identifier des bonnes pratiques, cette inconsistance met en avant la nécessité de comprendre les comportements derrière ces résultats

disparates. Une approche plus ciblée de la productivité est requise. L'unité de temps est alors réduite à l'heure, car elle est plus représentative de l'activité des ressources étudiées. En effet, sur une journée, la connaissance métier indique que différents types de produits peuvent être assemblés, menant à des temps d'opération différents selon la taille du produit. Cela requiert alors des changements de séries nécessitant eux-mêmes une période d'ajustement avant de relancer la production en cadence nominale. Par ailleurs, l'usine produit en continu, donc des quarts de travail sont à cheval sur deux journées. Ainsi, le processus ne se résume pas à l'assemblage du même produit sur une journée de 24h, sur la même machine, par la même ressource et en continu. Réduire l'unité de temps permet de s'adapter à ce contexte de production plus complexe.

L'idée est alors de définir des « *profils de production* », permettant d'avoir accès à diverses informations expliquant ce qui peut mener à de tels rendements. Un profil de production se veut être révélateur d'un comportement face à la machine. Un comportement sera alors un ensemble de pratiques. Des données complémentaires décrivant plus précisément les ressources, comme les quarts de travail ou les années d'expérience au contact de la machine sont autant d'informations qui permettent de compléter un profil de production. Afin de pouvoir par la suite distinguer différents profils, il est pertinent de s'assurer qu'ils peuvent être comparés. Ainsi, il est question d'évaluer la productivité de profils présentant des similarités. Il est rappelé qu'afin de suivre une analyse comparative, les différentes statistiques sont déterminées à partir d'un échantillon sélectionné pour correspondre à une période de production de trois mois et pour une même machine donnée.

4.3.4.2 Statistiques de production

Dans un premier temps, le nombre moyen de produits assemblés par heure de production est déterminé. Les heures de production correspondent au cumul des temps d'assemblage de chacun des produits sur la durée de l'échantillon choisi. Le choix de cette unité de temps permet de savoir le temps réel passé sur la machine à assembler des produits. Il permet de prendre en compte les pertes de cadence non planifiées. La connaissance métier indique que pour une période de trois mois, il n'est pas pertinent de regarder les profils d'assemblage ne dépassant pas 4000 produits assemblés, tous types de produits confondus. La table 4.3 montre le nombre de produits assemblés sur la période (tous types de produits confondus), le total des heures de production sur la période ainsi que le nombre moyen de produits assemblés par heure de production sur la période.

Tableau 4.3 Nombre moyen de produits assemblés par heure de production pour différents profils

Profil	Nombre de produits	Heures de production sur la période considérée	**Nombre moyen de produits assemblés par heure de production**
1	9804	213,559	**46**
2	9591	201,042	**48**
3	8022	168,093	**48**
4	7778	174,75	**45**
5	5177	120,482	**43**

Ainsi, 5 cas sont conservés (Table 4.3). Afin de conserver une cohérence dans la comparaison, les cas affichant une expérience éloignée des autres sont retirés, ainsi que les profils ayant réalisé moins de 150h de production sur la période sélectionnée. Il reste alors 3 cas qui constitueront 3 profils de production (Table 4.4). Pour l'exemple, la suite de l'analyse se concentrera sur ces 3 profils, la méthodologie restant la même pour étudier d'autres situations. La table 4.4 transmet alors les mêmes informations que la figure précédente, seulement pour les 3 profils sélectionnés.

Tableau 4.4 Nombre moyen de produits assemblés par heure de production pour les 3 profils sélectionnés

Profil	Nombre de produits	Heures de production sur la période considérée	**Nombre moyen de produits assemblés par heure de production**
1	9804	213,59	**46**
2	9591	201,042	**48**
3	8022	168,093	**48**

À première vue, la performance générale semble équivalente pour la machine choisie. Afin de mieux comprendre ce qu'impliquent les différentes valeurs de la métrique du nombre de produits assemblés par heure de production, le parti pris est de l'approfondir en la décomposant. Ainsi, la première analyse porte sur le nombre de produits assemblés. La figure 4.6 affiche alors la répartition des types de produits parmi la totalité des produits assemblés pour chacun des 3 profils.

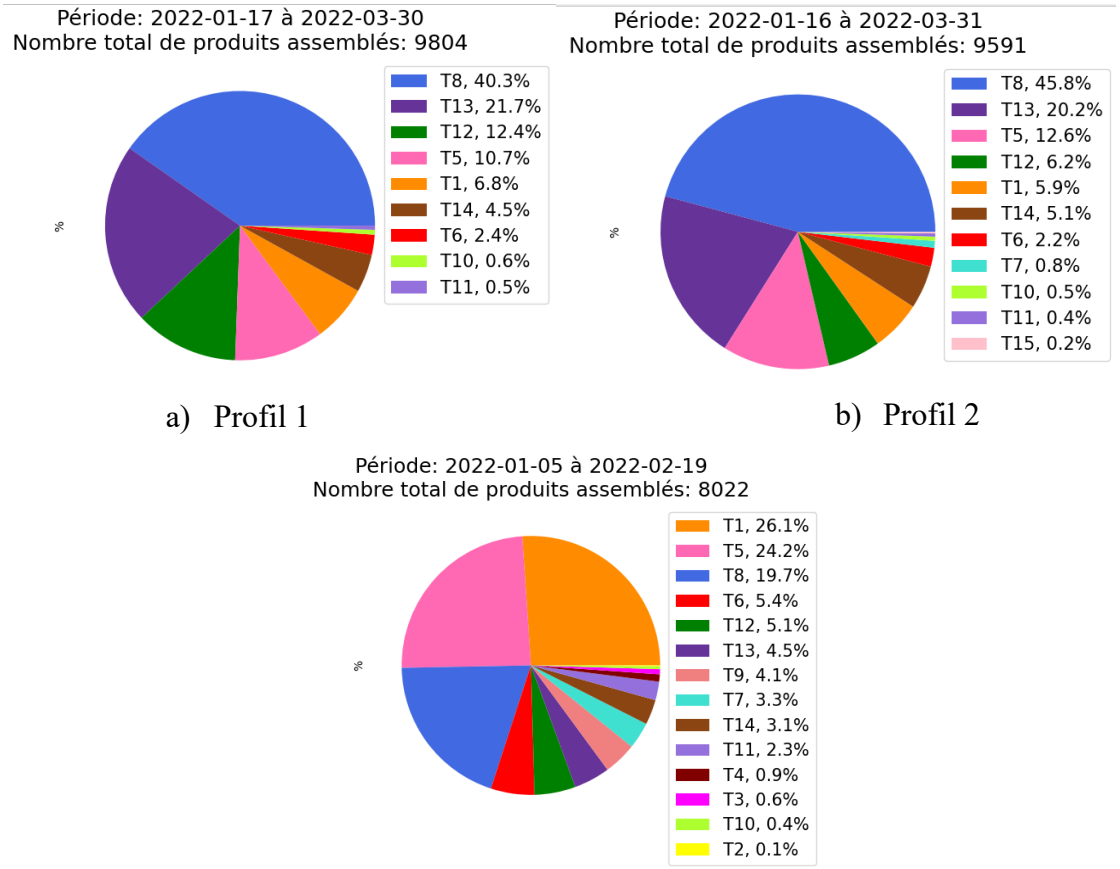


Figure 4.6 Proportions de types de produits assemblés dans les 3 profils étudiés

Les différentes couleurs sur chacun des graphes représentent les types de produits assemblés. Sur cette machine, les proportions de types de produits sont similaires pour deux des trois profils sélectionnés. Durant la période sélectionnée, le type T8 occupé une majorité de leurs assemblages, suivi du type T13. Pour le troisième profil, environ 70% de ses assemblages sont répartis sur les

types T13, T5 et T8. Afin d'avoir une base identique permettant la comparaison des profils d'assemblage, l'intérêt sera porté sur les profils présentant une répartition similaire de ces trois types de produits pour la suite de l'étude, à savoir les deux premiers profils. Les autres profils sont exclus dans cette étude, mais devront être étudiés de la même manière.

Ensuite, les figures 4.7 et 4.8 représentent la répartition des temps d'opération pour chacun des profils 1 et 2.

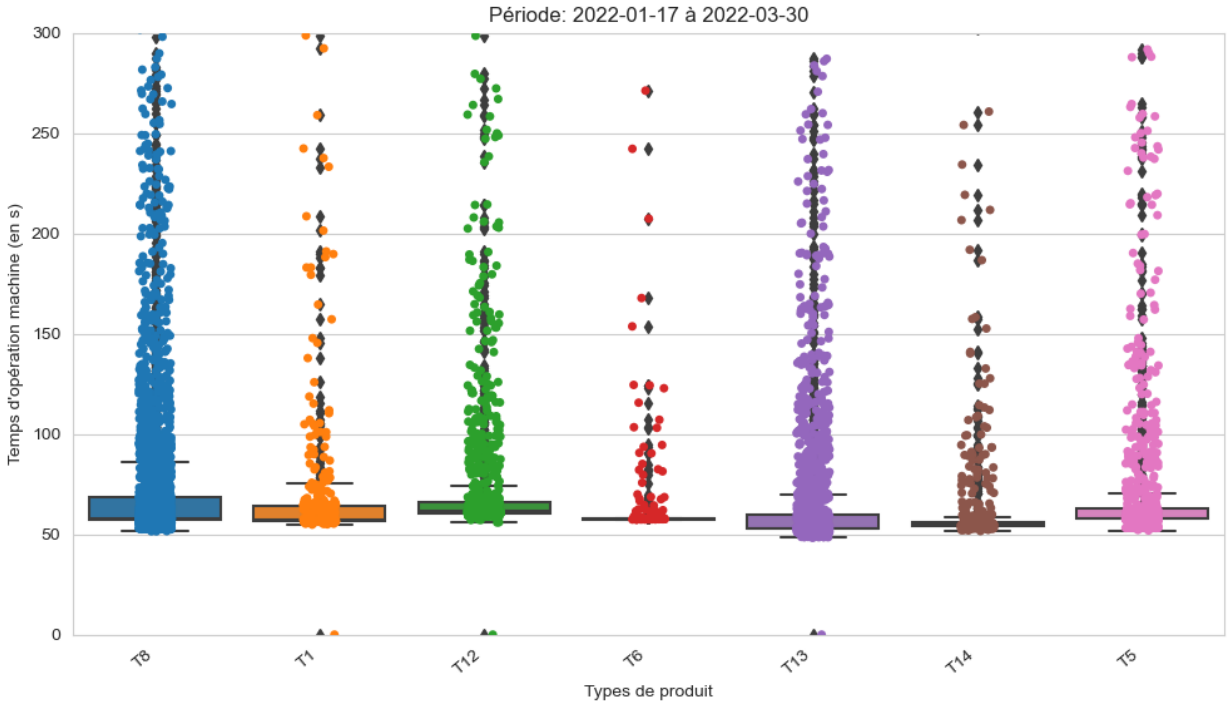


Figure 4.7 Distribution des temps d'opération par types de produits assemblés pour le profil 1

Tableau 4.5 Moyenne et médianes (en secondes) par types de produits assemblés dans le profil 1

Types de produits	Moyenne (s)	Médiane (s)
T1	50,19	56,52
T5	80,95	58,02
T6	69,81	57,70
T8	79,04	57,60

Tableau 4.5 Moyenne et médianes (en secondes) par types de produits assemblés dans le profil 1 (suite et fin)

T12	**80,70**	**61,58**
T13	83,50	52,94
T14	73,19	54,64

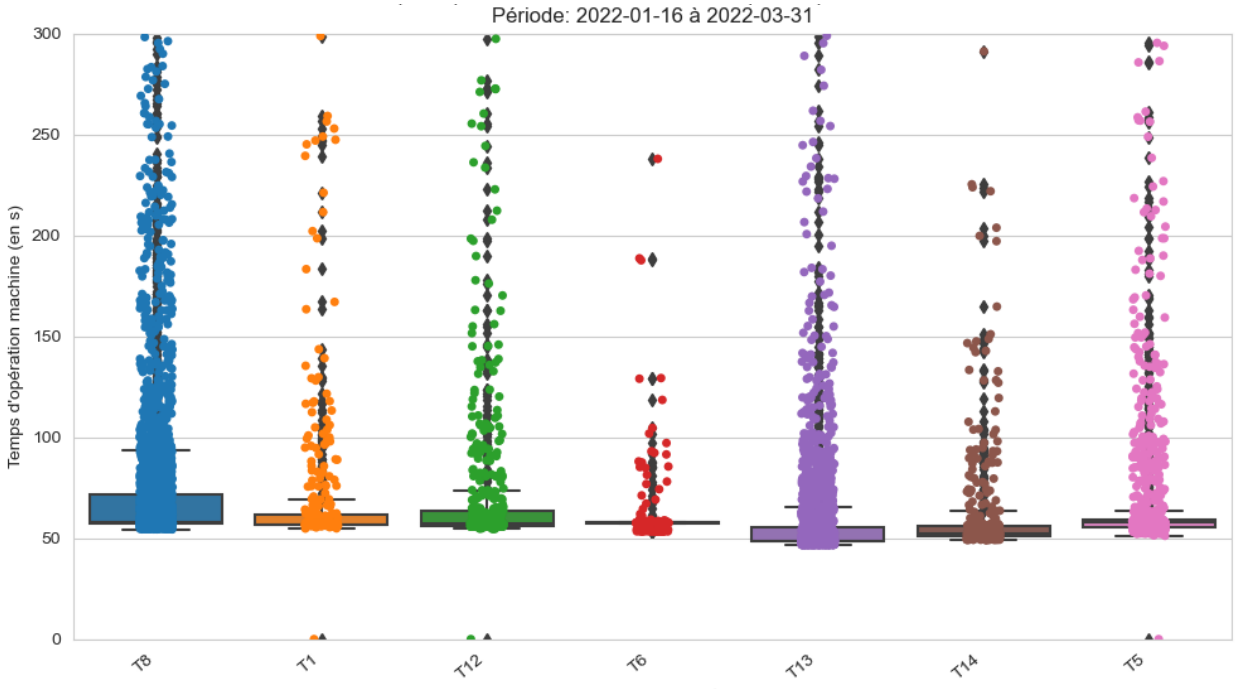


Figure 4.8 Distribution des temps d'opération par types de produits assemblés pour le profil 2

Tableau 4.6 Moyenne et médianes (en secondes) par types de produits assemblés dans le profil 2

Types de produits	Moyenne (s)	Médiane (s)
T1	73,98	56,48
T5	76,23	57,77

Tableau 4.6 Moyenne et médianes (en secondes) par types de produits assemblés dans le profil 2 (suite et fin)

T6	**67,72**	**57,65**
T8	80,36	57,57
T12	74,15	57,30
T13	67,14	48,75
T14	66,86	52,15

Les figures 4.7 et 4.8 montrent que pour chacun des deux profils, la dispersion des temps d'opération reste faible et indique une certaine régularité dans la capacité à réaliser des temps similaires pour chacun des types de produits sélectionnés. Cependant, des différences de dispersion sont observables d'un profil à l'autre. Par ailleurs, il existe une différence entre la médiane et la moyenne pour chacun des types de produits (Tables 4.5 et 4.6). Cela s'explique par la présence de nombreux points aberrants. Ces derniers peuvent être le résultat de changements de série qui nécessitent des temps d'opération plus longs qu'en cadence normale ou encore la conséquence de pertes de cadences. Les figures 4.7 et 4.8 pointent du doigt le point central de ce travail. Il existe des pertes de cadence lors de l'assemblage des différents produits sur cette machine semi-automatique. L'objectif est alors de les réduire en comprenant quels comportements mènent à de tels points aberrants ou permettent de les éviter.

Lorsque les types de produits assemblés sont identifiés et que les temps d'opération sont évalués, la métrique du nombre de produits assemblés par unité de temps est de nouveau calculée. Les tables 4.7 et 4.8 affichent le nombre moyen de produits assemblés par heure de production sur la période sélectionnée. Cette fois-ci, il est affiné par types de produits.

Tableau 4.7 Nombre moyen de produits assemblés par heure de production sur la période considérée, pour le profil 1

Types de produits	Nombre de produits assemblés	Heures de production cumulées	Nombre moyen de produits assemblés par heure de production
T8	2364	50,34	**47**
T13	1739	40,13	**43**
T14	190	3,52	**54**

Tableau 4.8 Nombre moyen de produits assemblés par heure de production sur la période considérée, pour le profil 2

Types de produits	Nombre de produits assemblés	Heures de production cumulées	Nombre moyen de produits assemblés par heure de production
T8	2762	59,99	**46**
T13	1499	28,33	**53**
T14	224	3,64	**61**

Après avoir effectué un filtrage sur les types de produits, les quantités moyennes produites restent du même ordre de grandeur lorsqu'il s'agit du nombre de produits assemblés sur la période, mais il est possible de distinguer des différences de productivité lorsque les heures de production des différents types de produits sont ciblées. Pour le type T8, majoritairement assemblé sur la machine étudiée, la productivité paraît similaire, mais un écart se creuse pour les types T13 et T14.

Cette première phase d'analyse exploratoire a permis de quantifier la production sur une période de production de trois mois, pour une machine d'assemblage sélectionnée. Elle permet notamment de mettre en lumière des profils de production comparables présentant par ailleurs des écarts de productivité qui nécessitent d'être approfondis. Pour cela, le journal des erreurs de chacune des ressources est étudié.

4.3.4.3 Statistiques d'événements

Après avoir fait un état des lieux de la productivité, l'analyse est portée sur les différents événements qui peuvent l'impacter. En effet, lors du processus d'assemblage, divers événements peuvent se produire sur la machine, qu'ils soient nécessaires au processus ou révélateurs d'une anomalie. L'objectif de cette partie est d'approfondir la mesure de la performance de production en observant les événements qui peuvent mener à une plus ou moins bonne productivité.

Dans un premier temps, la figure 4.9 donne accès à trois statistiques calculées à partir du journal d'événements.

```
Période: 2022-01-17 à 2022-03-30
Nombre total d'événements: 4778

95.1% du total d'événements, soit 4544 événements,
        apparaissent lorsque la machine est en fonctionnement contrôlé.
15.67% de ces événements, soit 712 événements, mènent à un ralentissement.
45.53% de ces événements, soit 2069 événements, mènent à un arrêt.
```

a) Profil 1

```
Période: 2022-01-16 à 2022-03-31
Nombre total d'événements: 4563

94.46% du total d'événements, soit 4310 événements,
        apparaissent lorsque la machine est en fonctionnement contrôlé.
15.17% de ces événements, soit 654 événements, mènent à un ralentissement.
48.68% de ces événements, soit 2098 événements, mènent à un arrêt.
```

b) Profil 2

Figure 4.9 Statistiques événementielles pour les deux profils étudiés sur la période considérée

Le journal d'événements recensant les différentes occurrences de codes relevés sur la machine, la figure 4.9 permet de savoir si le code indique que :

- la machine est à l'origine de l'occurrence d'un événement : l'événement peut se produire lorsque la machine est en fonctionnement contrôlé ou non;
- la machine est susceptible de ralentir : certains événements nécessitent un ralentissement de la cadence d'assemblage au-delà d'une certaine durée, comme les redémarrages après des changements de série; et
- la machine est susceptible de s'arrêter : certains événements nécessitent un arrêt total et soudain de la machine ou après une certaine durée, avec par exemple, le besoin d'une intervention extérieure, ou par manque de composantes.

Ces statistiques permettent d'avoir une première approche générale des différents types de pertes de cadence auxquels la machine est confrontée pour les deux profils sélectionnés. Plus de 90% des événements ont lorsque la machine est en fonctionnement contrôlé. Il est aussi indiqué, pour chacun des profils, les proportions d'événements menant à un ralentissement ou à un arrêt, parmi la totalité d'événements relevés sur la période d'étude. D'après la figure 4.9, le profil présentant le plus d'événements (4778 contre 4563) est le profil ayant les nombres moyens de produits assemblés par heure de production pour les types T13 et T14 les plus faibles (Tables 4.7 et 4.8).

Enfin, l'analyse porte sur la proportion des codes d'événements présents dans l'échantillon sélectionné. La figure 4.10 présente la proportion des occurrences de chacun des événements de chacun de profil, dès lors qu'elle est supérieure à 2% (pour une question de lisibilité).

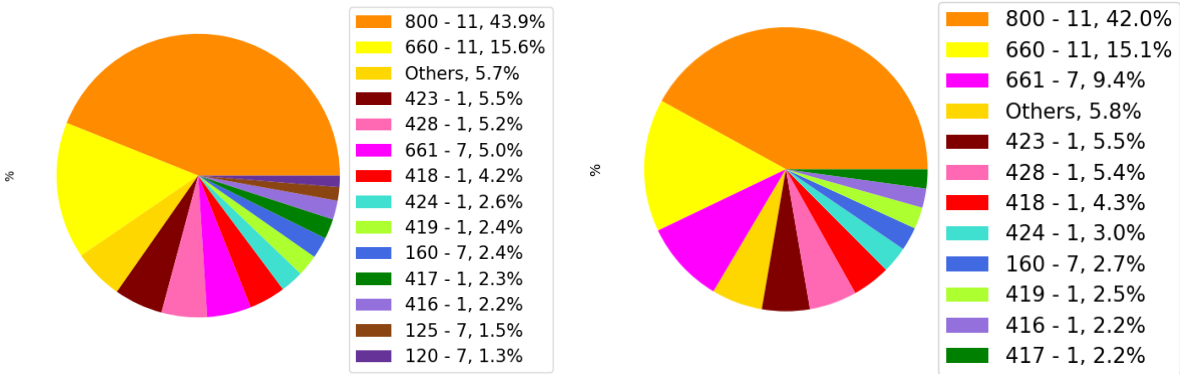


Figure 4.10 Proportions des événements rencontrés sur la période considérée, pour le profil 1 (gauche) et le profil 2 (droite)

Les deux profils affichent des ordres de grandeur d'événements similaires. Les codes les plus fréquents sont de la catégorie 11. Le code le plus récurrent est le code 800 avec plus de 40% d'occurrences. Il fait référence à un retour en production nominale. S'en suit le code 660 avec environ 15%, il correspond à une anomalie de matériel hors tolérance. Le code 661 de la catégorie 7 est en pratique celui indiquant la correction de l'anomalie révélée par le code 660. Il est présent en plus grande proportion pour l'un des deux profils. Une première distinction de comportements se dessine ici. Un second code de la catégorie 7, le code 160, apparaît, mais avec moins d'occurrences. Le reste des codes ayant une proportion inférieure à 6% appartiennent à la catégorie 1, relative à des besoins de changements de diverses composantes.

Concernant les codes de production nominale et de changements de composante et selon la connaissance métier, ces codes sont nécessaires au processus d'assemblage, mais peuvent s'avérer révélateurs d'une différence dans la manière d'appréhender la machine. Les codes relatifs à des anomalies sont alors les codes 660, 661 et 160. Un protocole est à suivre en cas d'apparition de matériel non conforme, à savoir qu'il doit être corrigé. L'un des profils semble effectuer plus de corrections que le second. Cette première distinction permet d'affiner les profils de production avec des réactions aux anomalies différentes.

Cette première phase d'analyse a permis de mettre en lumière différents profils de production, menant à des niveaux de productivité différents. Cependant, cette première approche n'est pas suffisante lorsqu'il s'agit de mettre en lumière des bonnes pratiques. En effet, elle permet de visualiser l'état de la productivité, mais ne permet pas de comprendre ce qui amène à différents niveaux de performance et ce qui permet de l'améliorer. Cette phase de mesure de productivité ouvre ainsi la porte à une nécessité de comprendre les réactions face aux événements ayant lieu sur la machine. Il s'agit alors de retracer les différents comportements qui peuvent mener à ces différents niveaux de productivité, notamment grâce à l'étude des processus.

4.3.5 Découverte de processus

Après avoir noté l'existence de différents niveaux de productivité, la suite de l'étude vise à analyser les évènements associés à ces situations, pour ensuite identifier les bonnes pratiques à reproduire. L'exploration de processus, via les graphes dirigés, est l'outil sélectionné dans cette démarche d'identification de bonnes pratiques. Elle permet ici de tracer un produit au cours du processus

d'assemblage, en suivant les différentes étapes par lesquelles il passe. Pour cela, une nouvelle bibliothèque du langage *Python* est utilisée, à savoir *PM4Py*.

4.3.5.1 Présentation du processus

Dans le cadre de cette étude, le suivi des différents états d'un produit permettra de tracer le processus réalisé. Les états sont décrits par les codes présentés précédemment. Ils peuvent faire référence à plusieurs types d'événements :

- en production;
- matériel non conforme;
- changement de composantes;
- maintenance et arrêts planifiés;
- ajustements de paramètres pour changement de série; et
- diverses anomalies.

Ainsi, lors de l'assemblage de plusieurs produits, divers événements peuvent se produire, altérant ainsi les temps d'opération et donc la productivité. Afin d'exploiter ces données grâce aux outils de découverte de processus, la table finale des données des journaux de production et d'événements est utilisée. Comme expliqué précédemment, elle permet d'associer l'événement ou la succession d'événements au produit assemblé auquel ils se réfèrent. Dans cette table, ne sont conservées que les lignes représentant l'occurrence d'un événement. Il en résulte alors une table recensant une succession d'événements chronologiques associés à un produit. Jusqu'à présent, il était possible de distinguer les codes apparaissant sur la machine lorsque cette dernière était à l'arrêt pour opération de maintenance ou au contraire en production (ou censée l'être). Dans cette partie de l'analyse, cette distinction n'est plus possible, car elle générerait des erreurs dans les données retournées. Dès lors que les informations sont rassemblées, l'objectif de cette étude de processus est d'analyser les différents codes ou séquences de codes qui peuvent intervenir lors de l'assemblage d'un produit, afin de déceler les transitions critiques entre états. L'apparition de certains codes nécessitant une intervention externe, les graphes permettront d'analyser des réactions et ainsi de comprendre avec plus de finesse les profils de productivité évalué au 4.3.4.

4.3.5.2 Modélisation du processus sous forme de graphes dirigés

Afin de poursuivre cette analyse, des graphes dirigés sont tracés à l'aide de la bibliothèque *PM4PY* de Python. Elle permet de transformer un *DataFrame* en « *journal d'événements* », recensant trois informations principales:

- l'identifiant unique de chaque produit;
- l'« événement » ou état par lequel le produit passe, c.-à-d. le code d'événement; et
- l'horodatage de l'événement.

Ces graphes permettent de visualiser les transitions entre les différents codes pour la machine étudiée, durant la période sélectionnée. Ils peuvent être décorés de deux informations : le nombre total d'occurrences de chacun des codes et chacune des transitions, ainsi que d'une information temporelle au choix. Ici, il s'agira de la durée moyenne de chacune des transitions. Elle correspond au temps passé dans l'état initial avant de passer à l'état suivant.

Dans un premier temps, les transitions sont évaluées par leur nombre d'occurrences. Le graphe dirigé des événements ayant eu lieu sur la machine et la période sélectionnées est affiché. La figure 4.11 présente le graphe dirigé obtenu avec les occurrences des événements et des transitions.

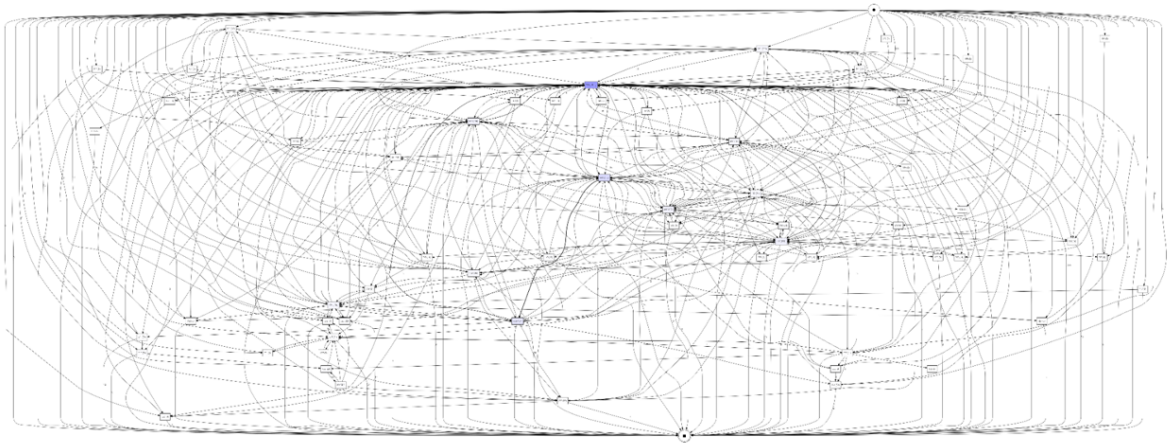
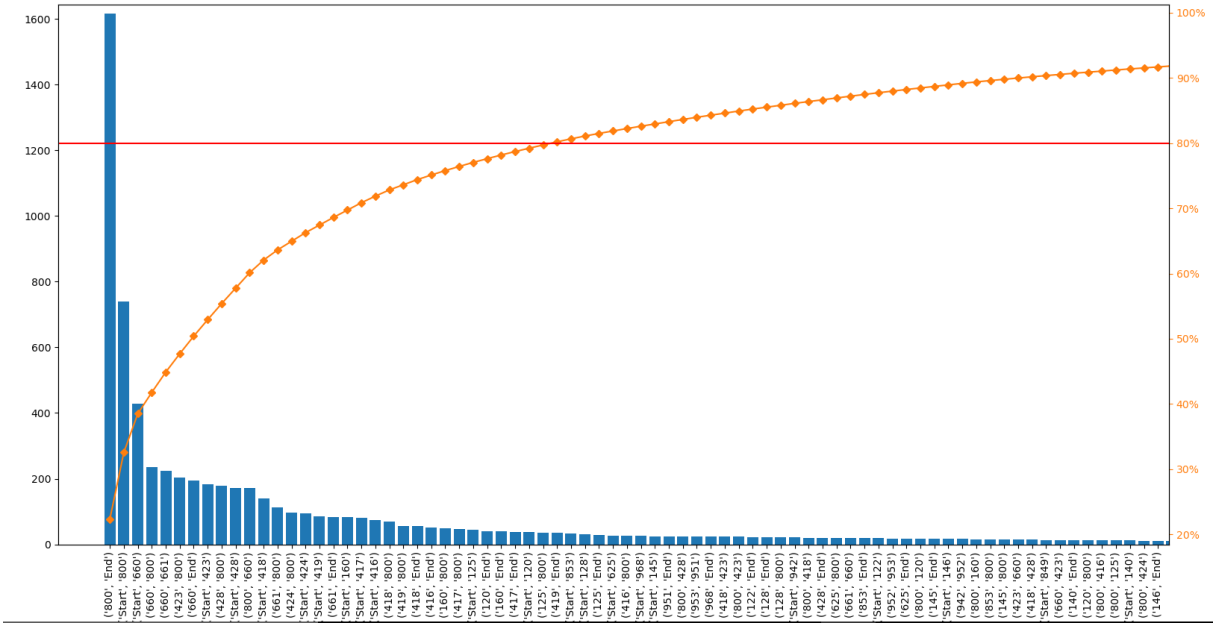


Figure 4.11 Graphes dirigés « spaghetti »

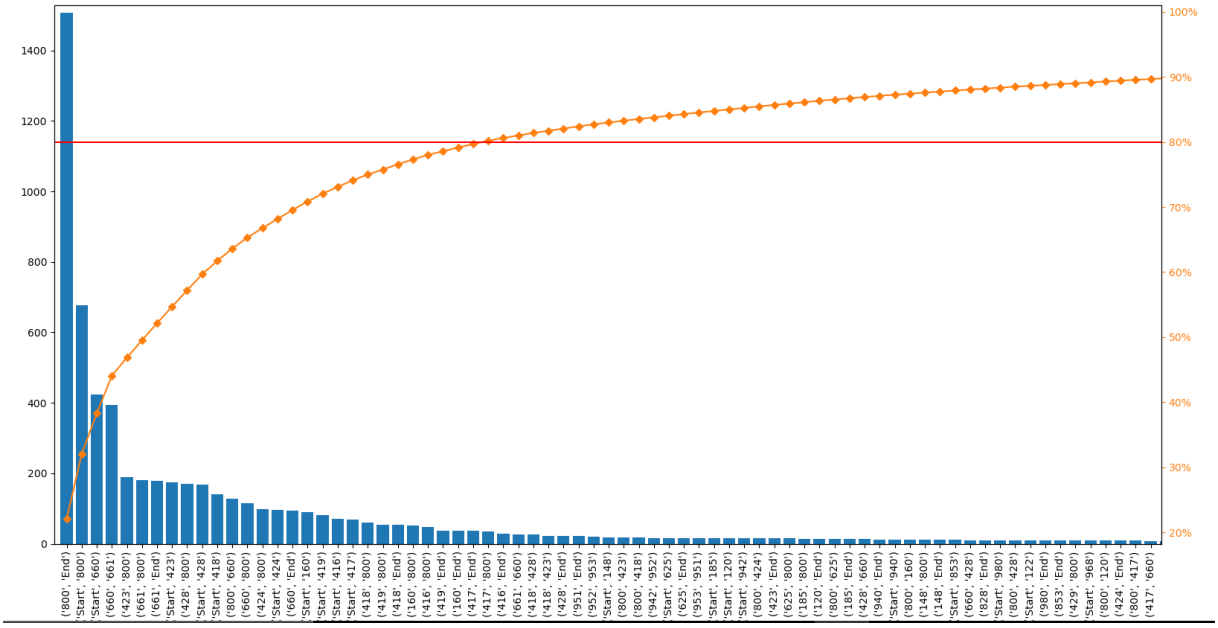
Lorsque ce graphe est agrandi, plusieurs transitions affichent des occurrences faibles au regard de la période d'étude. Ainsi, pour des raisons de lisibilité, le choix est fait de le filtrer. Cela permettra de mieux appréhender les transitions critiques affectant la cadence de production.

4.3.5.3 Filtrage des graphes

Le filtrage des graphes dirigés est une affaire délicate (Van Der Aalst, 2019). En effet, s'il est mal fait, il peut conduire à des conclusions erronées et donc à des interprétations questionnables. Par exemple, le filtrage concernant le nombre d'occurrences d'un événement vise à supprimer les événements au-dessous d'une certaine valeur. Si des événements dont supprimés, il est possible que de nouvelles transitions apparaissent bien qu'elles ne représentent pas le processus initial. Afin de pallier ce problème, le graphe dirigé est filtré dans le but de n'afficher que les informations qui sont jugées pertinentes. Dans le cas présent, deux filtrages seront appliqués aux transitions. Cela permet de conserver les transitions intactes, sans créer de raccourci et altérer les données. Dans un premier temps, seules les transitions les plus fréquentes sont conservées, puis les transitions les plus chronophages. D'après le principe de Pareto, 80% des conséquences sont issues de 20% des causes. Deux graphes de Pareto sont alors tracés (Figures 4.12 et 4.13).



a) Profil 1



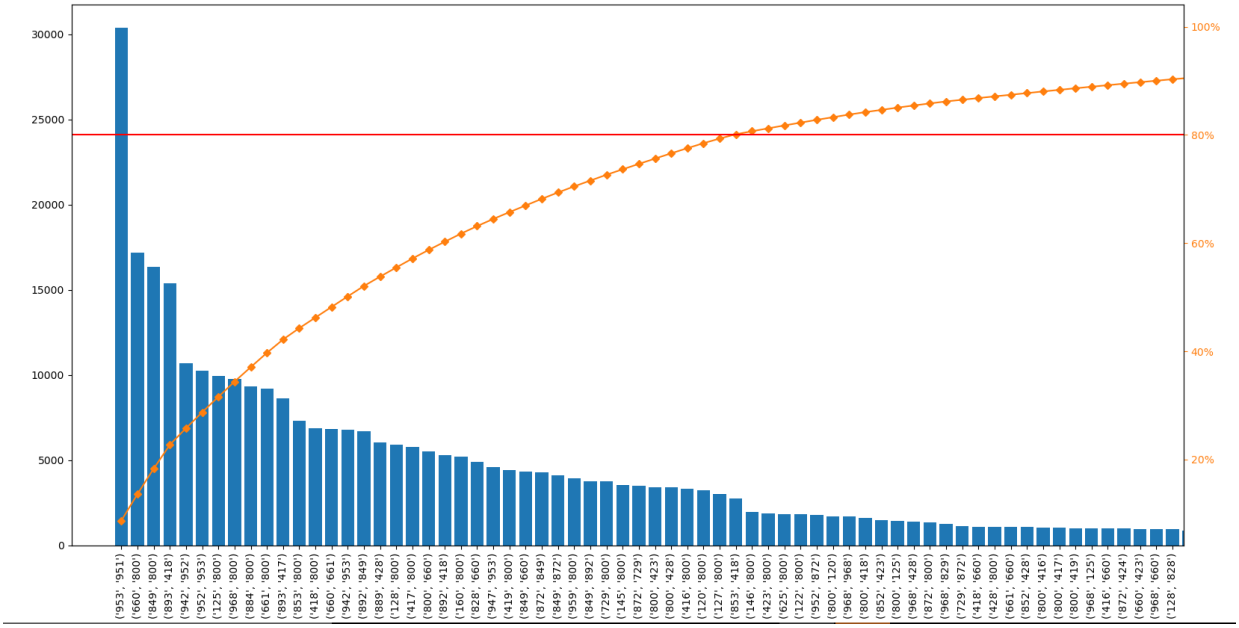
b) Profil 2

Figure 4.12 Graphe de Pareto des occurrences des transitions

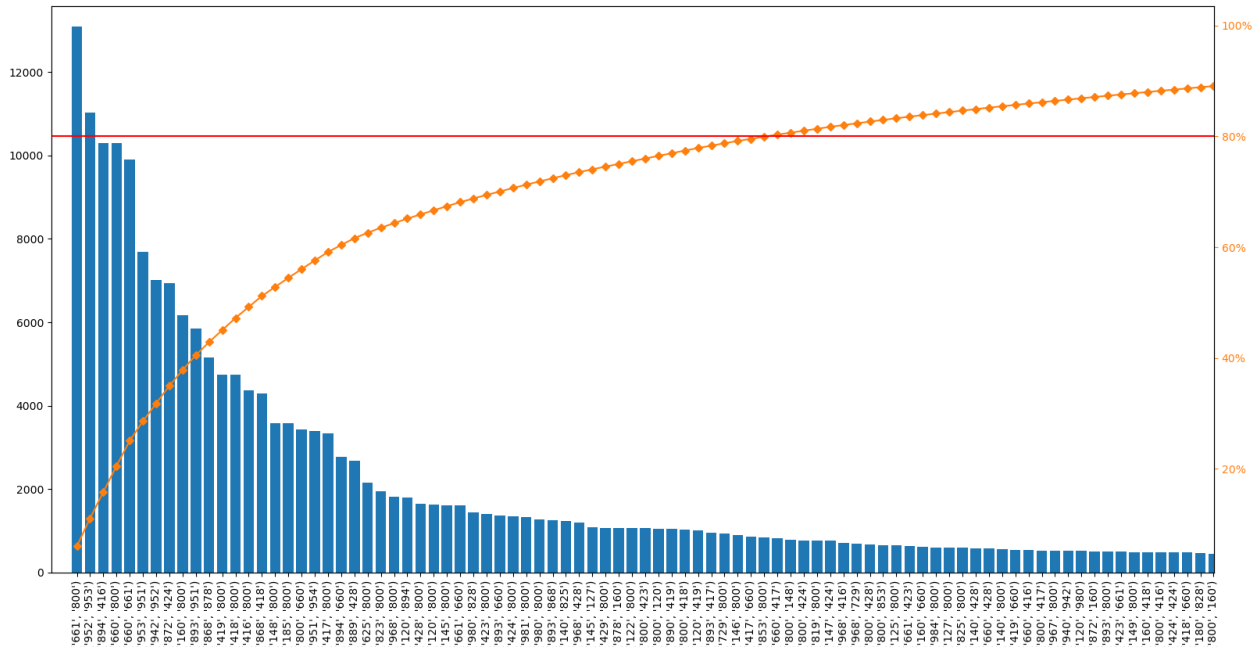
Ainsi, le graphe de Pareto affiche la distribution décroissante des occurrences des transitions relevées sur la machine et la période sélectionnées, ainsi que le pourcentage cumulé des proportions de chacune des transitions parmi l'intégralité des transitions sur la figure 4.11.

Pour chacun des profils, les transitions qui représentent 80% des transitions les plus fréquentes en cumulé sont conservées.

Le deuxième filtre à partir du principe de Pareto se fait sur la somme des durées des transitions (Figure 4.13). Ainsi, seules les transitions représentant en cumulé 80% de la durée totale des transitions sont conservées.



a) Profil 1



b) Profil 2

Figure 4.13 Graphe de Pareto des durées des transitions

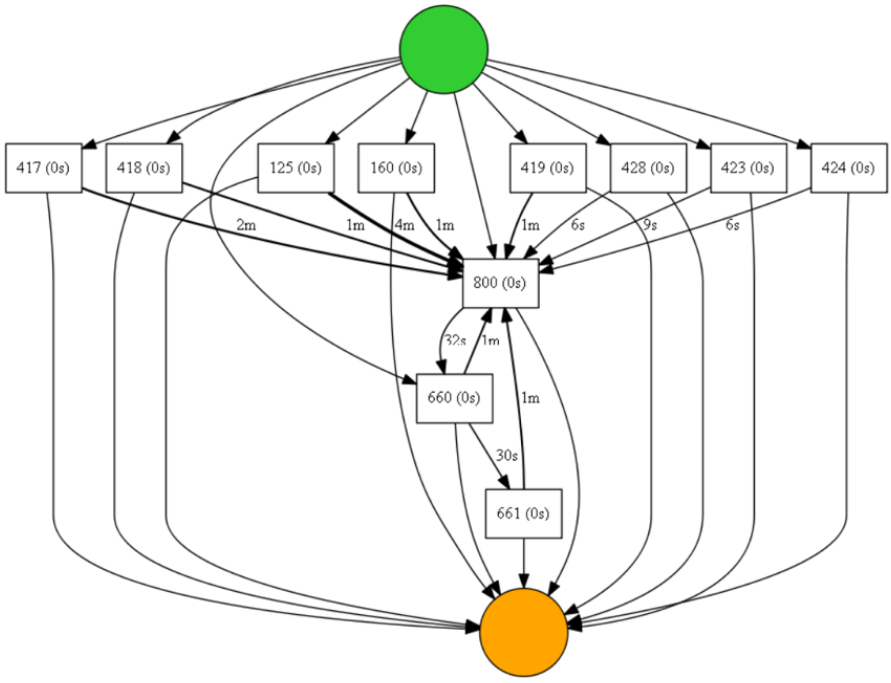
Ces graphes de Pareto affichent la distribution décroissante des durées des transitions relevées sur la machine et la période sélectionnées, ainsi que le temps cumulé des durées de chacune des transitions parmi l'intégralité des transitions.

Par souci de cohérence et afin de garder le graphe connecté, l'algorithme effectue un second tri parmi les transitions préalablement filtrées (Fraunhofer Institute for Applied Information Technology, 2021).

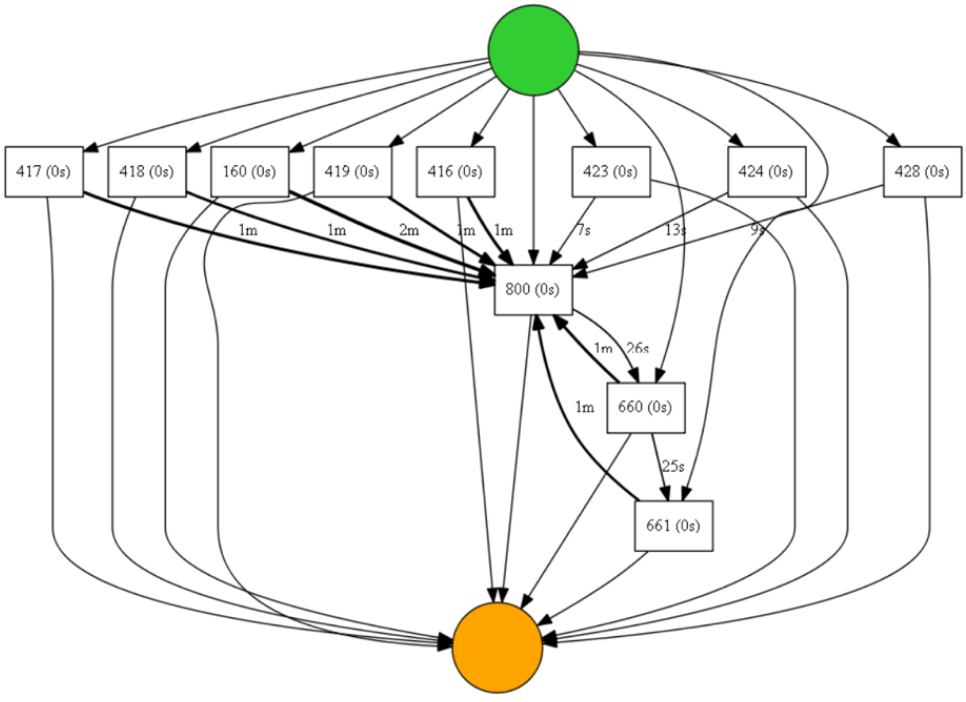
4.3.5.4 Identification et étude des transitions fréquentes

Évaluation des durées moyennes des transitions les plus fréquentes

Dans un premier temps, l'analyse porte sur la durée moyenne de chacune des transitions les plus fréquentes. Ainsi le graphe obtenu affiche les codes permettant le traçage des transitions les plus fréquentes (Figure 4.14).



a) Profil 1



b) Profil 2

Figure 4.14 DFGs filtrés selon es transitions les plus fréquentes, pondérés des durées moyennes

Tableau 4.9 Durées moyennes précisées en secondes

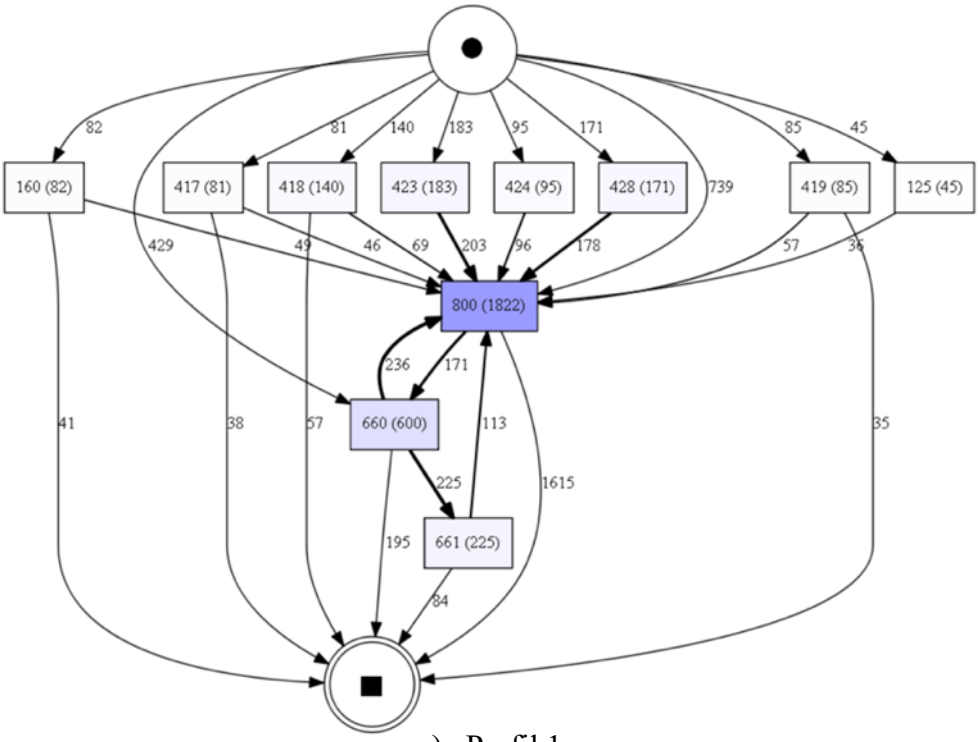
Profil 1		*Profil 2*	
Transition	**Durée moyenne (en s)**	**Transition**	**Durée moyenne (en s)**
417 – 800	126	**417 – 800**	96
418 – 800	99	**418 – 800**	79
160 - 800	106	**160 - 800**	121
419 – 800	77	**419 – 800**	88
660 – 800	72	**660 – 800**	89
661 – 800	81	**661 – 800**	72
125 – 800	277	**416 – 800**	91

Les figures 4.14.a et 4.14.b montrent différents éléments. Les points vert et orange représentent respectivement le début et la fin du processus d'assemblage. Chaque lien représente une transition directe d'un code vers un autre. Dans chacune des cases est inscrit un code événement. Les durées moyennes arrondies de transitions sont exprimées en minutes ou en secondes. Il est rappelé que la durée affichée indique, en réalité, le temps passé dans l'état initial de la transition avant de passer à l'état final. La table 4.10 précise alors les durées affichées en minutes sur les graphes. Certaines durées de transitions dépassent la minute, ce qui est significatif au regard d'un temps d'opération qui, pour tous types de produits confondus, ne dépasse pas 60s en cadence nominale. Par ailleurs, des écarts peuvent être observés entre les deux profils concernant des retours en production, identifiés par le code 800 (Table 4.10). Pour certains événements, le profil 1 semble mettre plus de temps que le profil 2 et inversement. La connaissance métier indique qu'une transition d'un changement de composante (codes 400) vers un retour en production peut mener à un arrêt de la machine si elle dépasse une certaine durée et témoigne alors d'une mauvaise anticipation. De la même façon, la transition d'une correction de matériel non conforme (code 661) vers un retour en

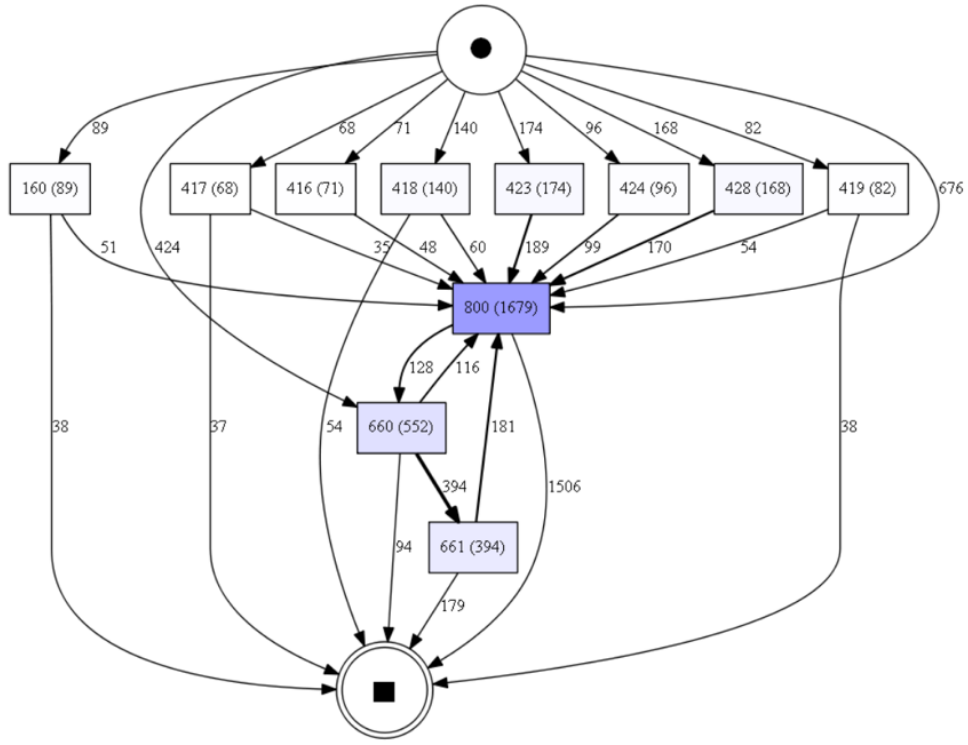
production peut être révélatrice d'une plus ou moins bonne gestion de la correction de cette non-conformité. Des différences de durées sont remarquables entre les deux profils et ouvrent une nouvelle piste d'exploration des divers comportements dans la gestion des événements qui ont lieu au cours du processus d'assemblage. Afin de savoir si ces transitions sont réellement critiques, il convient de vérifier leurs niveaux de fréquences.

Évaluation des occurrences des transitions les plus fréquentes

L'analyse porte, dans un second temps, sur le nombre d'occurrences de chacune des transitions étudiées. De la même façon que précédemment, la figure 4.15 affiche les graphes dirigés filtrés des occurrences pour chacun des deux profils.



a) Profil 1



b) Profil 2

Figure 4.15 DFGs filtrés selon les transitions les plus fréquentes, pondérés des occurrences

De la même façon que les graphes précédents, le point et le carré indiquent respectivement le début et la fin du processus d'assemblage. Les figures 4.15.a et 4.15.b indiquent respectivement les mêmes codes que les deux graphes précédents. Ils montrent cependant le nombre d'occurrences de chacune des transitions conservées, au lieu de la durée moyenne. On observe des ordres de grandeur similaires pour des événements inévitables du processus, à savoir les changements de composantes (codes 400). Cependant, des écarts se dessinent pour des événements relatifs à des anomalies (code 660, 661,160 et 125). Certains codes ne sont d'ailleurs pas partagés par les deux graphes. Cela vient ajouter une nouvelle distinction dans les profils de réactions. Les codes 800, 660 et 661 étant communs aux deux profils sélectionnés et permettant d'étudier une sous-partie complète du processus, le parti pris de cette étude est de s'intéresser aux transitions entre ces trois codes.

Matériel non conforme et correction

Dans cette partie, les graphes sont agrandis sur la boucle impliquant l'anomalie de matériel non-conforme (code 600), sa correction (code 661) ainsi que le retour en production (code 800).

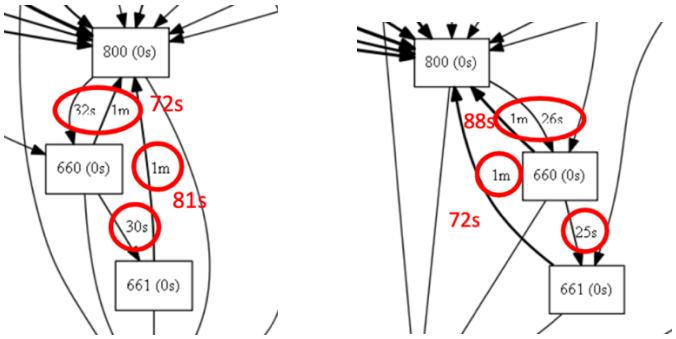


Figure 4.16 DFGs filtrés selon les transitions les plus fréquentes, pondérés des durées moyennes, agrandis pour le profil 1 (gauche) et le profil 2 (droite)

Le profil affichant un nombre moyen de produits assemblés par heure de production tous types de produits confondus supérieur (48 contre 46) semble :

- voir apparaître moins rapidement une non-conformité (800 vers 660);
- choisir plus rapidement de corriger la non-conformité (660 vers 661);
- corriger plus rapidement la non-conformité (661 vers 800); et

- choisir plus lentement de ne pas corriger la non-conformité (660 vers 800).

En ce qui concerne les nombres d'occurrences, les graphes dirigés agrandis communiquent les informations suivantes.

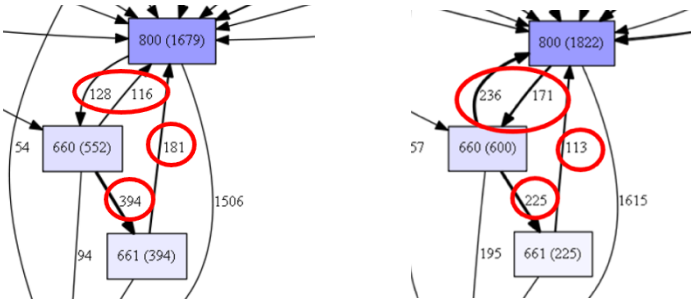


Figure 4.17 DFGs filtrés selon les transitions les plus fréquentes, pondérés des occurrences, agrandies pour le profil 1 (gauche) et le profil 2 (droite)

Le profil affichant un nombre moyen de produits assemblés par heure de production tous types de produits confondus supérieur (48 contre 46) semble :

- voir apparaître moins de non-conformités (800 vers 660);
- corriger plus de non-conformités (660 vers 661); et
- ignorer moins de non-conformités (660 vers 800).

Ces deux analyses permettent de mettre en lumière des comportements d'assemblage différents dans le cas particulier de la gestion des non-conformités. Afin d'approfondir l'analyse et de comprendre l'écart de productivité évoqué au 4.3.4.2 sur deux types de produits, les mêmes graphes sont affichés à partir de données filtrées par type de produits.

Étude par types de produit

Ici, le type de produits affichant le plus grand écart du nombre moyen de produits assemblés par heure de production est étudié (Table 4.11). La même démarche peut être utilisée pour les autres types de produits et peut mener à des conclusions différentes.

Tableau 4.10 Nombre moyen de produits assemblés par heure de production et par type de produits pour les deux profils étudiés

Type de produits	Profil 1	Profil 2
T8	46	47
T13	***53***	***43***
T14	54	61

De la même façon que précédemment, les graphes sont filtrés selon les transitions les plus fréquentes et les graphes dirigés agrandis sur la boucle relative au matériel non conforme sont affichés, pour le durées moyennes des transitions sélectionnées, ainsi que leurs occurrences.

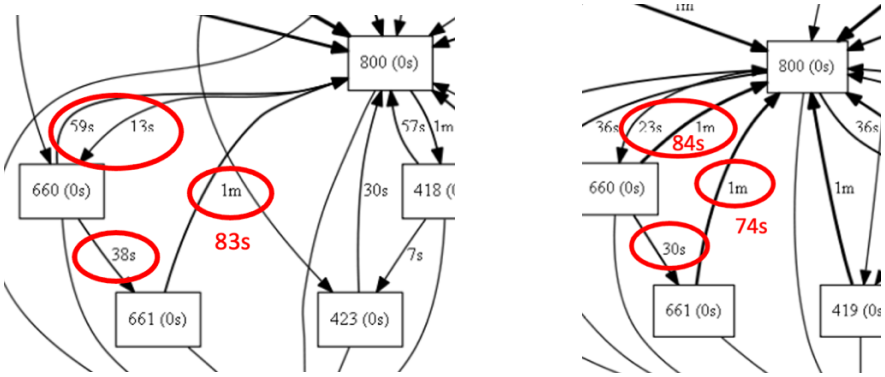


Figure 4.18 DFGs filtrés selon les transitions les plus fréquentes, pondérés des durées moyennes pour le type de produits T13, agrandis, pour le profil 1 (gauche) et le profil 2 (droite)

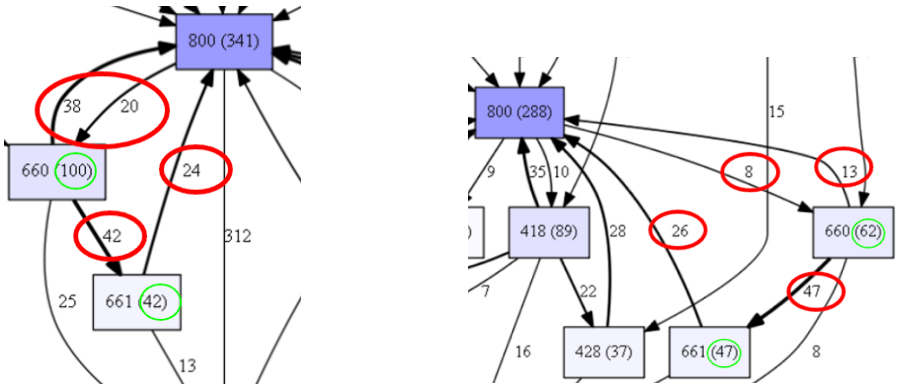
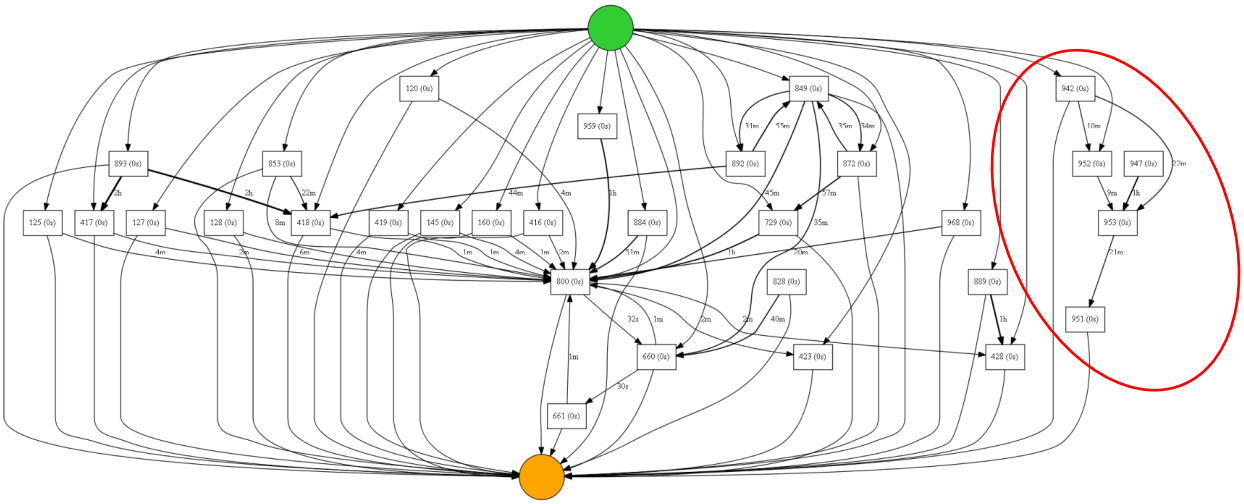


Figure 4.19 DFGs filtrés selon les transitions les plus fréquentes, pondérés des occurrences pour le type de produits T13, agrandis, pour le profil 1 (gauche) et le profil 2 (droite)

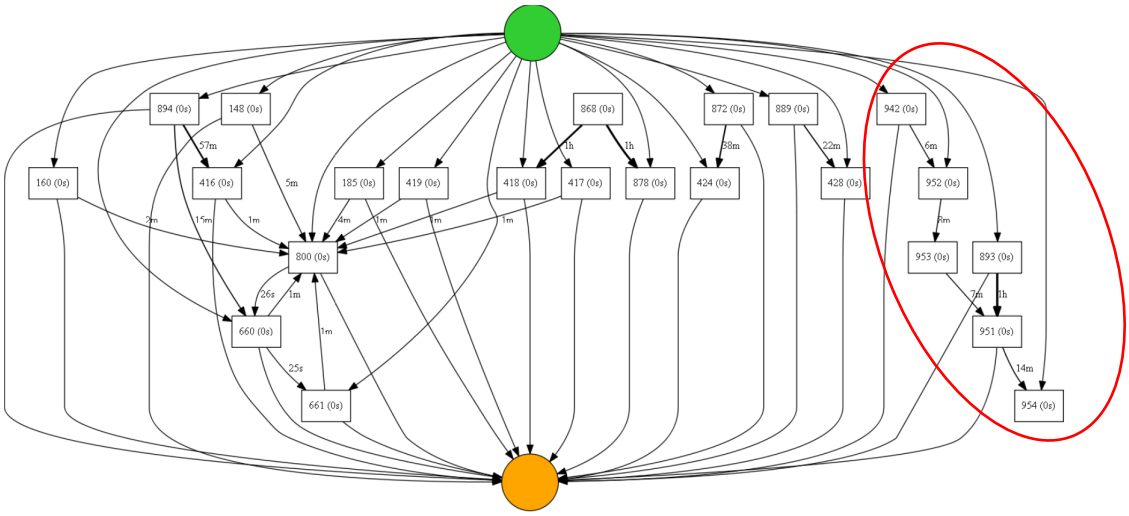
Les figures 4.18 et 4.19 affichent respectivement les durées moyennes de transitions et les occurrences des transitions pour le sous-processus sélectionné et pour le type T13. Les valeurs entourées en rouge sont celles sur lesquels l'intérêt doit être porté. Les valeurs inscrites en rouge sur la figure 4.18 sont les valeurs exactes des durées moyennes, arrondies sur les graphes initiaux. Les valeurs encerclées en rouge et en vert sont les données sur lesquelles il faut se concentrer. Bien que les écarts soient moins flagrants, car l'échantillon de données est plus petit, ces graphes confirment les conclusions faites depuis l'analyse sans distinction des types de produits. Il est ainsi possible d'aller jusqu'à distinguer les différents types de produits pour observer des comportements différents, menant à des niveaux différents de productivité. Il est donc possible de cibler des réactions différentes afin d'éventuellement mettre en place des pistes d'amélioration adéquates, allant jusqu'à préciser le type de produit sur lequel porter attention.

4.3.5.5 Identification et étude des transitions chronophages

Cette partie a pour objectif de s'assurer qu'au-delà des transitions les plus fréquentes, certaines transitions moins récurrentes n'ont pas une durée démesurée vis-à-vis du processus d'assemblage. La figure 4.20 affiche les transitions représentant 80% de la durée totale de toutes les transitions sur la période choisie pour chacun des deux profils d'étude.



a) Profil 1



b) Profil 2

Figure 4.20 DFGs filtrés selon les transitions les plus chronophages, pondérés des durées moyennes

De nouveau, ce sont les durées moyennes arrondies de chacune des transitions conservées qui sont affichées. De nouveaux codes sont apparus au regard des graphes des occurrences étudiés précédemment, notamment des codes d'anomalies, mais l'une des différences qui selon la connaissance métier pourrait s'avérer révélatrice est l'apparition de codes d'ajustements de paramètres lors de changement de série. Cette sous-partie du processus d'assemblage pourrait effectivement être explorée afin d'évaluer la capacité d'atteindre les paramètres de changements

de série adéquats plus ou moins rapidement. À titre d'exemple, des différences de durées importantes apparaissent. Elles pourraient être approfondies en observant les différents profils.

4.3.6 Conclusion

Finalement, ce cas d'étude a permis de mettre en œuvre, sur un cas réel, la démarche de valorisation de données industrielles pour l'évaluation des pertes de cadence, développée au chapitre 3, et d'en confirmer la faisabilité. Après la compréhension et la préparation des données, différents graphes résultant de l'utilisation d'outils d'analyse exploratoire ont permis de rassembler des caractéristiques de production et de construire des profils de production similaires présentant des niveaux de productivité différents. Afin de comprendre ces écarts, la découverte de processus a été mise en œuvre au moyen de graphes dirigés, mettant en relation différents événements ayant lieu sur une machine. Ces graphes affichent les transitions critiques du processus d'assemblage, susceptible de causer des pertes de cadence. L'apparition de divers types d'événements est finalement appréhendée différemment selon les profils et les réactions diffèrent selon leur récurrence et leur durée lors du processus d'assemblage. L'étude de graphes dirigés a donc mené à une meilleure compréhension et à l'affinement des profils d'assemblage en ajoutant des informations relatives aux différents comportements face à la machine. L'analyse de ce cas pratique avec la démarche proposée a mis en lumière différentes pistes d'amélioration des comportements d'assemblage. En effet, si les différences relevées entre deux profils sont observables sur le terrain, elles pourraient permettre d'effectuer du transfert de connaissances ciblé entre les différents profils, afin d'améliorer les pratiques et donc la productivité. Le chapitre suivant viendra conclure ce travail et y apportera des critiques et des perspectives d'approfondissement.

CHAPITRE 5 CONCLUSION ET RECOMMANDATIONS

L'objectif principal de ce mémoire était de *réduire les pertes de cadence dans un contexte de production semi-automatique.* Pour y arriver, nous avons proposé une démarche d'identification des bonnes pratiques qui s'inspire de la méthodologie CRISP-DM et repose sur des outils d'analyse exploratoire et de découverte de processus. Cette démarche se démarque des autres approches que nous retrouvons dans la littérature scientifique d'une part grâce à la combinaison de ces outils pour identifier des pratiques de production et, d'autre part, à l'intérêt porté à la visualisation de données, permettant une meilleure implication des experts du domaine. Pour valider cette démarche, nous l'avons appliquée à un cas d'étude. Nos résultats démontrent que la méthode permet effectivement d'identifier différents comportements face à des étapes ciblées du processus d'assemblage, se référant à des niveaux de productivité différents. Ainsi, la méthodologie développée dans ce mémoire permet d'identifier des pratiques ciblées qui pourraient être améliorées ou transférées afin de réduire les pertes de cadence. Des pistes d'approfondissement de ce travail sont proposées afin de le bonifier et d'ouvrir la porte à d'autres analyses. En revanche, nous n'avons pas été capables de vérifier si nous atteignions réellement une réduction des pertes de cadence en raison du temps requis, mais cette validation se fera ultérieurement chez notre partenaire industriel. Toutefois, la démarche a permis d'identifier de façon concrète des pistes de solutions.

La démarche proposée témoigne de certaines limitations. Tout d'abord, le cas d'étude repose sur différentes hypothèses. En effet, il ne concerne qu'une seule machine et se limite à une période de 3 mois. Une piste d'approfondissement consisterait en l'application de la méthode développée à d'autres machines d'assemblage et en la considération d'une période d'études différentes.

Par ailleurs, au fur et à mesure de l'étude, l'analyse est de plus en plus ciblée, arrivant à l'étude d'une sous-partie du processus. Une autre piste d'amélioration serait d'étudier d'autres parties du processus afin d'en acquérir de nouvelles connaissances pertinentes.

Ensuite, différents critères de regroupement et d'exclusion pour la sélection des profils de production sont utilisés dans ce travail. Ces critères sont choisis à partir de l'expérience des professionnels du domaine, mais ils pourraient être différents et mèneraient probablement à des conclusions différentes, pour un autre cas pratique par exemple. Une analyse systématique pourrait être développée afin de guider et clarifier le choix de ces différents critères.

La méthode présentée dans ce projet s'appuie sur un outil de découverte de processus, à savoir les graphes dirigés. Comme évoqué au chapitre 2, peu de travaux scientifiques utilisent des méthodes d'exploration de processus pour l'évaluation des pertes de cadence en production. Un article cependant utilise les graphes dirigés. En effet, Bhogal et Garg (2020) les utilisent afin d'analyser les performances de production et d'identifier des motifs de pannes. Ils souhaitent générer de la connaissance pour réduire les pertes de cadence engendrer par les pannes et le processus de réparation. Cependant, ils traitent peu de données, qui plus est relevées manuellement, donc ils ne font pas face à des problèmes de filtrage, comme présenté au chapitre précédent. Par ailleurs, ils présentent une évaluation de la performance épurée. En effet, la performance du processus repose sur la durée des différentes étapes, des différentes activités individuelles ainsi que du temps écoulé entre différents événements (Bhogal et Garg, 2020). Ce travail a montré que la simple lecture d'un graphe dirigé affichant la performance temporelle n'est pas suffisante dans une démarche d'évaluation de la performance. Par ailleurs, ils ne vont pas jusqu'à cibler des comportements permettant l'amélioration des pratiques.

Concernant les graphes dirigés, Van Der Aalst (2019) présente les limites de leur interprétation après des opérations de filtrage. En effet, comme évoqué au chapitre 3, un mauvais filtrage des données peut mener à la création de nouvelles transitions inexistantes dans les différents cas parcourant un processus étudié. L'algorithme menant à l'affichage de graphes dirigés possède des paramètres de seuil d'occurrence ou de durées, permettant de ne conserver que les événements qui satisfassent ces différents seuils. Ainsi, la suppression d'événements peut créer des raccourcis. Les occurrences et les durées des transitions peuvent être mal recalculées et l'interprétation de ces graphes erronée peut mener à de fausses conclusions. Ainsi, une grande vigilance doit être accordée au choix du filtre. Dans ce travail, le filtre utilisé est un filtre d'affichage. Les transitions les plus fréquentes ou les plus chronophages sont sélectionnées au préalable parmi toutes les transitions puis seules ces transitions sont affichées.

Le cas d'étude est mené sur une machine constituée de 3 parties fonctionnant en parallèle, complexifiant de manière significative le processus. Les temps d'opération utilisés dans cette analyse sont ceux calculés par la machine. Par ailleurs, la connaissance métier indique que cette machine assemble un produit en un nombre donné d'étapes. Pour avoir une véritable évaluation du temps d'opération, il faudrait étudier le retour de la machine à l'étape 1 du processus.

Le cas d'étude a permis de montrer que dans le cas du partenaire, il est possible d'identifier des pratiques à améliorer dans un but de réduction des pertes de cadence. Cependant, des pistes restent à creuser. En effet, dans le cas du partenaire industriel, les comportements adoptés face à des machines d'une telle envergure sont très normés. Par ailleurs, lors du processus d'assemblage, des corrections manuelles sont nécessaires, entraînant des arrêts machines de plusieurs secondes. C'est notamment le cas lorsque l'assemblage des produits doit répondre à des tolérances très strictes. Cependant, la connaissance métier indique qu'une correction manuelle peut parfois faire plus de dommage que réellement corriger le dépassement provoqué par la machine. Cela serait alors à l'origine d'un autre type de pertes, liées à la qualité. Dans ce cas de figure, l'évaluation de la performance peut être trompeuse. Une bonne pratique est-elle celle qui suit les normes ? Les pratiques sont-elles les entités à améliorer ou bien faut-il monter d'un cran et revoir les normes du processus afin de correspondre à la réalité de la production ? Encore une fois, l'évaluation des pertes des cadences par l'identification de comportements visant à réduire ces pertes n'est pas triviale.

Cette dernière limite en souligne une autre. En effet, ce travail a été volontairement mené pour se distinguer de la maintenance prédictive. Cependant, ce travail ne considère en aucun point des questions de qualité afin de se concentrer sur les pertes liées à des réduction ou interruption de la cadence. Une piste d'approfondissement serait alors de lier les différents produits aux données du département de la qualité, afin de savoir si beaucoup de produits ont été rejetés ou non.

Finalement, ce mémoire apporte une contribution scientifique et pratique au domaine de la valorisation de données industrielles visant à améliorer la performance des systèmes manufacturiers.

RÉFÉRENCES

Ahmed, Q., A Anifowose, F., & Khan, F. (2015). System availability enhancement using computational intelligence–based decision tree predictive model . *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability Institution of Mechanical Engineers* , 612-626.

Alfeo, A. L., Cimino, M. G., Manco, G., Ritacco, E., & Vaglini, G. (2020). Using an autoencoder in the design of an anomaly detector for smart manufacturing. *Pattern Recognition Letters*, 136, 272-278.

American Society for Quality. (s.d.). *Fishbone Diagram*. Récupéré sur ASQ: https://asq.org/quality-resources/fishbone

Bhogal, R., & Garg, A. (2020, June). Anomaly detection and fault prediction of breakdown to repair process using mining techniques. *In 2020 International Conference on Intelligent Engineering and Management (ICIEM)* (pp. 240-245).

Bouché, P., & Zanni-Merk, C. (2011). Improving the performance of production lines with an expert system using a stochastic approach. *Simulation*, *87*(5), 363-383.

Campos, J., Sharma, P., Gabiria, U. G., Jantunen, E., & Baglee, D. (2017). A big data analytical architecture for the Asset Management. *Procedia CIRP*, 64, 369-374.

Cerquitelli, T., Pagliari, D. J., Calimera, A., Bottaccioli, L., Patti, E., Acquaviva, A., & Poncino, M. (2021). Manufacturing as a data-driven practice: methodologies, technologies, and tools. *Proceedings of the IEEE*, 109(4), 399-422.

Chen, Y. C., Chien, C. F., Chen, Y. Y., & Wang, C. Y. (2019, April). A Data Mining Approach for Optimizing Manufacturing Parameters of Wire Bonding Process in IC Packaging Industry and Empirical Study. In *2019 IEEE International Conference on Smart Manufacturing, Industrial & Logistics Engineering (SMILE)* (pp. 53-57).

Çiflikli, C., & Kahya-Özyirmidokuz, E. (2010). Implementing a data mining solution for enhancing carpet manufacturing productivity. *Knowledge-Based Systems*, *23*(8), 783-788.

Dagnino, A. (2019, August). Data mining methods to analyze alarm logs in IoT process control systems. In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)* (pp. 323-330).

Denkena, B., Dittrich, M. A., Keunecke, L., & Wilmsmeier, S. (2020). Continuous modelling of machine tool failure durations for improved production scheduling. *Production Engineering*, *14*(2), 207-215.

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, 166, 114060.

Dupuis, A., Dadouchi, C., & Agard, B. (2022). Predicting crop rotations using process mining techniques and Markov principals. *Computers and Electronics in Agriculture*, 194, 106686.

Elangovan, M., Kumar, S., & Bharathi Ganesh, H. (2015). Condition monitoring of a valve in a reciprocating compressor using machine learning approach. *International Journal of Applied Engineering Research*, 33078-33081.

ELSEVIER. (s.d.). Récupéré sur Scopus: https://www.scopus.com

Fraunhofer Institute for Applied Information Technology. (2021). PM4Py Fraunhofer FIT. Récupéré sur PM4Py Documentation: https://pm4py.fit.fraunhofer.de/documentation

Google. (s.d.). Récupéré sur Google Scholar.

Haasbroek, A., Strydom, J. J., McCoy, J. T., & Auret, L. (2018). Fault Diagnosis for an Industrial High Pressure Leaching Process with a Monitoring Dashboard. *IFAC-PapersOnLine*, 51(21), 117-122.

Hrcka, L., Simoncicova, V., Tadanai, O., Tanuska, P., & Vazan, P. (2017, April). Using text mining methods for analysis of production data in automotive industry. In *Computer Science On-Line Conference* (pp. 393-403). Springer, Cham.

IBM. (2021, 08 17). *IBM SPSS Modeler documentation*. Récupéré sur Documentation IBM: https://www.ibm.com/docs/fr/spss-modeler/saas?topic=dm-crisp-help-overview

IBM. (s.d.). *Big data analytics*. Récupéré sur https://www.ibm.com/analytics/big-data-analytics

Kang, N., Zhao, C., Li, J., & Horst, J. A. (2016). A Hierarchical structure of key performance indicators for operation management and continuous improvement in production systems. *International Journal of Production Research*, *54*(21), 6333-6350.

Klaeger, T., Schult, A., & Oehm, L. (2019, July). Using anomaly detection to support classification of fast running packaging processes. In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)* (Vol. 1, pp. 343-348).

Lee, J., Davari, H., Singh, J., & Pandhare, V. (2018). Industrial Artificial Intelligence for industry 4.0-based manufacturing systems. *Manufacturing letters*, 18, 20-23.

Liang, Z., Wang, H., Ding, X., & Mu, T. (2021). Industrial time series determinative anomaly detection based on constraint hypergraph. *Knowledge-Based Systems*, 233, 107548.

Lindberg, C. F., Tan, S., Yan, J., & Starfelt, F. (2015). Key performance indicators improve industrial performance. *Energy Procedia*, *75*, 1785-1790.

Office Québecois de la Langue Française. (2019). *FICHE TERMINOLOGIQUE - Valorisation de données*. *Office Québecois de la Langue Française*: https://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26557454

Pabolu, V. K. R., Shrivastava, D., & Kulkarni, M. S. (2022). A Dynamic System to Predict an Assembly Line Worker's Comfortable Work-Duration Time by Using the Machine Learning Technique. *Procedia CIRP*, 106, 270-275.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526-534.

Soltanali, H., Khojastehpour, M., & Farinha, J. T. (2021). Measuring the production performance indicators for food processing industry. *Measurement*, *173*, 108394.

Souza, M. L. H., da Costa, C. A., de Oliveira Ramos, G., & da Rosa Righi, R. (2021). A feature identification method to explain anomalies in condition monitoring. *Computers in Industry*, 133, 103528.

Trunzer, E., Weiß, I., Folmer, J., Schrüfer, C., Vogel-Heuser, B., Erben, S., ... & Vermum, C. (2017, December). Failure mode classification for control valves for supporting data-driven

fault detection. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 2346-2350).

Van Der Aalst, W. M. (2019). A practitioner's guide to process mining: limitations of the directly-follows graph. *Procedia Computer Science*, 164, 321-328.

Van der Aalst, W. M. (2016). *Process mining: data science in action*. Springer.